A Review of Deep Learning Techniques for Speech Processing

AMBUJ MEHRISH, Singapore University of Technology and Design, Singapore
NAVONIL MAJUMDER, Singapore University of Technology and Design, Singapore
RISHABH BHARDWAJ, Singapore University of Technology and Design, Singapore
RADA MIHALCEA, University of Michigan, USA
SOUJANYA PORIA, Singapore University of Technology and Design, Singapore

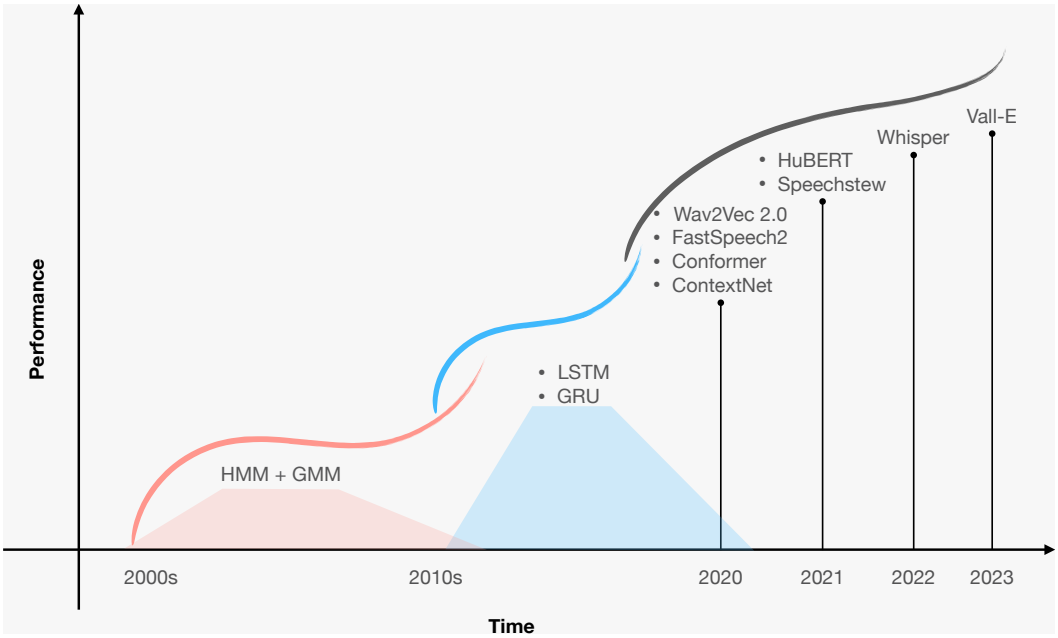

Fig. 1. Evolution of speech processing models over the years.

The field of speech processing has undergone a transformative shift with the advent of deep learning. The use of multiple processing layers has enabled the creation of models capable of extracting intricate features from speech data. This development has paved the way for unparalleled advancements in speech recognition, text-to-speech synthesis, automatic speech recognition, and emotion recognition, propelling the performance of these tasks to unprecedented heights. The power of deep learning techniques has opened up new avenues for research and innovation in the field of speech processing, with far-reaching implications for a range of industries and applications. This review paper provides a comprehensive overview of the key deep learning models and their applications in speech-processing tasks. We begin by tracing the evolution of speech processing research, from early approaches, such as MFCC and HMM, to more recent advances in deep learning architectures, such as CNNs, RNNs, transformers, conformers, and diffusion models. We categorize the approaches and compare their strengths and weaknesses for solving speech-processing tasks. Furthermore, we extensively cover various speech-processing tasks, datasets, and benchmarks used in the literature and describe how different deep-learning networks have been utilized to tackle these tasks. Additionally, we discuss the challenges and future directions of deep learning in speech processing, including the need for more parameter-efficient, interpretable models and the potential of deep learning for multimodal speech processing. By examining the field's evolution, comparing and contrasting different approaches, and highlighting future directions and challenges, we hope to inspire further research in this exciting and rapidly advancing field.

Contents

1 Introduction

Humans use language to express their feelings and emotions effectively. A language consists of a set of words (vocabulary) and grammar, which is a set of rules to tell us how we use them, and it can be in the form of text, sign, or speech. Speech or spoken language is defined as the way to communicate using language by a phonetic combination[1]ocally Phonetics refers to humans' way of producing and perceiving sounds. of consonant and vowel sounds, referring to a word in the vocabulary.

Speech processing is a study and processing method for speech signals. It covers various tasks, including automatic speech recognition (ASR). [382, 605], speaker recognition (SR) [30], and speech synthesis or text-to-speech [388]. In recent decades, speech processing has gained significant importance owing to its wide-ranging applications in telecommunications, healthcare, and entertainment. Statistical modelling methods, primarily Hidden Markov Models (HMMs), have played a pivotal role in propelling the field forward [144, 431].

Over the past few years, the field of speech processing has been transformed by introducing powerful tools, including deep learning. Figure 1 illustrates the evolution of speech processing models over the years, the rapid development of deep learning architecture for speech processing reflects the growing complexity and diversity of the field. This technology has revolutionized the analysis and processing of speech signals using deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). These architectures have proven highly effective in various speech-processing applications, such as speech recognition, speaker recognition, and speech synthesis. This study comprehensively overviews the most critical and emerging deep-learning techniques and their potential applications in various speech-processing tasks.

The success of deep learning in speech processing can be attributed to its ability to automatically learn relevant features from raw speech signals, thereby reducing the need for hand-crafted feature engineering. This has led to significant improvements in speech processing performance, particularly in noisy environments and in the presence of various accents and dialects.

Several recent studies have demonstrated the effectiveness of deep learning architectures in speech processing. For instance, in [179], the authors showed that DNNs could significantly improve speech recognition accuracy over traditional HMM-based systems. Similarly, in [3], the authors demonstrated the effectiveness of CNNs in speech recognition tasks, while in [156], RNNs were shown to be effective in speech recognition and synthesis. Furthermore, recent advances in deep learning, such as attention mechanisms [81] and transformers [532], have further improved the performance of speech processing systems. Attention mechanisms allow the model to selectively focus on relevant parts of the input signal, while transformers enable the modeling of long-range dependencies in the signal.

Despite these advancements, there remain challenges in deep learning-based speech processing, such as the need for large amounts of labeled data, the interpretability of the models, and their robustness to various environmental factors. This paper provides an extensive overview of deep learning architectures used in speech-processing applications. Speech processing involves analyzing, synthesizing, and recognizing speech signals, and deep learning techniques have shown significant advancements in this field.

The paper starts by covering the background including the definition of speech signals, speech features, and traditional models that are primarily non-neural architectures. Then the paper dives into deep-learning architectures for speech including but not limited to RNNs, CNNs, Transformers, GNNs, and diffusion models. As an important area of speech processing, the survey paper also

[1] v

dedicates a section to the representation of learning techniques. After discussing architectures and models, the paper talks about a wide range of tasks in speech processing where deep learning has shown significant improvements such as speech recognition, speech synthesis, speaker recognition, speech-to-speech translation, and speech synthesis. After discussing the basics, model architectures, and tasks in the domain, we discuss advanced transfer learning techniques such as domain adaptation, meta-learning, and parameter-efficient transfer learning. In the conclusion, we discuss potential future directions in the field.

Why this paper?. Deep learning has become a powerful tool in speech processing because it automatically learns high-level representations of speech signals from raw audio data. As a result, significant advancements have been made in various speech-processing tasks, including speech recognition, speaker identification, speech synthesis, and more. These tasks are essential in various applications, such as human-computer interaction, speech-based search, and assistive technology for people with speech impairments. For example, virtual assistants like Siri and Alexa use speech recognition technology, while audiobooks and in-car navigation systems rely on text-to-speech systems.

Given the wide range of applications and the rapidly evolving nature of deep learning, a comprehensive review paper that surveys the current state-of-the-art techniques and their applications in speech processing is necessary. Such a paper can help researchers and practitioners stay up-to-date with the latest developments and trends and provide insights into potential areas for future research. However, to the best of our knowledge, no current work covers a broad spectrum of speech-processing tasks.

A review paper on deep learning for speech processing can also be a valuable resource for beginners interested in learning about the field. It can provide an overview of the fundamental concepts and techniques used in deep learning for speech processing and help them gain a deeper understanding of the field.

While some survey papers focus on specific speech-processing tasks such as speech recognition, a broad survey would cover a wide range of other tasks such as speaker recognition speech synthesis, and more. A broad survey would highlight the commonalities and differences between these tasks and provide a comprehensive view of the advancements made in the field.

2 Background

Before moving on to deep neural architectures, we discuss basic terms used in speech processing, low-level representations of speech signals, and traditional models used in the field.

2.1 Speech Signals

In the context of sound, signals are defined as variations of pressure in the air, which can be perceived by humans as sound. When such signals are produced by humans to facilitate spoken communication, they are referred to as speech signals. For computers to process speech signals, they need to be converted from one quantity, air pressure, to another quantity, usually an electrical signal. This conversion is achieved with the help of transducers. In the context of speech processing, the shape of a periodic signal, which repeats after a fixed length of time, is called its waveform. The inverse of the period is known as the frequency of the signal. Timbre, on the other hand, refers to the way humans perceive the quality of sound. Speech signals are usually digitized to facilitate processing. This involves converting them to a series of numbers by measuring the amplitude of the signal at a fixed time interval. The number of samples collected per second defines the sampling rate.

2.2 Speech Features

Speech features are numerical representations of speech signals that are used for analysis, recognition, and synthesis. Broadly, speech signals can be classified into two categories: time-domain features and frequency-domain features.

Time-domain features are derived directly from the amplitude of the speech signal over time. These are simple to compute and often used in real-time speech-processing applications. Some common time-domain features include:

- Energy: Energy is a measure of the overall amplitude of the speech signal over time. It can be computed by taking the squared value of each sample in the signal and summing them over a window of time.
- Zero-crossing rate: The zero-crossing rate refers to the frequency at which the speech signal intersects the zero-axis within a given time frame. The calculation involves tallying the instances where the signal changes polarity during a specific window of time.
- Pitch: Pitch, the perceived tone of a speaker's voice, is quantified by measuring the fundamental frequency of the speech signal. This can be calculated through the use of pitch detection algorithms or by utilizing autocorrelation.
- Linear predictive coding (LPC): LPC is a technique that models the speech signal as a linear combination of past samples using an autoregressive model. The model parameters are estimated using methods such as the Levinson-Durbin algorithm, and the resulting coefficients are used as a feature representation for speech-processing tasks.

Frequency-domain features are derived from the signal represented in the frequency domain also known as its spectrum. A spectrum captures the distribution of energy as a function of frequency. Spectrograms are two-dimensional visual representations capturing the variations in a signal's spectrum over time. When compared against time-domain features, it is generally more complex to compute frequency-domain features as they tend to involve time-frequency transform operations such as Fourier transform.

- Mel-spectrogram: A time-frequency representation of a speech signal can be created by utilizing a mel-scale filter bank applied to the short-time Fourier transform (STFT) of the signal. This method allows for a comprehensive representation of the spectral content of the speech signal over time
- Mel-frequency cepstral coefficients (MFCCs):To extract features for speech processing, the Mel-frequency cepstral coefficients (MFCCs) technique is commonly utilized. This involves several processing steps applied to the short-time power spectrum of the audio signal, including filtering the spectrum through a mel-scale filterbank, taking the logarithm of the filtered spectrum, and utilizing the discrete cosine transform (DCT) to obtain the cepstral coefficients

Other types of speech features include formant frequencies, pitch contour, cepstral coefficients, wavelet coefficients, and spectral envelope. These features can be used for various speech-processing tasks, including speech recognition, speaker identification, emotion recognition, and speech synthesis.

In speech processing, frequency-based representations such as Mel spectrogram and MFCC are widely used since they are more robust to noise as compared to temporal variations of the sound. Time-domain features can be useful when the task warrants <this information (such as pauses, emotions, phoneme duration, speech segments)>. It is noteworthy that the time-domain and frequency-domain features tend to capture different sets of information and thus can be used in conjunction to solve a task.

2.3 Traditional models for speech processing

Traditional speech representation learning algorithms based on shallow models utilize basic non-parametric models for extracting features from speech signals. The primary objective of these models is to extract significant features from the speech signal through mathematical operations, such as Fourier transforms, wavelet transforms, and linear predictive coding (LPC). The extracted features serve as inputs to classification or regression models. The shallow models aim to extract meaningful features from the speech signal to enable the classification or regression model to learn and make accurate predictions.

- Gaussian Mixture Models (GMMs): GMMs are generative models that represent the probability distribution of a speech feature vector using a weighted sum of Gaussian distributions. GMMs are commonly used for speaker identification and speech recognition tasks. In speaker identification, GMMs are used to model the distribution of speaker-specific features, while in speech recognition, GMMs are used to model the acoustic properties of speech sounds.
- Support Vector Machines (SVMs): SVMs are a class of supervised learning algorithms that are commonly used for speech classification tasks, such as speaker identification and phoneme recognition. SVMs work by finding the optimal hyperplane that separates the different classes in the feature space.
- Hidden Markov Models (HMMs): Hidden Markov Models (HMMs) have become a popular tool for performing various speech recognition tasks, especially automatic speech recognition (ASR). In ASR, HMMs are used to model the probability distribution of speech sounds by incorporating a sequence of hidden states along with corresponding observations. The Baum-Welch algorithm, a variant of the Expectation-Maximization algorithm, is commonly used to train HMMs. By utilizing HMMs, it becomes possible to predict the most probable sequence of speech sounds given an input speech signal.
- The K-nearest neighbors (KNN) algorithm is a straightforward classification approach that involves identifying the K-nearest neighbors of a given input feature vector within the training data and assigning it to the class that appears most frequently among the neighbors. This algorithm has gained popularity in various speech-related applications such as speaker identification and language recognition. KNN offers a practical and intuitive method to classify speech data and has been employed successfully in many real-world scenarios.
- Decision trees: Decision trees are a class of supervised learning algorithms that are commonly used for speech classification tasks. They work by recursively partitioning the feature space into smaller regions based on the values of the features. Each partition is associated with a decision rule that assigns the corresponding input feature vector to a class.

In summary, traditional speech representation learning algorithms based on shallow models involve extracting features from the speech signal and using these features as inputs to a classification or regression model. These algorithms have been widely used in speech processing applications such as speech recognition, speaker identification, and speech synthesis, but have been largely replaced by more advanced representation learning algorithms such as deep neural networks.

3 Deep Learning Architectures and Their Applications in Speech Processing Tasks

3.1 Recurrent Neural Networks (RNNs)

It is natural to consider Recurrent Neural Networks for various speech processing tasks since the input speech signal is inherently a dynamic process [462]. RNNs can model a given time-varying (sequential) patterns that were otherwise hard to capture by standard feedforward neural architectures. Initially, RNNs were used in conjunction with HMMs where the sequential data is

first modeled by HMMs while localized classification is done by the neural network. However, such a hybrid model tends to inherit limitations of HMMs, for instance, HMM requires task-specific knowledge and independence constraints for observed states [42]. To overcome the limitations inherited by the hybrid approach, end-to-end systems completely based on RNNs became popular for sequence transduction tasks such as speech recognition and text[153, 240]. Next, we discuss RNN and it's variants:

3.1.1 RNN Models

Vanilla RNN. Give input sequence of T states $(x_1, \ldots, x_T)$ with $x_i \in \mathbb{R}^d$, the output state at time t can be computed as

$$h_t = \mathcal{H}(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{1}$$

$$y_t = W_{hy}h_t + b_y \tag{2}$$

where W_{hh}, W_{hx}, W_{yh} are weight matrices and b_h, b_y are bias vectors. $\mathcal{H}$ is non-linear activation functions such as Tanh, ReLU, and Sigmoid. RNNs are made of high dimensional hidden states, notice h_t in the above equation, which makes it possible for them to model sequences and help overcome the limitation of feedforward neural networks. The state of the hidden layer is conditioned on the current input and the previous state, which makes the underlying operation recursive. Essentially, the hidden state h_{t-1} works as a memory of past inputs $\{x_k\}_{k=1}^{t-1}$ that influence the current output y_t.

Bidirectional RNNs. For numerous tasks in speech processing, it is more effective to process the whole utterance at once. For instance, in speech recognition, one-shot input transcription can be more robust than transcribing based on the partial (i.e. previous) context information [156]. The vanilla RNN has a limitation in such cases as they are unidirectional in nature, that is, output y_t is obtained from $\{x_k\}_{k=1}^{t}$, and thus, agnostic of what comes after time t. Bidirectional RNNs (BRNNs) were proposed to overcome such shortcomings of RNNs [470]. BRRNs encode both future and past (input) context in separate hidden layers. The outputs of the two RNNs are then combined at each time step, typically by concatenating them together, to create a new, richer representation that includes both past and future context.

$$\overrightarrow{h_t} = \mathcal{H}(W_{\overrightarrow{hh}}\overrightarrow{h}_{t-1} + W_{\overrightarrow{xh}}x_t + b_{\overrightarrow{h}}) \tag{3}$$

$$\overleftarrow{h_t} = \mathcal{H}(W_{\overleftarrow{hh}}\overleftarrow{h}_{t+1} + W_{\overleftarrow{xh}}x_t + b_{\overleftarrow{h}}) \tag{4}$$

$$y_t = W_{\overrightarrow{hy}}\overrightarrow{h}_t + W_{\overleftarrow{hy}}\overleftarrow{h}_t + b_y \tag{5}$$

where high dimensional hidden states $\overrightarrow{h}_{t-1}$ and $\overleftarrow{h}_{t+1}$ are hidden states modeling the forward context from $1, 2, \ldots, t-1$ and backward context from $T, T-1, \ldots, t+1$, respectively.

Long Short-Term Memory. Vanilla RNNs are observed to face another limitation, that is, vanishing gradients that do not allow them to learn from long-range context information. To overcome this, a variant of RNN, named as LSTM, was specifically designed to address the vanishing gradient problem and enable the network to selectively retain (or forget) information over longer periods of time [181]. This attribute is achieved by maintaining separate purpose-built memory cells in the network: the long-term memory cell c_t and the short-term memory cell h_t. In Equation (2),

LSTM redefines the operator $\mathcal{H}$ in terms of forget gate f_t, input gate i_t, and output gate o_t,

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \tag{6}$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \tag{7}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{8}$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \tag{9}$$
$$h_t = o_t \odot \tanh(c_t), \tag{10}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is a logistic sigmoid activation function. c_t is a fusion of the information from the previous state of the long-term memory c_{t-1}, the previous state of short-term memory h_{t-1}, and current input x_t. W and b are weight matrices and biases. $\odot$ is the element-wise vector multiplication or Hadamard operator. Bidirectional LSTMs (BLSTMs) can capture longer contexts in both forward and backward directions [153].

Gated Recurrent Units. Gated Recurrent Units (GRU) aim to be a computationally-efficient approximate of LSTM by using only two gates (vs three in LSTM) and a single memory cell (vs two in LSTM). To control the flow of information over time, a GRU uses an update gate z_t to decide how much of the new input to be added to the previous hidden state and a reset gate r_t to decide how much of previous hidden state information to be forgotten.

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1}), \tag{11}$$
$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1}), \tag{12}$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_{xh}x_t + W_{rh}(r_t \odot h_{t-1})), \tag{13}$$

where $\odot$ is element-wise multiplication between the two vectors (Hadamard product).

RNNs and their variants are widely used in various deep learning applications like speech recognition, synthesis, and natural language understanding. Although seq2seq based on recurrent architectures such as LSTM/GRU has made great strides in speech processing, they suffer from the drawback of slow training speed due to internal recurrence. Another drawback of the RNN family is their inability to leverage information from long distant time steps accurately.

Connectionist Temporal Classification. Connectionist Temporal Classification (CTC) [154] is a scoring and output function commonly used to train LSTM networks for sequence-based problems with variable timing. CTC has been applied to several tasks, including phoneme recognition, ASR, and other sequence-based problems. One of the major benefits of CTC is its ability to handle unknown alignment between input and output, simplifying the training process. When used in ASR [100, 101, 371], CTC eliminates the need for manual data labeling by assigning probability scores to the output given any input signal. This is particularly advantageous for tasks such as speech recognition and handwriting recognition, where the input and output can vary in size. CTC also solves the problem of having to specify the position of a character in the output, allowing for more efficient training of the neural network without post-processing the output. Finally, the CTC decoder can transform the neural network output into the final text without post-processing.

3.1.2 Application

The utilization of RNNs in popular products such as Google's voice search and Apple's Siri to process user input and predict the output has been well-documented [171, 297]. RNNs are frequently utilized in speech recognition tasks, such as the prediction of phonetic segments from audio signals [403]. They excel in use cases where context plays a vital role in outcome prediction and are distinct

from CNNs as they utilize feedback loops to process a data sequence that informs the final output [403].

In recent times, there have been advancements in the architecture of RNNs, which have been primarily focused on developing end-to-end (E2E) models [295, 400] for ASR. These E2E models have replaced conventional hybrid models and have displayed substantial enhancements in speech recognition [295, 296]. However, a significant challenge faced by E2E RNN models is the synchronization of the input speech sequence with the output label sequence [153]. To tackle this issue, a loss function called CTC [154] is utilized for training RNN models, allowing for the repetition of labels to construct paths of the same length as the input speech sequence. An alternative method is to employ an Attention-based Encoder-Decoder (AED) model based on RNN architecture, which utilizes an attention mechanism to align the input speech sequence with the output label sequence. However, AED models tend to perform poorly on lengthy utterances.

The development of Bimodal Recurrent Neural Networks (BRNN) has led to significant advancements in the field of Audiovisual Speech Activity Detection (AV-SAD) [510]. BRNNs have demonstrated immense potential in improving the performance of speech recognition systems, particularly in noisy environments, by combining information from various sources. By integrating separate RNNs for each modality, BRNNs can capture temporal dependencies within and across modalities. This leads to successful outcomes in speech-based systems, where integrating audio and visual modalities is crucial for accurate speech recognition. Compared to conventional audio-only systems, BRNN-based AV-SAD systems display superior performance, particularly in challenging acoustic conditions where audio-only systems might struggle.

To enhance the performance of continuous speech recognition, LSTM networks have been utilized in hybrid architectures alongside CNNs [407]. The CNNs extract local features from speech frames that are then processed by LSTMs over time [407]. LSTMs have also been employed for speech synthesis, where they have been shown to enhance the quality of statistical parametric speech synthesis [407].

Aside from their ASR and speech synthesis applications, LSTM networks have been utilized for speech post-filtering. To improve the quality of synthesized speech, researchers have proposed deep learning-based post-filters, with LSTMs demonstrating superior performance over other post-filter types [95]. Bidirectional LSTM (Bi-LSTM) is another variant of RNN that has been widely used for speech synthesis [132]. Several RNN-based analysis/synthesis models such as WaveNet [394], SampleRNN [366], and Tacotron have been developed. These neural vocoder models can generate high-quality synthesized speech from acoustic features without requiring intermediate vocoding steps.

3.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a specialized class of deep neural architecture consisting of one or more pairs of alternating convolutional and pooling layers. A convolution layer applies filters that process small local parts of the input, where these filters are replicated along the whole input space. A pooling layer converts convolution layer activations to low resolution by taking the maximum filter activation within a specified window and shifting across the activation map. CNNs are variants of fully connected neural networks widely used for processing data with grid-like topology. For example, time-series data (1D grid) with samples at regular intervals or images (2D grid) with pixels constitute a grid-like structure.

As discussed in Section 2, the speech spectrogram retains more information than hand-crafted features, including speaker characteristics such as vocal tract length differences across speakers, distinct speaking styles causing formant to undershoot or overshoot, etc. Also, explicitly expressed these characteristics in the frequency domain. The spectrogram representation shows very strong

correlations in time and frequency. Due to these characteristics of the spectrogram, it is a suitable input for a CNN processing pipeline that requires preserving locality in both frequency and time axis. For speech signals, modeling local correlations with CNNs will be beneficial. The CNNs can also effectively extract the structural features from the spectrogram and reduce the complexity of the model through weight sharing. This section will discuss the architecture of 1D and 2D CNNs used in various speech-processing tasks.

3.2.1 CNN Model Variants

2D CNN. Since spectrograms are two-dimensional visual representations, one can leverage CNN architectures widely used for visual data processing (images and videos) by performing convolutions in two dimensions. The mathematical equation for a 2D convolutional layer can be represented as:

$$y_{i,j}^{(k)} = \sigma\Big(\sum_{l=1}^{L} \sum_{m=1}^{M} x_{i+l-1,j+m-1}^{(l)} w_{l,m}^{(k)} + b^{(k)} \Big) \tag{14}$$

Here, $x_{i,j}^{(l)}$ is the pixel value of the l^{th} input channel at the spatial location (i, j), $w_{l,m}^{(k)}$ is the weight of the m^{th} filter at the l^{th} channel producing the k^{th} feature map, and $b^{(k)}$ is the bias term for the k^{th} feature map.

The output feature map $y_{i,j}^{(k)}$ is obtained by convolving the input image with the filters and then applying an activation function σ to introduce non-linearity. The convolution operation involves sliding the filter window over the input image, computing the dot product between the filter and the input pixels at each location, and producing a single output pixel.

However, there are some drawbacks to using a 2D CNN for speech processing. One of the main issues is that 2D convolutions are computationally expensive, especially for large inputs. This is because 2D convolutions involve many multiplications and additions, and the computational cost grows quickly with the input size.

To address this issue, a 1D CNN can be designed to operate directly on the speech signal without needing a spectrogram. 1D convolutions are much less computationally expensive than 2D convolutions because they only operate on one dimension of the input. This reduces the multiplications and additions required, making the network faster and more efficient. In addition, 1D feature maps require less memory during processing, which is especially important for real-time applications. A neural network's memory requirements are proportional to its feature maps' size. By using 1D convolutions, the size of the feature maps can be significantly reduced, which can improve the efficiency of the network and reduce the memory requirements.

1D CNN. 1D CNN is essentially a special case of 2D CNN where the height of the filter is equal to the height the spectogram. Thus, the filter only slides along the temporal dimension and the height of the resultant feature maps is one. As such, 1D convolutions are computationally less expensive and memory efficient [254], as compared to 2D CNNs. Several studies [6, 239, 255] have shown that 1D CNNs are preferable to their 2D counterparts in certain applications. For example, Alsabhan [11] found that the performance of predicting emotions with a 2D CNN model was lower compared to a 1D CNN model.

1D convolution is useful in speech processing for several reasons:

- Since, speech signals are sequences of amplitudes sampled over time, 1D convolution can be applied along temporal dimension to capture temporal variations in the signal.
- *Robustness to distortion and noise:* Since, 1D convolution allows local feature extraction, the resultant features are often resilient to global distortions of the signal. For instance,

a speaker might be interrupted in the middle of an utterance. Local features would still produce robust representations for those relevant spans, which is key to ASR, among many speech processing task. On the other hand, speech signals are often contaminated with noise, making extracting meaningful information difficult. 1D convolution followed by pooling layers can mitigate the impact of noise [174], improving speech recognition systems' accuracy.

The basic building block of a 1D CNN is the convolutional layer, which applies a set of filters to the input data. A convolutional layer employs a collection of adjustable parameters called filters to carry out convolution operations on the input data, resulting in a set of feature maps as the output, which represent the activation of each filter at each position in the input data. The size of the feature maps depends on the size of the input data, the size of the filters, and the number of filters used. The activation function used in a 1D CNN is typically a non-linear function, such as the rectified linear unit (ReLU) function.

Given an input sequence x of length N, a set of K filters W_k of length M, and a bias term b_k, the output feature map y_k of the k^{th} filter is given by

$$y_k[n] = \text{ReLU}(b_k + \sum_{m=0}^{M-1} W_k[m] * x[n-m]) \tag{15}$$

where n ranges from $M-1$ to $N-1$, and $*$ denotes the convolution operation.

After the convolutional layer, the output tensor is typically passed through a pooling layer, reducing the feature maps' size by down-sampling. The most commonly used pooling operation is the max-pooling, which keeps the maximum value from a sliding window across each feature map.

CNNs typically previously popular approaches, such as HMMs and GMM-UBM, in many cases. Furthermore, CNNs have the potential to learn features that are robust to variations in speech signals caused by different speakers, accents, and background noise due to their three key properties: locality, weight sharing, and pooling. The locality property allows for greater robustness against non-white noise, as good features can be computed locally from cleaner parts of the spectrum. This means that only a smaller number of features are affected by the noise, giving higher network layers a better chance to handle the noise by combining higher-level features computed for each frequency band. This is a clear improvement over standard, fully connected neural networks that handle all input features in the lower layers. Hence, locality reduces the network weights that need to be learned.

3.2.2 Application

CNNs have proven to be versatile tools for a range of speech-processing tasks. They have been successfully applied to speech recognition [4, 382], including in hybrid NN-HMM models for speech recognition, and can be used for multi-class classification of words [5]. In addition, CNNs have been proposed for speaker recognition in an emotional speech, with a constrained CNN model presented in [481].

CNNs, both 1D and 2D, have emerged as the core building block for various speech processing models, including acoustic models [157, 266, 468] in ASR systems. For instance, in 2021, researchers from Facebook AI proposed wav2vec2.0 [468], a hybrid ASR system based on CNNs for learning representations of raw speech signals that were then fed into a transformer-based language model. The system achieved state-of-the-art results on several benchmark datasets.

Similarly, Google's VGGVox [88] used a CNN with VGG architecture to learn speaker embeddings from Mel spectrograms, achieving state-of-the-art results in speaker recognition. CNNs have also been widely used in developing state-of-the-art speech enhancement and text-to-speech

architectures. For instance, the architecture proposed in [304, 520] for Deep Noise Suppression (DNS) [446] challenge and Google's Tacotron2 [476] are examples of models that use CNNs as their core building blocks. In addition to traditional tasks like ASR and speaker identification, CNNs have also been applied to non-traditional speech processing tasks like emotion recognition [224], Parkinson's disease detection [218], language identification [483] and sleep apnea detection [482]. In all these tasks, CNN extracted features from speech signals and fed them into the task classification model.

3.2.3 Temporal Convolution Neural Networks

Recurrent neural networks, including RNNs, LSTMs, and GRUs, have long been popular for deep-learning sequence modeling tasks. They are especially favored in the speech-processing domain. However, recent studies have revealed that certain CNN architectures can achieve state-of-the-art accuracy in tasks such as audio synthesis, word-level language modelling, and machine translation, as reported in [98, 227, 228]. The advantage of convolutional neural networks is that they enable faster training by allowing parallel computation. They can avoid common issues associated with recurrent models, such as the vanishing or exploding gradient problem or the inability to retain long-term memory.

In a recent study by Bai et al. [29], they proposed a generic Temporal Convolutional Neural Network (TCNN) architecture that can be applied to various speech-related tasks. This architecture combines the best practices of modern CNNs and has demonstrated comparable performance to recurrent architectures such as LSTMs and GRUs. The TCN approach could revolutionize speech processing by providing an alternative to the widely used recurrent neural network models.

3.2.4 TCNN Model Variants

The architecture of TCNN is based upon two principles:(1) There is no information "leakage" from future to past;(2) the architecture can map an input sequence of any length to an output sequence of the same length, similar to RNN. TCN consists of dilated, causal 1D fully-convolutional layers with the same input and output lengths to satisfy the above conditions. In other words, TCNN is simply a 1D fully-convolutional network (FCN) with casual convolutions as shown in Figure 2.

- *Causal Convolution [394]:* Causal convolution convolves the input at a specific time point t solely with the temporally-prior elements.
- *Dilated Convolution [606]:* By itself, causal convolution filters have a limited range of perception, meaning they can only consider a fixed number of elements in the past. Therefore, it is challenging to learn any dependency between temporally distant elements for longer sequences. Dilated convolution ameliorates this limitation by repeatedly applying dilating filters to expand the range of perception, as shown in Figure 2. The dilation is achieved by uniformly inserting zeros between the filter weights.
  Consider a 1-D sequence $x \in \mathbf{R}^n$ and a filter: $f : \{0, ..., k-1\} \rightarrow \mathbf{R}$, the dilated convolution operation F_d on an element y of the sequence is defined as

$$F_d(y) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i).x_{y-d.i}, \tag{16}$$

  where k is filter size, d is dilation factor, and $y - d.i$ is the span along the past. The dilation step introduces a fixed step between every two adjacent filter taps. When $d = 1$, a dilated convolution acts as a normal convolution. Whereas, for larger dilation, the filter acts on a wide but non-contiguous range of inputs. Therefore, dilation effectively expands the receptive field of the convolutional networks.

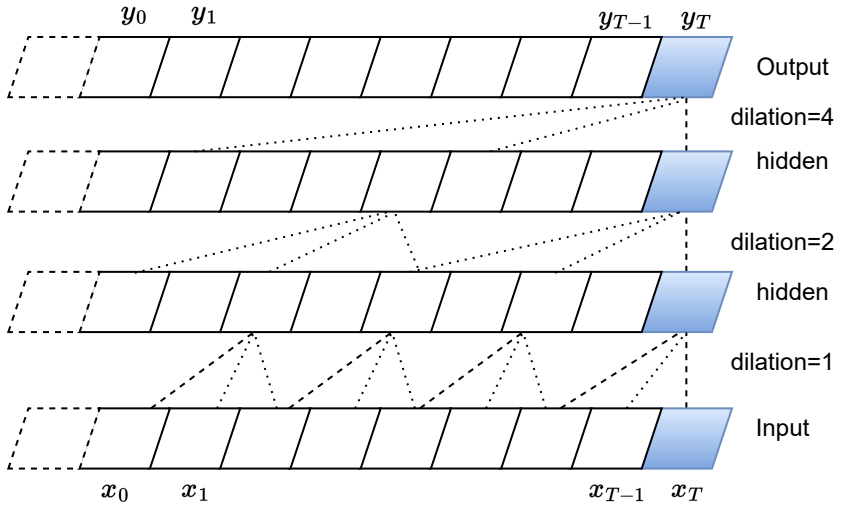

Fig. 2. TCNNs leverage causal and dilated convolutions to model temporal dependencies in sequential data. Causal convolutions ensure that future information is not used during training, while dilated convolutions increase the receptive field without increasing computational complexity. This makes TCNNs an effective and efficient solution for a wide range of tasks, including speech recognition, action recognition, and music analysis.

3.2.5 Application

Recent studies have shown that the TCNN architecture not only outperforms traditional recurrent networks like LSTMs and GRUs in terms of accuracy but also possesses a set of advantageous properties, including:

- Parallelism is a key advantage of TCNN over RNNs. In RNNs, time-step predictions depend on their predecessors' completion, which limits parallel computation. In contrast, TCNNs apply the same filter to each span in the input, allowing parallel application thereof. This feature enables more efficient processing of long input sequences compared to RNNs that process sequentially.
- The receptive field size can be modified in various ways to enhance the performance of TCNNs. For example, incorporating additional dilated convolutional layers, employing larger dilation factors, or augmenting the filter size are all effective methods. Consequently, TCNNs offer superior management of the model's memory size and are highly adaptable to diverse domains.
- When dealing with lengthy input sequences, LSTM and GRU models tend to consume a significant amount of memory to retain the intermediate outcomes for their numerous cell gates. On the other hand, TCNNs utilize shared filters throughout a layer, and the back-propagation route depends solely on the depth of the network. This makes TCNNs a more memory-efficient alternative to LSTMs and GRUs, especially in scenarios where memory constraints are a concern.

TCNNs can perform real-time speech enhancement in the time domain [402]. They have much fewer trainable parameters than earlier models, making them more efficient. TCNs have also been used for speech and music detection in radio broadcasts [206, 290]. They have been used for single

channel speech enhancement [315, 452] and are trained as filter banks to extract features from waveform to improve the performance of ASR [300].

3.3 Transformers

While recurrence in RNNs (Section 3.1) is a boon for neural networks to model sequential data, it is also a bane as the recurrence in time to update the hidden state intrinsically precludes parallelization. Additionally, although dedicated gated RNNs such as LSTM and GRU have helped to mitigate the vanishing gradient problem to some extent, it can still be a challenge to maintain long-term dependencies in RNNs.

Proposed by Vaswani et al. [532], Transformer solved a critical shortcoming of RNNs by allowing parallelization within the training sample, that is, facilitating the processing of the entire input sequence at once. Since then, the primary idea of using only the attention mechanism to construct an encoder and decoder has served as the basic recipe for many state-of-the-art architectures across the domains of machine learning. In this survey, we use **transformer** to denote architectures that are inspired by Transformer [45, 105, 162, 433, 434]. This section overviews the transformer's fundamental design proposed by Vaswani et al. [532] and its adaptations for different speech-related applications.

3.3.1 Basic Architecture

Transformer architecture [532] comprises an attention-based encoder and decoder, with each module consisting of a stack of identical blocks. Each block in the encoder and decoder consists of two sub-layers: a multi-head attention (MHA) mechanism and a position-wise fully connected feedforward network as described in Figure 3. The MHA mechanism in the encoder allows each input element to attend to every other element in the sequence, enabling the model to capture long-range dependencies in the input sequence. The decoder typically uses a combination of MHA and encoder-decoder attention to attend to both the input sequence and the previously generated output elements. The feedforward network in each block of the Transformer provides non-linear transformations to the output of the attention mechanism. Next, we discuss operations involved in transformer layers, that is, multi-head attention and position-wise feedforward network:

Attention in Transformers. Attention mechanism, first proposed by Bahdanau et al. [27], has revolutionized sequence modeling and transduction models in various tasks of NLP, speech, and computer vision [57, 76, 143, 544]. Broadly, it allows the model to focus on specific parts of the input or output sequence, without being limited by the distance between the elements. We can describe the attention mechanism as the mapping of a query vector and set of key-value vector pairs to an output. Precisely, the output vector is computed as a weighted summation of value vectors where the weight of a value vector is obtained by computing the compatibility between the query vector and key vector. Let, each query Q and key K are d_k dimensional and value V is d_v dimensional. Specific to the Transformer, the compatibility function between a query and each key is computed as their dot product between scaled by $\sqrt{d_k}$. To obtain the weights on values, the scaled dot product values are passed through a softmax function:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{17}$$

Here multiple queries, keys, and value vectors, are packed together in matrix form respectively denoted by $\mathbf{Q} \in \mathbb{R}^{N \times d_k}$, $\mathbf{K} \in \mathbb{R}^{M \times d_k}$, and $\mathbf{V} \in \mathbb{R}^{M \times d_v}$. N and M represent the lengths of queries and keys (or values). Scaling of dot product attention becomes critical to tackling the issue of small gradients with the increase in d_k [532].

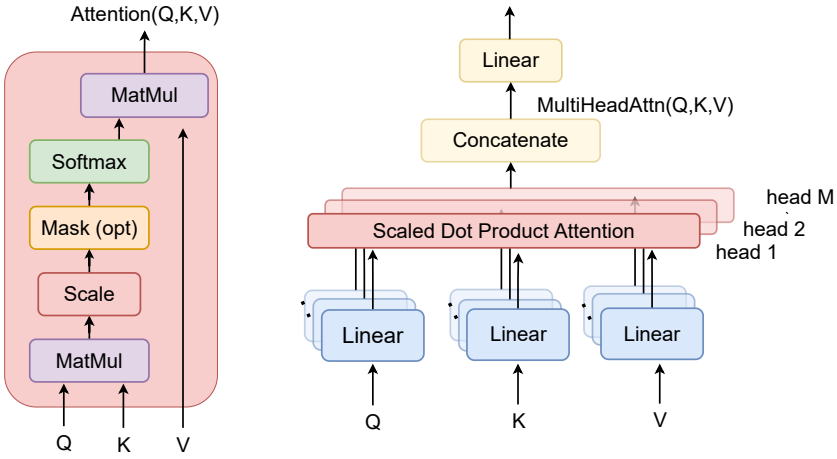

Fig. 3. Illustrations of attention (left) and multi-headed attention (right).

Instead of performing single attention in each transformer block, multiple attentions in lower-dimensional space have been observed to work better [532]. This observation gave rise to **Multi-Head Attention**: For h heads and dimension of tokens in the model d_m, the d_m-dimensional query, key, and values are projected h times to d_k, d_k, and d_v dimensions using learnable linear projections[2]. Each head performs attention operation as per Equation (17). The h d_v-dimensional are concatenated and projected back to d_m using another projection matrix:

$$\text{MultiHeadAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ... \text{head}_h)\mathbf{W}^O, \tag{18}$$

$$\text{with head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \tag{19}$$

Where $\mathbf{W}^Q, \mathbf{W}^K \in \mathbb{R}^{d_{model} \times d_k}, \mathbf{W}^V \in \mathbb{R}^{d_{model} \times d_v}, \mathbf{W}^O \in \mathbb{R}^{hd_v \times d_{model}}$ are learnable projection matrices. Intuitively, multiple attention heads allow for attending to parts of the sequence differently (e.g., longer-term dependencies versus shorter-term dependencies). Intuitively, multiple attention heads allow for attending in different representational spaces jointly.

Position-wise FFN. The position-wise FNN consists of two dense layers. It is referred to position-wise since the same two dense layers are used for each positioned item in the sequence and are equivalent to applying two 1×1 convolution layers.

Residual Connection and Normalization. Residual connection and Layer Normalization are employed around each module for building a deep model. For example, each encoder block output can be defined as follows:

$$H^{'} = \text{LayerNorm}(\text{SelfAttention}(X) + X) \tag{20}$$

$$H = \text{LayerNorm}(\text{FFN}(H^{'}) + H^{'}) \tag{21}$$

SelfAttention(.) denotes attention module with $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{X}$, where $\mathbf{X}$ is the output of the previous layer.

[2]Projection weights are neither shared across heads nor query, key, and values.

Transformer-based architecture turned out to be better than many other architectures such as RNN, LSTM/GRU, etc. One of the major difficulties when applying a Transformer to speech applications that it requires more complex configurations (e.g., optimizer, network structure, data augmentation) than the conventional RNN-based models. Speech signals are continuous-time signals with much higher dimensionality than text data. This high dimensionality poses significant computational challenges for the Transformer architecture, originally designed for sequential text data. Speech signals also have temporal dependencies, which means that the model needs to be able to process and learn from the entire signal rather than just a sequence of inputs. Also, speech signals are inherently variable and complex. The same sentence can be spoken differently and even by the same person at different times. This variability requires the model to be robust to differences in pitch, accent, and speed of speech.

3.3.2 Application

Recent advancements in NLP which lead to a paradigm shift in the field are highly attributed to the foundation models that are primarily a part of the transformers category, with self-attention being a key ingredient [41]. The recent models have demonstrated human-level performance in several professional and academic benchmarks. For instance, GPT4 scored within the top 10% of test takers on a simulated version of the Uniform Bar Examination [396]. While speech processing has not yet seen a shift in paradigm as in NLP owing to the capabilities of foundational models, even so, transformers have significantly contributed to advancement in the field including but not limited to the following tasks: automatic speech recognition, speech translation, speech synthesis, and speech enhancement, most of which we discuss in detail in Section 5.

RNNs and Transformers are two widely adopted neural network architectures employed in the domain of Natural Language Processing (NLP) and speech processing. While RNNs process input words sequentially and preserve a hidden state vector over time, Transformers analyze the entire sentence in parallel and incorporate an internal attention mechanism. This unique feature makes Transformers more efficient than RNNs [238]. Moreover, Transformers employ an attention mechanism that evaluates the relevance of other input tokens in encoding a specific token. This is particularly advantageous in machine translation, as it allows the Transformer to incorporate contextual information, thereby enhancing translation accuracy [238]. To achieve this, Transformers combine word vector embeddings and positional encodings, which are subsequently subjected to a sequence of encoders and decoders. These fundamental differences between RNNs and Transformers establish the latter as a promising option for various natural language processing tasks [238].

A comparative study on transformer vs. RNN [238] in speech applications found that transformer neural networks achieve state-of-the-art performance in neural machine translation and other natural language processing applications [238]. The study compared and analysed transformer and conventional RNNs in a total of 15 ASR, one multilingual ASR, one ST, and two TTS applications. The study found that transformer neural networks outperformed RNNs in most applications tested. Another survey of transformer-based models in speech processing found that transformers have an advantage in comprehending speech, as they analyse the entire sentence simultaneously, whereas RNNs process input words one by one.

Transformers have been successfully applied in end-to-end speech processing, including automatic speech recognition (ASR), speech translation (ST), and text-to-speech (TTS) [302]. In 2018, the Speech-Transformer was introduced as a no-recurrence sequence-to-sequence model for speech recognition. To reduce the dimension difference between input and output sequences, the model's architecture was modified by adding convolutional neural network (CNN) layers before feeding the features to the transformer. In a later study [380], the authors proposed a method to improve

the performance of end-to-end speech recognition models based on transformers. They integrated the connectionist temporal classification (CTC) with the transformer-based model to achieve better accuracy and used language models to incorporate additional context and mitigate recognition errors.

In addition to speech recognition, the transformer model has shown promising results in TTS applications. The transformer-based TTS model generates mel-spectrograms, followed by a WaveNet vocoder to output the final audio results [302]. Several neural network-based TTS models, such as Tacotron 2, DeepVoice 3, and transformer TTS, have outperformed traditional concatenative and statistical parametric approaches in terms of speech quality [302, 416, 476].

One of the strengths of Transformer-based architectures for neural speech synthesis is their high efficiency while considering the global context [157, 477]. The Transformer TTS model has shown advantages in training and inference efficiency over RNN-based models such as Tacotron 2 [476]. The efficiency of the Transformer TTS network can speed up the training about 4.25 times [302]. Moreover, Multi-Speech, a multi-speaker TTS model based on the Transformer [302], has demonstrated the effectiveness of synthesizing a more robust and better quality multi-speaker voice than naive Transformer-based TTS.

In contrast to the strengths of Transformer-based architectures in neural speech synthesis, large language models based on Transformers such as BERT [105], GPT [433], XLNet [595], and T5 [437] have limitations when it comes to speech processing. One of the issues is that these models require discrete tokens as input, necessitating using a tokenizer or a speech recognition system, introducing errors and noise. Furthermore, pre-training on large-scale text corpora can lead to domain mismatch problems when processing speech data. To address these limitations, dedicated frameworks have been developed for learning speech representations using transformers, including wav2vec [468], data2vec [23], Whisper [432], VALL-E [540], Unispeech [543], SpeechT5 [15] etc. We discuss some of them as follows.

- Speech representation learning frameworks, such as wav2vec, have enabled significant advancements in speech processing tasks. One recent framework, w2v-BERT [562], combines contrastive learning and MLM to achieve self-supervised speech pre-training on discrete tokens. Fine-tuning wav2vec models with limited labeled data has also been demonstrated to achieve state-of-the-art results in speech recognition tasks [24]. Moreover, XLS-R [19], another model based on wav2vec 2.0, has shown state-of-the-art results in various tasks, domains, data regimes, and languages, by leveraging multilingual data augmentation and contrastive learning techniques on a large scale. These models learn universal speech representations that can be transferred across languages and domains, thus representing a significant advancement in speech representation learning.
- Transformers have been increasingly popular in the development of frameworks for learning representations from multi-modal data, such as speech, images, and text. Among these frameworks, Data2vec [23] is a self-supervised training approach that aims to learn joint representations to capture cross-modal correlations and transfer knowledge across modalities. It has outperformed other unsupervised methods for learning multi-modal representations in benchmark datasets. However, for tasks that require domain-specific models, such as speech recognition or speaker identification, domain-specific models may be more effective, particularly when dealing with data in specific domains or languages. The self-supervised training approach of Data2vec enables cost-effective and scalable learning of representations without requiring labeled data, making it a promising framework for various multi-modal learning applications.

- The field of speech recognition has undergone a revolutionary change with the advent of the Whisper model [432]. This innovative solution has proven to be highly versatile, providing exceptional accuracy for various speech-related tasks, even in challenging environments. The Whisper model achieves its outstanding performance through a minimalist approach to data pre-processing and weak supervision, which allows it to deliver state-of-the-art results in speech processing. The model is capable of performing multilingual speech recognition, translation, and language identification, thanks to its training on a diverse audio dataset. Its multitasking model can cater to various speech-related tasks, such as transcription, voice assistants, education, entertainment, and accessibility. One of the unique features of Whisper is its minimalist approach to data pre-processing, which eliminates the need for significant standardization and simplifies the speech recognition pipeline. The resulting models generalize well to standard benchmarks and deliver competitive performance without fine-tuning, demonstrating the potential of advanced machine learning techniques in speech processing.
- Text-to-speech synthesis has been a topic of interest for many years, and recent advancements have led to the development of new models such as VALL-E [540]. VALL-E is a novel text-to-speech synthesis model that has gained significant attention due to its unique approach to the task. Unlike traditional TTS systems, VALL-E treats the task as a conditional language modeling problem and leverages a large amount of semi-supervised data to train a generalized TTS system. It can generate high-quality personalized speech with a 3-second acoustic prompt from an unseen speaker and provides diverse outputs with the same input text. VALL-E also preserves the acoustic environment and the speaker's emotions about the acoustic prompt, without requiring additional structure engineering, pre-designed acoustic features, or fine-tuning. Furthermore, VALL-E X [636] is an extension of VALL-E that enables cross-lingual speech synthesis, representing a significant advancement in TTS technology.

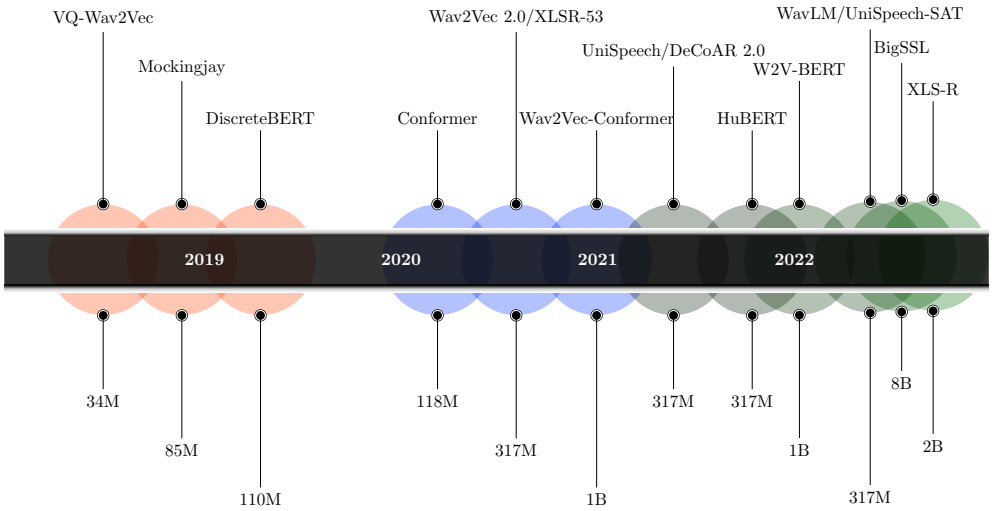

Fig. 4. Timeline for large Transformer Models for Speech

The timeline highlights the development of large Transformer-based models for speech processing is shown in Figure 4. The size of the models has grown exponentially, with significant breakthroughs

achieved in speech recognition, synthesis, and translation. These large models have set new performance benchmarks in the field of speech processing, but also pose significant computational and data requirements for training and inference.

3.4 Conformer

The Transformer architecture, which utilizes a self-attention mechanism, has successfully replaced recurrent operations in previous architectures. Over the past few years, various Transformer variants have been proposed [157]. Architectures combining Transformers and CNNs have recently shown promising results on speech-processing tasks [559]. To efficiently model both local and global dependencies of an audio sequence, several attempts have been made to combine CNNs and Transformers. One such architecture proposed by the authors is the Conformer [157], a convolution-augmented transformer for speech recognition. Conformer outperforms RNNs, previous Transformers, and CNN-based models, achieving state-of-the-art performance in speech recognition. The Conformer model consists of several building blocks, including convolutional layers, self-attention layers, and feedforward layers. The architecture of the Conformer model can be summarized as follows:

- Input Layer: The Conformer model inputs a sequence of audio features, such as MFCCs or Mel spectrograms.
- Convolutional Layers: Local features are extracted from the audio signal by processing the input sequence through convolutional layers.
- Self-Attention Layers: The Conformer model incorporates self-attention layers following the convolutional layers. Self-attention is a mechanism that enables the model to focus on various sections of the input sequence while making predictions. This is especially advantageous for speech recognition because it facilitates capturing long-term dependencies in the audio signal.
- Feedforward Layers: After the self-attention layers, the Conformer model applies a sequence of feedforward layers intended to process the output of the self-attention layers further and ready it for the ultimate prediction.
- Output Layer: Finally, the output from the feedforward layers undergoes a softmax activation function to generate the final prediction, typically representing a sequence of character labels or phonemes.

The conformer model has emerged as a promising neural network architecture for various speech-related research tasks, including but not limited to speech recognition, speaker recognition, and language identification. In a recent study by Gulati et al. [157], the conformer model was demonstrated to outperform previous state-of-the-art models, particularly in speech recognition significantly. This highlights the potential of the conformer model as a key tool for advancing speech-related research.

3.4.1 Application

The Conformer model stands out among other speech recognition models due to its ability to efficiently model both local and global dependencies of an audio sequence. This is crucial for speech recognition, language translation, and audio classification [1, 2, 157]. The model achieves this through self-attention and convolution modules, combining the strengths of CNNs and Transformers. While CNNs capture local information in audio sequences, the self-attention mechanism captures global dependencies [2]. The Conformer model has achieved remarkable performance in speech recognition tasks, setting benchmarks on datasets such as LibriSpeech and AISHELL-1.

Despite these successes, speech synthesis and recognition challenges persist, including difficulties generating natural-sounding speech in non-English languages and real-time speech generation. To address these limitations, Wang et al. [635] proposed a novel approach that combines noisy student training with SpecAugment and large Conformer models pre-trained on the Libri-Light dataset using the wav2vec 2.0 pre-training method. This approach achieved state-of-the-art word error rates on the LibriSpeech dataset. Recently, Wang et al. [552] developed Conformer-LHUC, an extension of the Conformer model that employs learning hidden unit contribution (LHUC) for speaker adaptation. Conformer-LHUC has demonstrated exceptional performance in elderly speech recognition and shows promise for the clinical diagnosis and treatment of Alzheimer's disease.

Several enhancements have been made to the Conformer-based model to address high word error rates without a language model, as documented in [329]. Wu [575] proposed a deep sparse Conformer to improve its long-sequence representation capabilities. Furthermore, Burchi and Timofte [48] have recently enhanced the noise robustness of the Efficient Conformer architecture by processing both audio and visual modalities. In addition, models based on Conformer, such as Transducers [246], have been adopted for real-time speech recognition [403] due to their ability to process audio data much more quickly than conventional recurrent neural network (RNN) models.

3.5 Sequence to Sequence Models

The sequence-to-sequence (seq2seq) model in speech processing is popularly used for ASR, ST, and TTS tasks. The general architecture of the seq2seq model involves an encoder-decoder network that learns to map an input sequence to an output sequence of varying lengths. In the case of ASR, the input sequence is the speech signal, which is processed by the encoder network to produce a fixed-length feature vector representation of the input signal. The decoder network inputs this feature vector and produces the corresponding text sequence. This can be achieved through a stack of RNNs [424], Transformer [112] or Conformer [157] in the encoder and decoder networks.

The sequence-to-sequence model has emerged as a potent tool in speech translation. It can train end-to-end to efficiently map speech spectrograms in one language to their corresponding spectrograms in another. The notable advantage of this approach is eliminating the need for an intermediate text representation, resulting in improved efficiency. Additionally, the Seq2seq models have been successfully implemented in speech generation tasks, where they reverse the ASR approach. In such applications, the input text sequence serves as the input, with the encoder network creating a feature vector representation of the input text. The decoder network then leverages this representation to generate the desired speech signal.

Karita et al. [238] conducted an extensive study comparing the performance of transformer and traditional RNN models on 15 different benchmarks for Automatic Speech Recognition (ASR), including a multilingual ASR benchmark, a Speech Translation (ST) benchmark, and two Text-to-Speech (TTS) benchmarks. In addition, they proposed a shared Sequence-to-Sequence (S2S) architecture for AST, TTS, and ST tasks, which is depicted in Figure 5.

- Encoder

$$\begin{aligned} X_0 &= \text{Encoder-PreNet}(X), \\ X_e &= \text{Encoder-Main}(X_0) \end{aligned} \tag{22}$$

  where X is the sequence of speech features (e.g. Mel spectrogram) for AST and ST and phoneme or character sequence for TTS.

- Decoder

$$\begin{aligned} Y_0[1:t-1] &= \text{Decoder-PreNet}(Y[1:t-1]), \\ Y_d[t] &= \text{Decoder-Main}(X_e, Y_0[1:t-1]), \\ Y_{post}[1:t] &= \text{Decoder-PostNet}(Y_d[1:t]), \end{aligned} \tag{23}$$

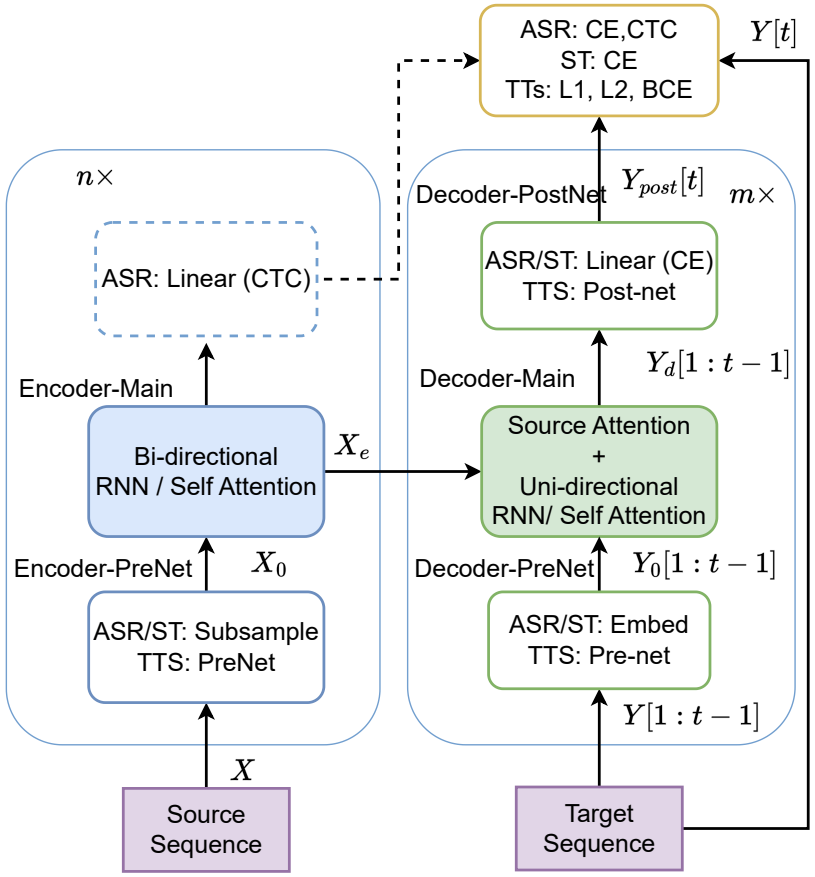

Fig. 5. Unified formulation for Sequence-to-Sequence architecture in speech applications [238]. X and Y are source and target sequences respectively.

During the training stage, input to the decoder is ground truth target sequence $Y[1:t-1]$. The Decoder-Main module is utilized to produce a subsequent target frame. This is accomplished by utilizing the encoded sequence X_e and the prefix of the target prefix $Y_0[1:t-1]$. The decoder is mostly unidirectional for sequence generation and often uses an attention mechanism [27] to produce the output.

Seq2seq models have been widely used in speech processing, initially based on RNNs. However, RNNs face the challenge of processing long sequences, which can lead to the loss of the initial context by the end of the sequence [238]. To overcome this limitation, the transformer architecture has emerged, leveraging self-attention mechanisms to handle sequential data. The transformer has shown remarkable performance in tasks such as ASR, ST, and speech synthesis. As a result, the use of RNN-based seq2seq models has declined in favour of the transformer-based approach.

3.5.1 Application

Seq2seq models have been used for speech processing tasks such as voice conversion [204, 508], speech synthesis [204, 390, 391, 545, 560], and speech recognition. The field of ASR has seen significant progress, with several advanced techniques emerging as popular options. These include

the CTC approach, which has been further developed and improved upon through recent advancements [155], as well as attention-based approaches that have also gained traction [81]. The growing interest in these techniques has increased the use of seq2seq models in the speech community.

- Attention-based Approaches: The attention mechanism is a crucial component of sequence-to-sequence models, allowing them to effectively weigh input acoustic features during decoding [27, 348]. Attention-based Seq2seq models utilize previously generated output tokens and the complete input sequence to factorize the joint probability of the target sequence into individual time steps. The attention mechanism is conditioned on the current decoder states and runs over the encoder output representations to incorporate information from the input sequence into the decoder output. Incorporating attention mechanisms in Seq2Seq models has resulted in an impressive performance in various speech processing tasks, such as speech recognition [381, 424, 518, 568], text-to-speech [392, 476, 597], and voice conversion [204, 508]. These models have demonstrated competitiveness with traditional state-of-the-art approaches. Additionally, attention-based Seq2Seq models have been used for confidence estimation tasks in speech recognition, where confidence scores generated by a speech recognizer can assess transcription quality [305]. Furthermore, these models have been explored for few-shot learning, which has the potential to simplify the training and deployment of speech recognition systems [177].
- Connectionist Temporal Classification: While attention-based methods create a soft alignment between input and target sequences, approaches that utilize CTC loss aim to maximize log conditional likelihood by considering all possible monotonic alignments between them. These CTC-based Seq2Seq models have delivered competitive results across various ASR benchmarks [157, 176, 358, 505] and have been extended to other speech-processing tasks such as voice conversion [332, 625, 632], speech synthesis [625] etc. Recent studies have concentrated on enhancing the performance of Seq2Seq models by combining CTC with attention-based mechanisms, resulting in promising outcomes. This combination remains a subject of active investigation in the speech-processing domain.

3.6 Reinforcement Learning

Reinforcement learning (RL) is a machine learning paradigm that trains an agent to perform discrete actions in an environment and receive rewards or punishments based on its interactions. The agent aims to learn a policy that maximizes its long-term reward. In recent years, RL has become increasingly popular and has been applied to various domains, including robotics, game playing, and natural language processing. RL has been utilized in speech recognition, speaker diarization, and speech enhancement tasks in the speech field. One of the significant benefits of using RL for speech tasks is its ability to learn directly from raw audio data, eliminating the need for hand-engineered features. This can result in better performance compared to traditional methods that rely on feature extraction. By capturing intricate patterns and relationships in the audio data, RL-based speech systems have the potential to enhance accuracy and robustness.

3.6.1 Basic Models

The utilization of deep reinforcement learning (DRL) in speech processing involves the environment (a set of states S), agent, actions (A), and reward (r). The semantics of these components depends on the task at hand. For instance, in ASR tasks, the environment can be composed of speech features, the action can be the choices of phonemes, and the reward could be the correctness of those phonemes given the input. Audio signals are one-dimensional time-series signals that undergo pre-processing and feature extraction procedures. Pre-processing steps include noise suppression, silence removal, and channel equalization, improving audio signal quality and creating robust and

efficient audio-based systems. Previous research has demonstrated that pre-processing improves the performance of deep learning-based audio systems [281].

Feature extraction is typically performed after pre-processing to convert the audio signal into meaningful and informative features while reducing their number. MFCCs and spectrograms are popular feature extraction choices in speech-based systems [281]. These features are then given to the DRL agent to perform various tasks depending on the application. For instance, consider the scenario where a human speaks to a DRL-trained machine, where the machine must act based on features derived from audio signals.

- *Value-based DRL:* Given the state of the environment (s), a value function $Q : S \times A \rightarrow \mathbb{R}$ is learned to estimate overall future reward $Q(s, a)$ should an action a be taken. This value function is parameterized with deep networks like CNN, Transformers, etc.
- *Policy-based DRL:* As opposed to value-based RL, policy-based RL methods learns a policy function $\pi : S \rightarrow A$ that chooses the best possible action (a) based on reward.
- *Model-based DRL:* Unlike the previous two approaches, model-based RL learns the dynamics of the environment in terms of the state transition probabilities, i.e., a function $M : S \times A \times S \rightarrow \mathbb{R}$. Given such a model, policy, or value functions are optimized.

3.6.2 Application

In speech-related research, deep reinforcement learning can be used for several purposes, including:

Speech recognition and Emotion modeling. Deep reinforcement learning (DRL) can be used to train speech recognition systems [84, 85, 225, 440, 513] to transcribe speech accurately. In this case, the system receives an audio input and outputs a text sequence corresponding to the spoken words. The environmental states might be learned from the input audio features. The actions might be the generated phonemes. The reward could be the similarity between the generated and gold phonemes, quantified in edit distance. Several works have also achieved promising results for non-native speech recognition [435]

DRL pre-training has shown promise in reducing training time and enhancing performance in various Human-Computer Interaction (HCI) applications, including speech recognition [440]. Recently, researchers have suggested using a reinforcement learning algorithm to develop a Speech Enhancement (SE) system that effectively improves ASR systems. However, ASR systems are often complicated and composed of non-differentiable units, such as acoustic and language models. Therefore, the ASR system's recognition outcomes should be employed to establish the objective function for optimizing the SE model. Other than ASR, SE, some studies have also focused on SER using DRL algorithms [237, 275, 441]

Speaker identification. Similarly, for speaker identification tasks, the actions can be the speaker's choices, and a binary reward can be the correctness of choice.

Speech synthesis and coding. Likewise, the states can be the input text, the actions can be the generated audio, and the reward could be the similarity between the gold and generated mel-spectrogram.

Deep reinforcement learning has several advantages over traditional machine learning techniques. It can learn from raw data without needing hand-engineered features, making it more flexible and adaptable. It can also learn from feedback, making it more robust and able to handle noisy environments.

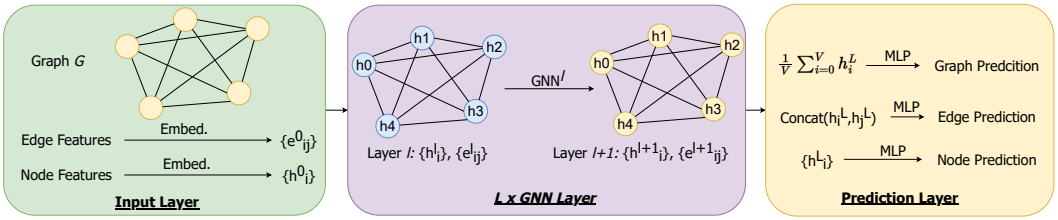

Fig. 6. A standard experimental pipeline for GCNs, which embeds the graph node and embeds the graph node edge features, performs several GNN layers to compute convolutional features, and finally predicts a task-specific MLP layer.

However, deep reinforcement learning also has some challenges that must be addressed. It requires a lot of data to train and can be computationally expensive. It also requires careful selection of the reward function to ensure that the system learns the desired behavior.

3.7 Graph Neural Network

Over the past few years, the field of Graph Neural Networks (GNNs) has witnessed a remarkable expansion as a widely adopted approach for analysing and learning from data on graphs. GNNs have demonstrated their potential in various domains, including computer science, physics, mathematics, chemistry, and biology, by delivering successful outcomes. Furthermore, in recent times, the speech-processing domain has also witnessed the growth of GNNs.

3.7.1 Basic Models

Speech processing involves analysing and processing audio signals, and GNNs can be useful in this context when we represent the audio data as a graph. In this answer, we will explain the architecture of GNNs for speech processing. The standard GNN pipeline is shown in Figure 6, according to the application the GNN layer can consist of Graph Convolutional Layers [629], Graph Attention Layers [533], or Graph Transformer [609].

Graph Representation of Speech Data. The first step in using GNNs for speech processing is representing the speech data as a graph. One way to do this is to represent the speech signal as a sequence of frames, each representing a short audio signal segment. We can then represent each frame as a node in the graph, with edges connecting adjacent frames.

Graph Convolutional Layers. Once the speech data is represented as a graph, we can use graph convolutional layers to learn representations of the graph nodes. Graph convolutional layers are similar to traditional ones, but instead of operating on a grid-like structure, they operate on graphs. These layers learn to aggregate information from neighboring nodes to update the features of each node.

Graph Attention Layers. Graph attention layers can be combined with graph convolutional layers to give more importance to certain nodes in the graph. Graph attention layers learn to assign weights to neighbor nodes based on their features, which can help capture important patterns in speech data. Several works have used graph attention layers for neural speech synthesis [331] or speaker verification [221] and diarization [270].

Recurrent Layers. Recurrent layers can be used in GNNs for speech processing to capture temporal dependencies between adjacent frames in the audio signal. Recurrent layers allow the network to maintain an internal state that carries information from previous time steps, which can be useful for modeling the dynamics of speech signals.

Output Layers. The output layer of a GNN for speech processing can be a classification layer that predicts a label for the speech data (e.g., phoneme or word) or a regression layer that predicts a continuous value (e.g., pitch or loudness). The output layer can be a traditional fully connected layer or a graph pooling layer that aggregates information from all the nodes in the graph.

3.7.2 Application

The advantages of using GNNs for speech processing tasks include their ability to represent the dependencies and interrelationships between various entities, which is suitable for speech processing tasks such as speaker diarization [484, 485, 548], speaker verification [222, 479], speech synthesis [331, 501, 502], or speech separation [536, 553], which require the analysis of complex data representations. GNNs retain a state representing information from their neighborhood with arbitrary depth, unlike standard neural networks. GNNs can be used to model the relationship between phonemes and words. GNNs can learn to recognize words in spoken language by treating the phoneme sequence as a graph. GNNs can also be used to model the relationship between different acoustic features, such as pitch, duration, and amplitude, in speech signals, improving speech recognition accuracy.

GNNs have shown promising results in multichannel speech enhancement, where they are used for extracting clean speech from noisy mixtures captured by multiple microphones [521]. The authors of a recent study [383] propose a novel approach to multichannel speech enhancement by combining Graph Convolutional Networks (GCNs) with spatial filtering techniques such as the Minimum Variance Distortionless Response (MVDR) beamformer. The algorithm aims to extract speech and noise from noisy signals by computing the Power Spectral Density (PSD) matrices of the noise and the speech signal of interest and then obtaining optimal weights for the beam former using a frequency-time mask. The proposed method combines the MVDR beam former with a super-Gaussian joint maximum a posteriori (SGJMAP) based SE gain function and a GCN-based separation network. The SGJMAP-based SE gain function is used to enhance the speech signals, while the GCN-based separation network is used to separate the speech from the noise further.

3.8 Diffusion Probabilistic Model

Diffusion probabilistic models, inspired by non-equilibrium thermodynamics [180, 492], have proven to be highly effective for generating high-quality images and audio. These models create a Markov chain of diffusion steps ($x_t \sim q(x_t|x_{t-1})$) from the original data (x_0) to the latent variable $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ by gradually adding pre-scheduled noise to the data. The reverse diffusion process then reconstructs the desired data samples (x_0) from the noise x_T, as shown in Figure 7. Unlike VAE or flow models, diffusion models keep the dimensionality of the latent variables fixed. While mostly used for image and audio synthesis, diffusion models have potential applications in speech-processing tasks, such as speech synthesis and enhancement. This section offers a comprehensive overview of the fundamental principles of diffusion models and explores their potential uses in the speech domain.

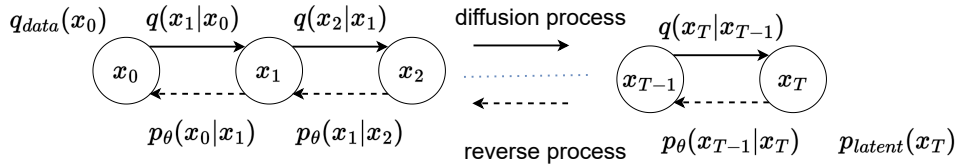

Fig. 7. The Diffusion Probabilistic Model is a generative model that progressively transforms a noise distribution into the target data distribution through a series of diffusion steps, where the noise level decreases as the process continues. The model is trained by maximizing the likelihood of the data distribution and can be used for tasks such as speech synthesis, enhancement, and denoising.

Forward diffusion process. Given a clean speech data $x_0 \sim q_{data}(x_0)$,

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{24}$$

At every time step t, $q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$ where $\{\beta_t \in (0,1)\}_{t=1}^{T}$. As the forward process progresses, the data sample x_0 losses its distinguishable features, and as $T \to \infty$, x_T approaches a standard Gaussian distribution.

Reverse diffusion process. The reverse diffusion process is defined by a Markov chain from $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to x_0 and parameterized by θ:

$$p_\theta(x_0, ..., x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{25}$$

where $x_T \sim \mathcal{N}(0, I)$ and the transition probability $p_\theta(x_{t-1}|x_t)$ is learnt through noise-estimation. This process eliminates the Gaussian noise added in the forward diffusion process.

3.8.1 Application

Diffusion models have emerged as a leading approach for generating high-quality speech in recent years [65, 198, 212, 262, 421, 422]. These non-autoregressive models transform white noise signals into structured waveforms via a Markov chain with a fixed number of steps. One such model, FastDiff, has achieved impressive results in high-quality speech synthesis [198]. By leveraging a stack of time-aware diffusion processes, FastDiff can generate high-quality speech samples 58 times faster than real-time on a V100 GPU, making it practical for speech synthesis deployment for the first time. It also outperforms other competing methods in end-to-end text-to-speech synthesis. Another powerful diffusion probabilistic model proposed for audio synthesis is DiffWave [262]. It is non-autoregressive and generates high-fidelity audio for different waveform generation tasks, such as neural vocoding conditioned on mel spectrogram, class-conditional generation, and unconditional generation. DiffWave delivers speech quality on par with the strong WaveNet vocoder [394] while synthesizing audio much faster.

Diffusion models have shown great promise in speech processing, particularly in speech enhancement [340, 341, 430, 472]. Recent advances in diffusion probabilistic models have led to the development of a new speech enhancement algorithm that incorporates the characteristics of the noisy speech signal into the diffusion and reverses processes [342]. This new algorithm is a generalized form of the probabilistic diffusion model, known as the conditional diffusion probabilistic model. During its reverse process, it can adapt to non-Gaussian real noises in the estimated speech signal. In addition, Qiu et al. [430] propose SRTNet, a novel method for speech enhancement that

uses the diffusion model as a module for stochastic refinement. The proposed method comprises a joint network of deterministic and stochastic modules, forming the "enhance-and-refine" paradigm. The paper also includes a theoretical demonstration of the proposed method's feasibility and presents experimental results to support its effectiveness.

4 Speech Representation Learning

The process of speech representation learning is essential for extracting pertinent and practical characteristics from speech signals, which can be utilized for various downstream tasks such as speaker identification, speech recognition, and emotion recognition. While traditional methods for engineering features have been extensively used, recent advancements in deep-learning-based techniques utilizing supervised or unsupervised learning have shown remarkable potential in this field. Nonetheless, a novel approach founded on self-supervised representation learning has surfaced, aiming to unveil the inherent structure of speech data and acquire representations that capture the underlying structure of the data. This approach surpasses traditional feature engineering methods and can significantly increase the accuracy to a considerable extent and effectiveness of downstream tasks. The primary objective of this new paradigm is to uncover informative and meaningful features from speech signals and outperform existing approaches. Therefore, this approach is considered a promising direction for future research in speech representation learning.

This section provides a comprehensive overview of the evolution of speech representation learning with neural networks. We will examine various techniques and architectures developed over the years, including the emergence of unsupervised representation learning methods like autoencoders, generative adversarial networks (GANs), and self-supervised representation learning frameworks. We will also examine the difficulties and constraints associated with these techniques, such as data scarcity, domain adaptation, and the interpretability of learned representations. Through a comprehensive analysis of the advantages and limitations of different representation learning approaches, we aim to provide insights into how to harness their power to improve the accuracy and robustness of speech processing systems.

4.1 Supervised Learning

In supervised representation learning, the model is trained using annotated datasets to learn a mapping between input data and output labels. The set of parameters that define the mapping function is optimized during training to minimize the difference between the predicted and true output labels in the training data. The goal of supervised representation learning is to enable the model to learn a useful representation or features of the input data that can be used to accurately predict the output label for new, unseen data. For instance, supervised representation learning in speech processing using CNNs learn speech features from spectrograms. CNNs can identify patterns in spectrograms relevant to speech recognition, such as those corresponding to different phonemes or words. Unlike CNNs, which typically require spectrogram input, RNNs can directly take in the raw speech signals as input and learn to extract features or representations that are relevant for speech recognition or other speech-processing tasks. Learning speaker representations typically involves minimizing a loss function. Chung et al. [87] compares their effectiveness for speaker recognition tasks, we distill it in Table 1 to present an overview of commonly used loss functions. Additionally, a new angular variant of the prototypical loss is introduced in their work. Results from extensive experimental validation on the VoxCeleb1 test set indicate that the GE2E and prototypical networks outperform other models in terms of performance.

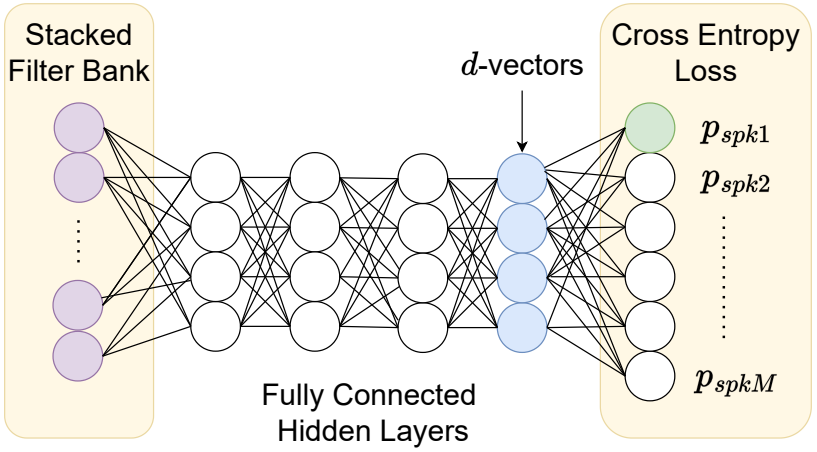

Fig. 8. *d*-vector model architecture.

4.1.1 Deep speaker representations

Speaker representation is a critical aspect of speech processing, allowing machines to analyze and process various parts of a speaker's voice, including pitch, intonation, accent, and speaking style. In recent years, deep neural networks (DNNs) have shown great promise in learning robust features for speaker recognition. This section reviews deep learning-based techniques for speaker representation learning that have demonstrated significant improvements over traditional methods.

These deep speaker representations can be applied to a range of speaker-recognition tasks beyond verification and identification, including diarization [280, 549, 614], voice conversion [82, 316, 571], multi-speaker TTS [408, 461, 584], speaker adaptation [80] etc. To provide a comprehensive overview, we analyzed deep embeddings from the perspectives of input raw [220, 443] or mel-spectogram [491], network architecture [104, 318], temporal pooling strategies [376], and loss functions [87, 489, 546]. In the following subsection, we introduce two representative deep embeddings: *d*-vector [530] and *x*-vector [490, 491]. These embeddings have been widely adopted recently and have demonstrated state-of-the-art performance in various speaker-recognition tasks. By understanding the strengths and weaknesses of different deep learning-based techniques for speaker-representation learning, we can better leverage their power to improve the accuracy and robustness of speaker-recognition systems.

- ***d*-vector** [530] is a frame-level speaker embedding technique as shown in Figure 8. It operates by assigning the speaker's true identity of a training utterance as the label for each frame belonging to that utterance during the training phase. This transforms the training process into a classification task, enabling a maxout DNN to classify the frames of a training utterance based on the speaker's identity. The DNN employs softmax as the output layer to minimize the cross-entropy loss between the ground-truth frame labels and the network output. During the testing phase, the *d*-vector extracts the output activation of each frame from the last hidden layer of the DNN as the deep embedding feature of the frame. The *d*-vector then computes the average of all frames' deep embedding features in an utterance to produce a compact representation known as the *d*-vector. The hypothesis behind the *d*-vector is that the compact representation space developed from a development set can generalize well to unseen speakers during the testing phase [530].

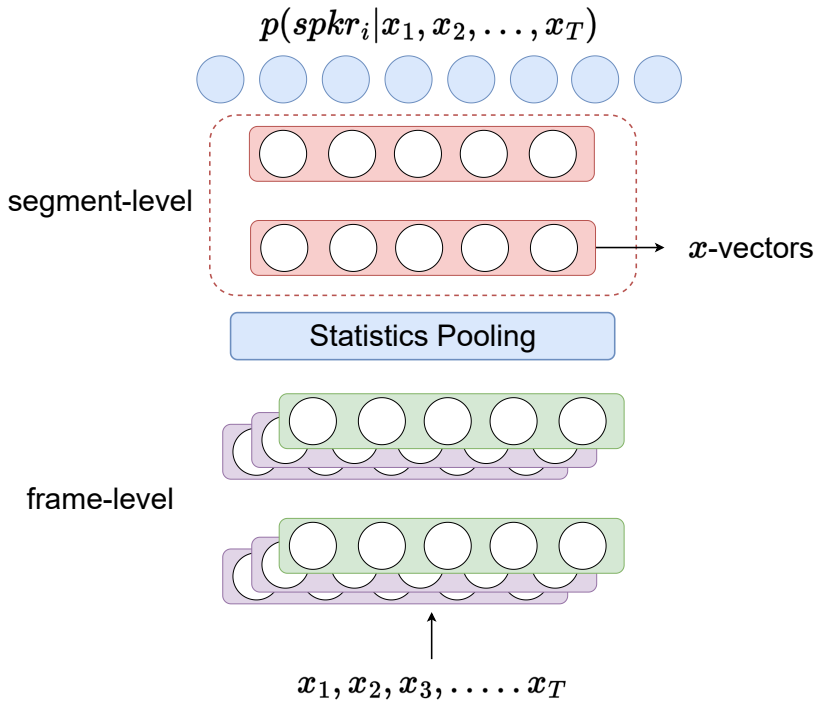

Fig. 9. x-vector model architecture.

- ***x*-vector** [490, 491] is a segment-level speaker embedding and an advancement over the d-vector method as it incorporates additional modeling of temporal information and phonetic information in speech signals, resulting in improved performance compared to the d-vector. x-vector employs an aggregation process to move from frame-by-frame speaker labeling to utterance-level speaker labeling as highlighted in Figure 9. The network structure of the -vector is depicted in a figure, which consists of time-delay layers for extracting frame-level speech embeddings, a statistical pooling layer for concatenating mean and standard deviation of embeddings as a segment-level feature, and a standard feedforward network for classifying the segment-level feature to its speaker. x-vector is the segment-level speaker embedding generated from the feedforward network's second-to-last hidden layer. The authors in [457, 593] have also discovered the significance of data augmentation in enhancing the performance of the x-vector.

4.2 Unsupervised learning

Unsupervised representation learning for speech processing has gained significant emphasis over the past few years. Similar to visual modality in CV and text modality in NLP, speech i.e. audio modality introduces unique challenges. Unsupervised speech representation learning is concerned with learning useful speech representations without using annotated data. Usually, the model is first pre-trained on the task where plenty of data is available. The model is then fined tuned or used to extract input representations for a small model, specifically targeting tasks with limited data.

One approach to addressing the unique challenges of unsupervised speech representation learning is to use probabilistic latent variable models (PLVM), which assume an unknown generative

Table 1. The table summarizes various loss functions used in training the speaker recognition models including their formulation [87].

Loss Function	Objective Type	Description
Softmax	Classification	$L_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp W_{y_i}^T x_i + b_{y_i}}{\sum_{j=1}^{C}\exp W_j^T x_i + b_{y_i}}$
AM-Softmax (CosFace) [546]	Classification	$L_C = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp s(\cos\theta_{y_i,i} - m)}{\exp s(\cos\theta_{y_i,i} - m) + \sum_{j\neq y_i}\exp s(\cos\theta_{j,i})}$
AAM-Softmax (ArcFace) [99]	Classification	$L_A = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp s(\cos\theta_{y_i,i} - m)}{\exp s(\cos\theta_{y_i,i} + m) + \sum_{j\neq y_i}\exp s(\cos\theta_{j,i})}$
Triplet [469]	Metric learning [617]	$L_T = \frac{1}{N}\sum_{j=1}^{N}\max(0, \|\|x_{j,0} - x_{j,1}\|\|_2^2, \|\|x_{j,0} - x_{k\neq j,1}\|\|_2^2 + m)$
Prototypical [489]	Metric learning [489]	$L_P = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{\exp \mathbf{S}_{j,j}}{\sum_{k=1}^{N}\exp \mathbf{S}_{j,k}}$
Generalized end-to-end (GE2E) [539]	Metric learning [550]	$L_G = -\frac{1}{N}\sum_{j,i}\log\frac{\exp \mathbf{S}_{j,i,j}}{\sum_{k=1}^{N}\exp \mathbf{S}_{j,i,k}}$
Angular Prototypical	Metric learning	$L_{AP} = -\frac{1}{N}\sum_{j,i}\log\frac{\exp \mathbf{S}_{j,i,j}}{\sum_{k=1}^{N}\exp \mathbf{S}_{j,i,k}}$

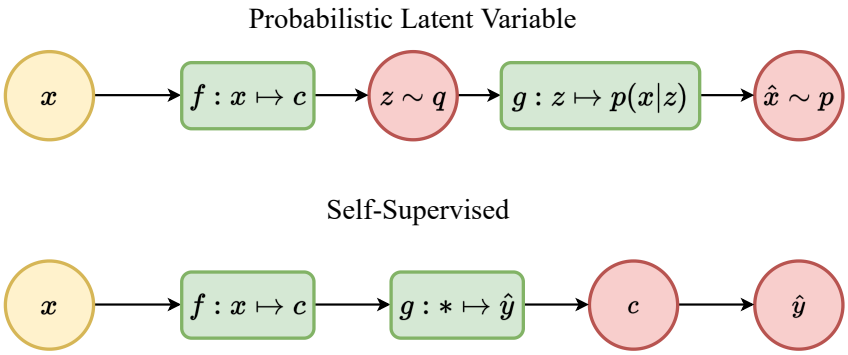

Fig. 10. Overview of difference between probabilistic latent variable models and self-supervised learning. In latent variable models learn the functions $f(.)$ and $g(.)$ learn the parameters of distribution p and q. The latent variable z is used for representing learning.

process produces the data and enables the learning of rich structural representations and reasoning about observed and unobserved factors of variation in complex datasets such as speech within a probabilistic framework. PLVM specified a joint distribution $p(x, z)$ over unobserved stochastic latent variable z and observed variables x. By factorizing the joint distribution into modular components, it becomes possible to learn rich structural representations and reason about observed and unobserved factors of variation in complex datasets such as speech within a probabilistic framework. The likelihood of a PLVM given a data x can be written as

$$p(x) = \int p(x|z)p(z)dz. \tag{26}$$

Probabilistic latent variable models provide a powerful way to learn a representation that captures the underlying relationships between observed and unobserved variables, without requiring explicit supervision or labels. These models involve unobserved latent variables that must be inferred from the observed data, typically using probabilistic inference techniques such as Markov Chain Monte Carlo (MCMC) methods. In the context of representation learning, Variational autoencoders (VAE) are commonly used with latent variable models for various speech processing tasks, leveraging the power of probabilistic modeling to capture complex patterns in speech data.

4.3 Semi-supervised Learning

Semi-supervised learning can be viewed as a process of optimizing a model using both labeled and unlabeled data. The set of labeled data points, denoted by X_L, contains N_L items, where each item is represented as (x_i, y_i) with y_i being the label of x_i. On the other hand, the set of unlabeled data points, denoted by X_U, consists of N_U items, represented as $x_{N_L+1}, x_{N_L+2}, ..., x_{N_L+N_U}$.

In semi-supervised learning, the objective is to train a model f_θ with parameters θ that can minimize the expected loss over the entire dataset. The loss function $L(y, f_\theta(x))$ is used to quantify the deviation between the model's prediction $f_\theta(x)$ and the ground truth label y. The expected loss can be mathematically expressed as:

$$L(y, f_\theta(x)) = E_{(x,y)\sim p_{data}(x,y)}[L(y, f_\theta(x))] \tag{27}$$

where $p_{data}(x, y)$ is the underlying data distribution.In semi-supervised learning, the loss function is typically decomposed into two parts: a supervised loss term that is only defined on the labeled data, and an unsupervised loss term that is defined on both labeled and unlabelled data. The supervised loss term is calculated as follows:

$$\mathcal{L}_{sup} = \frac{1}{N_L} \sum_{(x,y)\in X_L} L(y, f_\theta(x)) \tag{28}$$

The unsupervised loss term leverages the unlabelled data to encourage the model to learn meaningful representations that capture the underlying structure of the data. One common approach is to use a regularization term that encourages the model to produce similar outputs for similar input data. This can be achieved by minimizing the distance between the output of the model for two similar input data points. One such regularization term is the entropy minimization term, which can be expressed as:

$$\mathcal{L}_{unsup} = \frac{1}{N_U} \sum_{(x_i)\in X_U} \sum_{j=1}^{|y|} p_\theta(y_j, x_i) \log p_\theta(y_j, x_i) \tag{29}$$

where $p_\theta(y_j|x_i)$ is the predicted probability of the j-th label for the unlabelled data point x_i. Finally the overall objective function for semi-supervised learning can be expressed as $\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{unsup}$, α is a hyperparameter that controls the weight of the unsupervised loss term. The goal is to find the optimal parameters θ that minimize this objective function. Semi-supervised learning involves learning a model from both labelled and unlabelled data by minimizing a combination of supervised and unsupervised loss terms. By leveraging the additional unlabelled data, semi-supervised learning can improve the generalization and performance of the model in downstream tasks.

Semi-supervised learning techniques are increasingly being employed to enhance the performance of DNNs across a range of downstream tasks in speech processing, including ASR, TTS, etc. The primary objective of such approaches is to leverage large unlabelled datasets to augment the performance of supervised tasks that rely on labelled datasets. The recent advancements in speech recognition have led to a growing interest in the integration of semi-supervised learning methods to improve the performance of ASR and TTS systems [33, 85, 223, 582, 634, 635]. This approach is particularly beneficial in scenarios where labelled data is scarce or expensive to acquire. In fact, for many languages around the globe, labelled data for training ASR models are often inadequate, making it challenging to achieve optimal results. Thus, using a semi-supervised learning model trained on abundant resource data can offer a viable solution that can be readily extended to low-resource languages.

Semi-supervised learning has emerged as a valuable tool for addressing the challenges of insufficient annotations and poor generalization [160]. Research in various domains, including image quality assessment [334], has demonstrated that leveraging both labelled and unlabelled data through semi-supervised learning can lead to improved performance and generalization. In the domain of speech quality assessment, several studies [473] have exploited the generalization capabilities of semi-supervised learning to enhance performance.

Moreover, semi-supervised learning has gained significant attention in other areas of speech processing, such as end-to-end speech translation [418]. By leveraging large amounts of unlabelled data, semi-supervised learning approaches have demonstrated promising results in improving the performance and robustness of speech translation models. This highlights the potential of semi-supervised learning to address the limitations of traditional supervised learning approaches in a variety of speech processing tasks.

4.4 Self-supervised representation learning (SSRL)

Self-supervised representation learning (SSRL) is a machine learning approach that focuses on achieving robust and in-depth feature learning while minimizing reliance on extensively annotated datasets, thus reducing the annotation bottleneck [128, 282]. SSRL comprises various techniques that allow models to be trained without needing human-annotated labels [128, 282]. One of the key advantages of SSRL is its ability to operate on unlabelled datasets, which reduces the need for large annotated datasets [128, 282]. In recent years, self-supervised learning has progressed rapidly, with some methods approaching or surpassing the efficacy of fully supervised learning methods. Self-supervised learning methods typically involve pretext tasks that generate pseudo labels for discriminative model training without actual labeling. The difference between self-supervised representation learning and unsupervised representation is highlighted in Figure 10. In contrast to unsupervised representation learning, SSRL techniques are designed to generate these pseudo labels for model training. The ability of SSRL to achieve robust and in-depth feature learning without relying heavily on annotated datasets holds great promise for the continued development of machine learning techniques.

SSRL differs from supervised learning mainly in terms of its data requirements. While supervised learning relies on labeled data, where the model learns from input-output pairs, SSL generates its own labels from the input data, eliminating the need for labeled data [282]. The SSL approach trains the model to predict a portion of the input data, which is then utilized as a label for the task at hand [282]. Although SSRL is an unsupervised learning technique, it seeks to tackle tasks commonly associated with supervised learning without relying on labeled data [282].

4.4.1 Generative Models

This method involves instructing a model to produce samples resembling the input data without explicitly learning the labels, creating valuable representations applicable to other tasks. The detailed architecture for generative models with three different variants is shown in Figure 11. The earliest self-supervised method, predicting masked inputs using surrounding data, originated from the text field in 2013 with word2vec. The continuous bag of words (CBOW) concept of word2vec predicts a central word based on its neighbors, resembling ELMo and BERT's masked language modeling (MLM). These non-autoregressive generative approaches differ in their use of advanced structures, such as bidirectional LSTM (for ELMo) and transformer (for BERT), with recent models producing contextual embeddings. In the context of the speech, Mockingjay [323] applied masking to all feature dimensions in the speech domain, whereas TERA [322] applied masking only to a particular subset of feature dimensions.

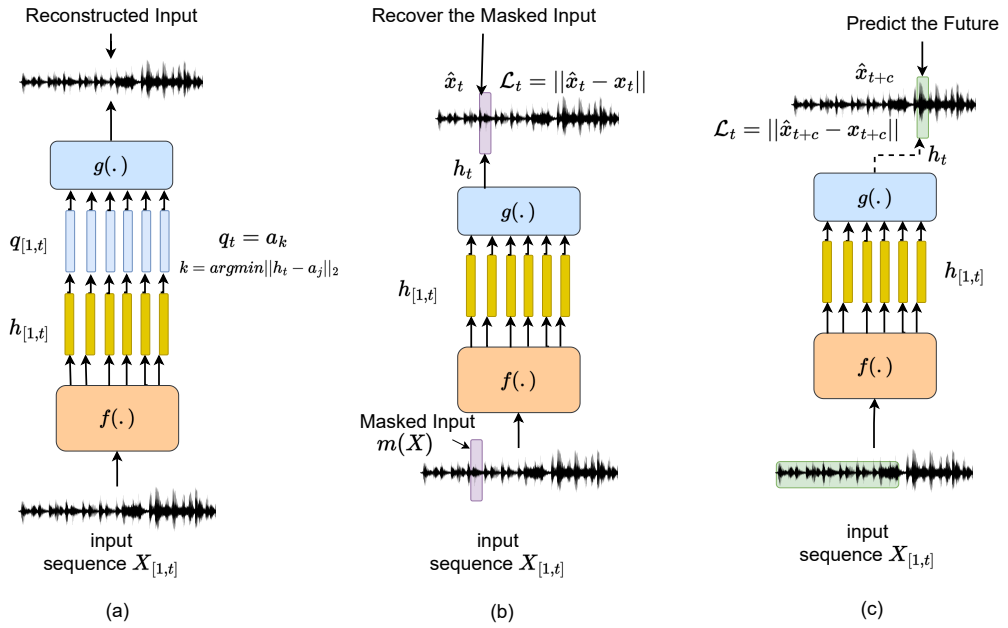

Fig. 11. Generative approaches to self-supervised learning.

- Auto-encoding Models: Autoencoders (AEs) and Variational Autoencoders (VAEs) are popular generative models in self-supervised learning. AEs consist of an encoder and a decoder, with the task of reconstructing input while discarding low-level details to learn significant features. VAEs are a probabilistic version of AEs, widely used in speech modeling. The vector-quantized variational autoencoder (VQ-VAE) [529] is another generative model extending VAEs by parameterizing discrete latent representations' posterior distribution. VQ-VAE has been successful in generative spoken language modeling, and combining discrete latent space with self-supervised learning improves its results further.
- Autoregressive models: Autoregressive generative self-supervised learning uses autoregressive prediction coding technique [91] to model the probability distribution of a sequence of data points. This approach aims to predict the next data point in a sequence based on the previous data points. Autoregressive models typically use RNNs or a transformer architecture as a basic model.
  The authors in paper [394] introduce a generative model for raw audio called WaveNet, based on PixelCNN [528]. To enhance the model's ability to handle long-range temporal dependencies, the authors incorporate dilated causal convolutions [394]. They also utilize Gated Residual blocks and skip connections to improve the model's expressivity.
- Masked Reconstruction: The concept of masked reconstruction is influenced by the masked language model (MLM) task proposed in BERT [105]. This task involves masking specific tokens in input sentences with learned masking tokens or other input tokens, and training the model to reconstruct these masked tokens from the non-masked ones. Recent research has explored similar pretext tasks for speech representation learning that help models develop contextualized representations capturing information from the entire input, like

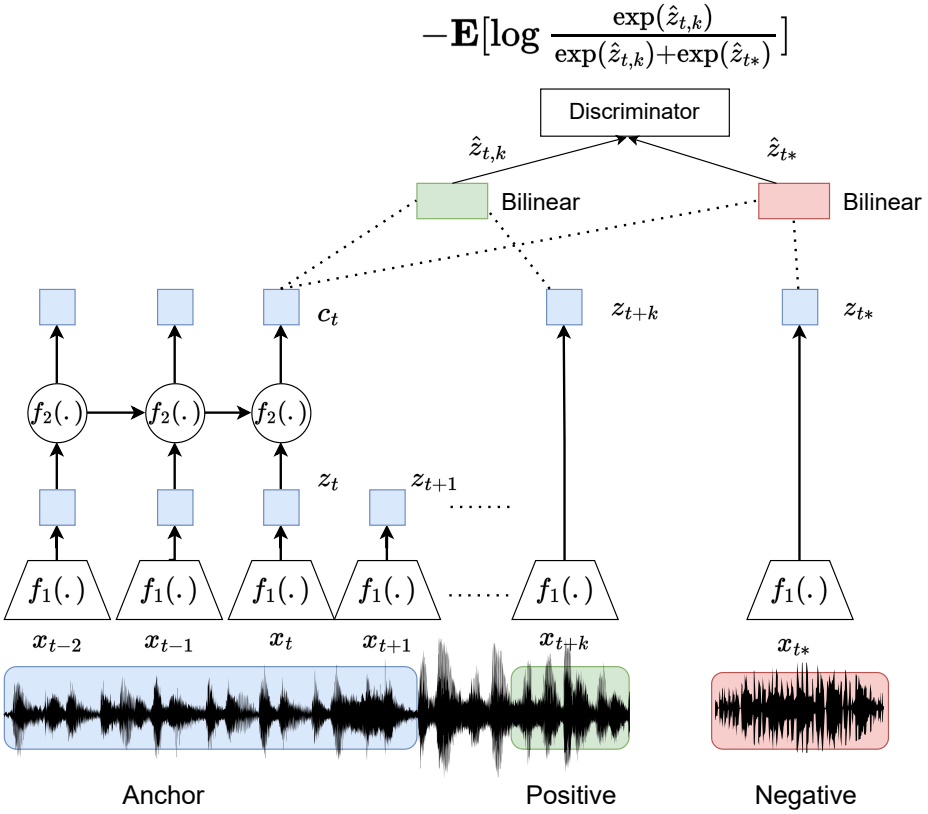

Fig. 12. Contrastive Self-supervised learning: Contrastive Predictive Coding.

the DeCoAR model [319]. This approach assists the model in comprehending input data better, leading to more precise and informative representations.

4.4.2 Contrastive Models

The technique involves training a model to differentiate between similar and dissimilar pairs of data samples, which helps the model acquire valuable representations that can be utilized for various tasks, as shown on Figure 12. The fundamental principle of contrastive learning is to generate positive and negative pairs of training samples based on the comprehension of the data. The model must learn a function that assigns high similarity scores to two positive samples and low similarity scores to two negative samples. Therefore, generating appropriate samples is crucial for ensuring that the model comprehends the fundamental features and structures of the data.

- Wav2Vec 2.0 [25] is a framework for self-supervised learning of speech representations that is one of the current state-of-the-art models for ASR [25]. The training of the model occurs in two stages. Initially, the model operates in a self-supervised mode during the first phase, where it uses unlabelled data and aims to achieve the best speech representation possible. The second phase is fine-tuning a particular dataset for a specific purpose. Wav2Vec 2.0

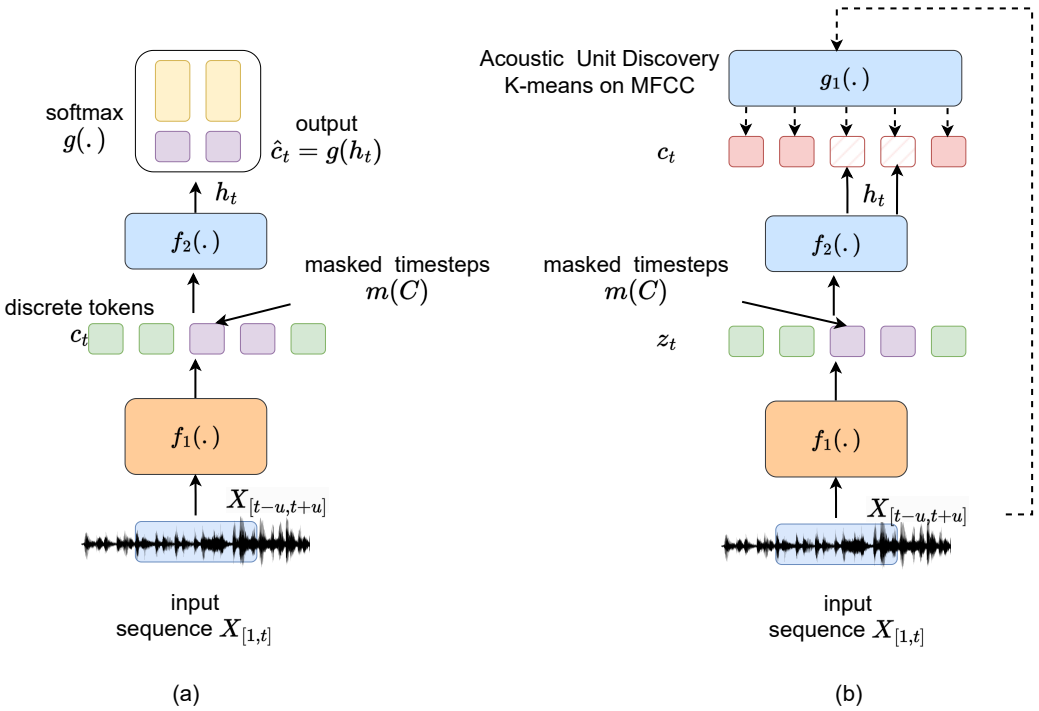

Fig. 13. Predictive Self-supervised learning: (a) Discrete BERT (b) HuBERT.

takes advantage of self-supervised training and uses convolutional layers to extract features from raw audio.

In the speech field, researchers have explored different approaches to avoid overfitting, including augmentation techniques like Speech SimCLR [214] and the use of positive and negative pairs through methods like Contrastive Predictive Coding (CPC) (Ooster and Meyer [395]), Wav2vec (v1, v2.0) (Schneider et al. [468]), VQ-wav2vec (Baevski et al. [24]), and Discrete BERT [22]." "In the graph field, researchers have developed approaches like Deep Graph Infomax (DGI) (Velickovic et al., 2019 [534]) to learn representations that maximize the mutual information between local patches and global structures while minimizing mutual information between patches of corrupted graphs and the original graph's global representation.

4.4.3 Predictive Models

The primary concept involves creating simpler objectives or targets to minimize the need for data generation. However, the most critical and difficult aspect is ensuring that the task's difficulty level is appropriate for the model to learn effectively.

SSRL has been leveraged in ASR through transformer-based models to acquire meaningful representations [22, 187, 322]. The rise of Self-Supervised Representation Learning (SSRL) has been transformative in exploiting the growing abundance of data [145].

- Directly applying BERT-type training to speech input is difficult due to its unsegmented and unstructured nature. A new model, Discrete BERT [22], has been proposed to solve this challenge. The model converts the continuous speech input into a sequence of discrete codes that can be used for code representation learning. The discrete units are extracted

from a pre-trained vq-wav2vec [24] model and utilized as inputs and targets in a standard BERT model, which includes a softmax normalized output layer. The detail architecture of Discrete BERT is shown in Figure 13 (a). The model is trained using categorical cross-entropy loss and a masked view of the original speech input to predict code representations. The Discrete BERT model was the first to demonstrate the efficacy of self-supervised speech representation learning, achieving a WER of 25% on the standard test-other subset with only a 10-minute fine-tuning set. This approach solves the challenge of applying BERT-type training directly to continuous speech input and exhibits the potential to enhance speech recognition accuracy significantly.

- The HuBERT [187] and TERA [322] models are two self-supervised approaches for speech representation learning. HuBERT uses an offline clustering step to align target labels with a BERT-like prediction loss, with the prediction loss applied only over the masked regions as outlined in Figure 13 (b). This encourages the model to learn a combined acoustic and language model over the continuous inputs. On the other hand, TERA is a self-supervised speech pre-training method that reconstructs acoustic frames from their altered counterparts using a stochastic policy to alter along various dimensions, including time, frequency, and tasks. These alterations help extract feature-based speech representations that can be fine-tuned as part of downstream models.

Microsoft has introduced UniSpeech-SAT [69] and WavLM [68] models, which follow the HuBERT framework. These models have been designed to enhance speaker representation and improve various downstream tasks. The key focus of these models is data augmentation during the pre-training stage, resulting in superior performance. WavLM model has exhibited outstanding effectiveness in diverse downstream tasks, such as automatic speech recognition, phoneme recognition, speaker identification, and emotion recognition. It is worth highlighting that this model currently holds the top position on the SUPERB leaderboard [592], which evaluates speech representations' performance in terms of reusability

Self-supervised learning has emerged as a widely adopted and effective technique for speech processing tasks due to its ability to train models with large amounts of unlabeled data. A comprehensive overview of self-supervised approaches, evaluation metrics, and training data is provided in Table 2 for speech recognition, speaker recognition, and speech enhancement. Researchers and practitioners can use this resource to select appropriate self-supervised methods and datasets to enhance their speech-processing systems. As self-supervised learning techniques continue to advance and refine, we can expect significant progress and advancements in speech processing.

5 Speech Processing Tasks

In recent times, the field of speech processing has gained significant attention due to its rapid evolution and its crucial role in modern technological applications. This field involves the use of diverse techniques and algorithms to analyse and understand spoken language, ranging from basic speech recognition to more complex tasks such as spoken language understanding and speaker identification. Since speech is one of the most natural forms of communication, speech processing has become a critical component of many applications such as virtual assistants, call centers, and speech-to-text transcription. In this section, we provide a comprehensive overview of the various speech-processing tasks and the techniques used to achieve them, while also discussing the current challenges and limitations faced in this field and its potential for future development.

The assessment of speech-processing models depends greatly on the caliber of datasets employed. By utilizing standardized datasets, researchers are enabled to objectively gauge the efficacy of varying approaches and identify scopes for advancement. The selection of evaluation metrics plays

Table 2. Summary of different self-supervised approaches and proposed models for speech processing with associated metrics and training Data. **ASR**: Automatic Speech Recognition, **PR**: Phoneme Recognition. **PC**: Phoneme Classification, **SR**: Speaker Recognition

SSL Approaches	Model Name	Task (Metric)	Pre-Training Dataset (hours)	Dataset	
				Training	Test
Contrastive Models	CPC	PC	LS (100h)	LS (100h)	LS (100h)
		SR	LS (100h)	LS (100h)	LS (100h)
	Modified CPC	PC	LS (100h,360h) Zerospeech2017 (45h)	CV-Dataset	CV-Dataset
	Bidirectional CPC	ASR	WSJ (80h) LS (960h) TIMIT (5h) SSA (1h) TED3 (440h) SwithBoard (310h)	WSJ (80h) LS (960h) TIMIT (5h) SSA (1h) TED3 (440h) SwithBoard (310h)	WSJ (test92, test93) LS (test-clean, test-other) TED3 (dev, test) SwithBoard (eval2000)
		ASR-Multi	Audio Set (2500h) AVSpeech (3100h) CV-Dataset (430h)	Audio Set (2500h) AVSpeech (3100h) CV-Dataset (430h)	OpenSLR ALFFA
	wav2vec	ASR	LS 80/860h LS 960h + WSJ (si284)	WSJ (si284)	WSJ (eval92)
		PR	TIMIT	TIMIT	TIMIT
	wav2vec 2.0	ASR	LS (960h) LL (60000h)	LS (960h)	LS (test-clean) LS (test-other)
		PR	LS (960h) LL (60000h)	TIMIT	TIMIT
	vq-wav2vec 2.0	ASR	LS (960h)	WJS (si284)	WJS (eval92)
		PR	LS (960h)	TIMIT	TIMIT
	wav2vec-C	ASR	Alexa-10k	Alexa-eval	Alexa-eval
	w2v-BERT	ASR	LL (60000h)	LS (960h)	LS (test) LS (test-other) LS (dev) LS (dev-other)
	Speech SimCLR	ASR	LS (960h) WJS (si284) TED2	WJS (si284)	WJS (si284)
		PR	LS (960h) WJS (si284) TED2	TIMIT	TIMIT
	UnSpeech	ASR-Multi	LL (60000h) GigaSpeech (10000h) VP (24000h)	SUPERB	SUPERB
Predictive Models	data2vec	ASR	LS (960h)	LS (10m, 1h, 100h, 960h)	LS (test-other)
	BEST-RQ	ASR	LL (60000h)	LS (960h)	LS (test) LS (test-other) LS (dev) LS (dev-other)
		ASR-Multi	XLS-R -U (429000h)	LS (MLS-10h) LS (MLS-Full)	LS (MLS-10h) LS (MLS-Full)
	HuBERT	ASR	LS (960h) LL (60000h)	LS (960h)	LS (test-clean) LS (test-other)
	WavLM	ASR-Multi	LL (60000h) GigaSpeech (10000h) VP (24000h)	SUPERB	SUPERB
	Discrete BERT	ASR	LS (960h)	LS (100h)	LS (test-clean) LS (test-other)

a critical role in this process, hinging on the task at hand and the desired outcome. Therefore, it is essential that researchers conduct a meticulous appraisal of different metrics to make informed decisions. This paper offers a thorough summary of frequently utilized datasets and metrics across diverse downstream tasks, as presented in Table 3 and, Table 4.

5.1 Automatic speech recognition (ASR) /conversational multi-speaker AST

5.1.1 Task Description

Automatic speech recognition (ASR) technology enables machines to convert spoken language into text or commands, serving as a cornerstone of human-machine communication and facilitating a wide range of applications such as speech-to-speech translation and information retrieval [338]. ASR involves multiple intricate steps, starting with the extraction and analysis of acoustic features, including spectral and prosodic features, which are then employed to recognize spoken words. Next, an acoustic model matches the extracted features to phonetic units, while a language model predicts the most probable sequence of words based on the recognized phonetic units. Ultimately,

Table 3. Comparative analysis of speech processing datasets: This table summarizes the essential features of different speech-processing datasets, including their typical applications in various speech-processing tasks. **ASR**: Automatic Speech Recognition, **PR**: Phoneme Recognition. **PC**: Phoneme Classification, **SR**: Speaker Recognition, **SV**: Speaker Verification, **SER**: Speech Emotion Recognition, **IC**: Intent Classification, **TTS**: Text-to-Speech, **VC**: Voice Conversion, **ST**: Speech Translation, **SS**: Speech Separation

Dataset	Language	Lenght (hours)	ASR	PR	PC	SR	SV	SER	IC	TTS	VC	ST	SS
TIMIT Acoustic-Phonetic Continuous Speech Corpus	English	5.4	✓	✓	✓								
Lip Reading Sentences 2 (LRS2)	English		✓										
LibriSpeech (LS)	English	1000	✓	✓	✓	✓							
GigaSpeech	English	10000	✓										
Fleurs	Multilingual	12	✓										
LibriTTS	English	585								✓	✓		
L2ARCTIC	English	11.2								✓	✓		
CMUARCTIC	English	20								✓	✓		
Wall Street Journal (WSJ)	English		✓	✓	✓								
VoxPopuli (VP)	Multilingual	1800	✓										
BABEL (BBL)	Multilingual		✓										
Common Voice (CV-dataset)	Multilingual	9283	✓	✓	✓								
CSTR VCTK	English												
HUB 5	English	2000	✓										
CHiME-5	English	50.12											
TED-LIUM 3 (TED 3)	English	452											
TED-LIUM 2 (TED 2)	English	118											
AISHELL-1	Mandarin	520	✓										
AISHELL-3	Mandarin	85								✓	✓		
AISHELL-4	Mandarin	120								✓	✓		
Arabic Speech Corpus	Arabic	3.7	✓	✓	✓								
Persian Consonant Vowel Combination	Persian	-	✓	✓	✓								
ALFFA	Multilingual	5.2-18.3											
OpenSLR-multi	Multilingual	4.4-265.9											
VCTK	English	44											
VoxCeleb1/2	English					✓	✓						
Fluent Speech Commands (FSC)	English	14.7							✓				
Emotional Speech Dataset (ESD)	English	29						✓					
Interactive Emotional Dyadic Motion Capture (IEMOCAP)	English	12						✓					
Multimodal EmotionLines Dataset (MELD)	English							✓					
LibraSeepch En-Fr	English/French											✓	
CoVoST-2	Multilingual	2880										✓	

the acoustic and language model outcomes are merged to produce the transcription of spoken words. Deep learning techniques have gained popularity in recent years, allowing for improved accuracy in ASR systems [25, 432]. This paper provides an overview of the key components involved in ASR and highlights the role of deep learning techniques in enhancing the technology's accuracy.

Most speech recognition systems that use deep learning aim to simplify the processing pipeline by training a single model to directly map speech signals to their corresponding text transcriptions. Unlike traditional ASR systems that require multiple components to extract and model features, such as HMMs and GMMs, end-to-end models do not rely on hand-designed components [18, 298]. Instead, end-to-end ASR systems use DNNs to learn acoustic and linguistic representations

Table 4. Comprehensive Evaluation Metrics for Speech Processing Tasks: This table provides a comprehensive overview of the evaluation metrics used to assess the performance of speech-based systems across various tasks such as ASR, speaker verification, and TTS. The table highlights the specific metrics employed for each task along with the score range and commonly used datasets

Tasks	Metric	Description	Score range	Evaluation dataset
Automatic speech recognition	WER	Word Error Rate	0-1	TIMIT
	CER	Character Error Rate	0-1	LibriSpeech
Phoneme recognition	Accuracy	Classification accuracy	0-1	TIMIT
Phoneme classification	F1-score	Harmonic mean of precision and recall	0-1	TIMIT
Speaker recognition	EER	Equal Error Rate	0-1	VoxCeleb1
Speaker verification	FAR/FRR	False Acceptance Rate / False Rejection Rate	0-1	VoxCeleb1
Speech emotion recognition	Accuracy	Classification accuracy	0-1	IEMOCAP,ESD
Intent classification	F1-score	Harmonic mean of precision and recall	0-1	ATIS,SNIPS
Text-to-speech	MOS	Mean Opinion Score	1-5	LJSpeech, LibriTTS
Voice conversion	MOS	Mean Opinion Score	1-5	VCC 2016
Speech translation	BLEU	Bilingual Evaluation Understudy	0-1	MuST-C
Speech separation	SI-SDRi	Signal to Distortion Ratio	-20-30	WSJ0-2mix
Speech enhancement	PESQ	Perceptual Evaluation of Speech Quality	-0.5-4.5	NOIZEUS
Voice activity detection	F1-score	Harmonic mean of precision and recall	0-1	QUT-NOISE

directly from the input speech signals [298]. One popular type of end-to-end model is the encoder-decoder model with attention. This model uses an encoder network to map input audio signals to hidden representations, and a decoder network to generate text transcriptions from the hidden representations. During the decoding process, the attention mechanism enables the decoder to selectively focus on different parts of the input signal [298].

End-to-end ASR models can be trained using various techniques such as CTC [239], which is used to train models without explicit alignment between the input and output sequences, and RNNs, which are commonly used to model temporal dependencies in sequential data such as speech signals. Transfer learning-based approaches can also improve end-to-end ASR performance by leveraging pre-trained models or features [102, 320, 474]. While end-to-end ASR models have shown promising results in various applications, there is still room for improvement to achieve human-level performance [102, 133, 230, 231, 320, 602]. Nonetheless, deep learning-based end-to-end ASR architecture offers a promising and efficient approach to speech recognition that can simplify the processing pipeline and improve recognition accuracy.

5.1.2 Dataset

The development and evaluation of ASR systems are heavily dependent on the availability of large datasets. As a result, ASR is an active area of research, with numerous datasets used for this purpose. In this context, several popular datasets have gained prominence for use in ASR systems.

- Common Voice: Mozilla's Common Voice project [16] is dedicated to producing an accessible, unrestricted collection of human speech for the purpose of training speech recognition systems. This ever-expanding dataset features contributions from more than 9, 000 speakers spanning 60 different languages.

- LibriSpeech: LibriSpeech [401] is a corpus of approximately 1,000 hours of read English speech created from audiobooks in the public domain. It is widely used for speech recognition research and is notable for its high audio quality and clean transcription.
- VoxCeleb: VoxCeleb [88] is a large-scale dataset containing over 1 million short audio clips of celebrities speaking, which can be used for speech recognition and recognition research. It includes a diverse range of speakers from different backgrounds and professions.
- TIMIT: The TIMIT corpus [148] is a widely used speech dataset consisting of recordings consisting of 630 speakers representing eight major dialects of American English, each reading ten phonetically rich sentences. It has been used as a benchmark for speech recognition research since its creation in 1986.
- CHiME-5: The CHiME-5 dataset [32] is a collection of recordings made in a domestic environment to simulate a real-world speech recognition scenario. It includes 6.5 hours of audio from multiple microphone arrays and is designed to test the performance of ASR systems in noisy and reverberant environments.

Other notable datasets include Google's Speech Commands Dataset [566], the Wall Street Journal dataset[3], and TED-LIUM [455].

5.1.3 Models

The use of RNN-based architecture in speech recognition has many advantages over traditional acoustic models. One of the most significant benefits is their ability to capture long-term temporal dependencies [238] in speech data, enabling them to model the dynamic nature of speech signals. Additionally, RNNs can effectively process variable-length audio sequences, which is essential in speech recognition tasks where the duration of spoken words and phrases can vary widely. RNN-based models can efficiently identify and segment phonemes, detect and transcribe spoken words, and can be trained end-to-end, eliminating the need for intermediate steps. These features make RNN-based models particularly useful in real-time applications, such as speech recognition in mobile devices or smart homes [113, 172], where low latency and high accuracy are crucial.

In the past, RNNs were the go-to model for ASR. However, their limited ability to handle long-range dependencies prompted the adoption of the Transformer architecture. For example, in 2019, Google's Speech-to-Text API transitioned to a Transformer-based architecture that surpassed the previous RNN-based model, especially in noisy environments and for longer sentences, as reported in [628]. Additionally, Facebook AI Research introduced wav2vec 2.0, a self-supervised learning approach that leverages a Transformer-based architecture to perform unsupervised speech recognition. wav2vec 2.0 has significantly outperformed the previous RNN-based model and achieved state-of-the-art results on several benchmark datasets.

Transformer for the ASR task is first proposed in [112], where authors include CNN layers before submitting preprocessed speech features to the input. By incorporating more CNN layers, it becomes feasible to diminish the gap between the sizes of the input and output sequences, given that the number of frames in audio exceeds the number of tokens in text. This results in a favorable impact on the training process. The change in the original architecture is minimal, and the model achieves a competitive word error rate (WER) of 10.9% on the Wall Street Journal (WSK) speech recognition dataset (Table 5). Despite its numerous advantages, Transformers in its pristine state has several issues when applied to ASR. RNN, with its overall training speed (i.e., convergence) and better WER because of effective joint training and decoding methods, is still the best option.

The authors in [112] propose the Speech Transformer, which has the advantage of faster iteration time, but slower convergence compared to RNN-based ASR. However, integrating the Speech

[3]https://www.ldc.upenn.edu/

Table 5. Table summarizing the performance of different ASR models in terms of WER% on five different datasets (LibriSpeech test, LibriSpeech clean, TIMIT, Common Voice, WSJ eval92, and GigaSpeech) also highlighting the use of extra data during training. ZS stands for Zero-Shot Performance.

Model	Architecture	Extra Training Data	WER% ↓	WER% ↓	Model	Architecture	Extra Training Data	WER% ↓
LibriSpeech test			clean	others	TIMIT			
Conformer + Wav2vec 2.0 [635]	Conformer + wav2vec2.0	Y	1.4	2.6	wav2vec 2.0 [25]	Transformer + CNN	Y	8.3
w2v-BERT XXL [92]	CNN+Transformer	Y	1.4	2.5	vq-wav2vec [24]	Transformer + CNN	Y	11.6
SpeechStew (1B)[55]	Conformer	Y	1.7	3.3	LSTM + Monophone Reg [444]	LSTM	N	14.5
SpeechStew (100M) [55]	Conformer	N	2.0	4.0	Common Voice			
ContextNet + SpecAugment [404]	LSTM+CNN	Y	1.7	3.4	SpeechStew (1B) [55]	Conformer	N	10.8
Conformer (L) [157]	Conformer	N	1.9	4.1	Whisper [432]		N	9.5
ContextNet [164]	Conformer + wav2vec2.0	N	1.9	3.4	WSJ eval92			
Squeezeformer [249]	Conformer	N	2.47	5.97	SpeechStew (100M) [55]	Conformer	N	1.3
LSTM Transducer [613]	LSTM	N	2.23	5.6	tdnn+chain [423]	TDNN	N	2.32
Transformer Transducer [324]	Transformer	N	2.0	4.2	GigaSpeech			
Whisper [432]		N	2.7 (ZS)	5.6 (ZS)	Conformer/Transformer-AED [58]	Conformer	N	10.80

Transformer with the naive language model (LM) is challenging. To address this issue, various improvements in the Speech Transformer architecture have been proposed in recent years. For example, [239] suggests incorporating the Connectionist Temporal Classification (CTC) loss into the Speech Transformer. CTC is a popular technique used in speech recognition to align input and output sequences of varying lengths and one-to-many or many-to-one mappings. It introduces a blank symbol representing gaps between output symbols and computes the loss function by summing probabilities across all possible paths. The loss function encourages the model to assign high probabilities to correct output symbols and low probabilities to incorrect output symbols and the blank symbol, allowing the model to predict sequences of varying lengths. The CTC loss is commonly used with RNNs such as LSTM and GRU, which are well-suited for sequential data. CTC loss is a powerful tool for training neural networks to perform sequence-to-sequence tasks where the input and output sequences have varying lengths and mappings between them are not one-to-one.

Various other improvements have also been proposed to enhance the performance of Speech Transformer architecture and integrate it with the naive language model, as the use of the transformer directly for ASR has not been effective in exploiting the correlation among the speech frames. The sequence order of speech, which the recurrent processing of input features can represent, is an important distinction. The degradation in performance for long sentences is reported using absolute positional embedding (AED) [81]. The problems associated with long sequences can become more acute for transformer [649]. To address this issue, a transition was made from absolute positional encoding to relative positional embeddings [649]. Whereas authors in [516] replace positional embeddings with pooling layers. In a considerably different approach, the authors in [375] propose a novel way of combining positional embedding with speech features by replacing positional encoding with trainable convolution layers. This update further improves the stability of optimization for large-scale learning of transformer networks. The above works confirmed the superiority of their techniques against sinusoidal positional encoding.

In 2016, Baidu introduced a hybrid ASR model called Deep Speech 2 [12] that uses both RNNs and Transformers. The model also uses CNNs to extract features from the audio signal, followed by a stack of RNNs to model the temporal dependencies and a Transformer-based decoder to generate the output sequence. This approach achieved state-of-the-art results on several benchmark datasets such as LibriSpeech, VoxForge, WSJeval92 etc. The transition of ASR models from RNNs

Table 6. When comparing the performance of wav2vec2.0 large and whisper on different datasets, it is evident that the zero-shot Whisper model outperforms wav2vec2.0 large on several datasets. A detailed analysis reveals significant differences in their respective performances.

Dataset	wav2vec2.0 Large	Whisper Large
Common Voice	29.9	**9.0**
Fleurs En	14.6	**4.4**
Tedlium	10.5	**4.0**
CHiME6	65.8	**25.5**
VoxPopuli En	17.9	**7.3**
Switchboard	28.3	**13.8**
CallHome	34.8	**17.6**
LibriSpeech Clean	2.7	**2.7**
LibriSpeech Other	6.2	**5.2**

to Transformers has significantly improved performance, especially for long sentences and noisy environments.

The Transformer architecture has been widely adopted by different companies and research groups for their ASR models, and it is expected that more organizations will follow this trend in the upcoming years. One of the advanced speech models that leverage this architecture is the Universal Speech Model (USM) [633] developed by Google, which has been trained on over 12 million hours of speech and 28 billion sentences of text in more than 300 languages. With its 2 billion parameters, USM can recognize speech in both common languages like English and Mandarin and less-common languages. Other popular acoustic models for speech recognition include Quartznet [266], Citrinet [358], and Conformer [157]. These models can be chosen and switched based on the specific use case and performance requirements of the speech recognition pipeline. For example, Conformer-based acoustic models are preferred for addressing robust ASR, as shown in a recent study. Another study found that Conformer-1[4] is more effective in handling real-world data and can produce up to 43% fewer errors on noisy data than other popular ASR models. Additionally, fine-tuning pre-trained models such as BERT [105] and GPT [433] has been explored for ASR tasks, leading to state-of-the-art performance on benchmark datasets like LibriSpeech (refer to Table 5). An open-source toolkit called Vosk[5] provides pre-trained models for multiple languages optimized for real-time and efficient performance, making it suitable for applications that require such performance.

The field of speech recognition has made significant progress by adopting unsupervised pre-training techniques, such as those utilized by Wav2Vec 2.0 [25]. Another recent advancement in automatic speech recognition (ASR) is the whisper model, which has achieved human-level accuracy when transcribing the LibriSpeech dataset. These two cutting-edge frameworks, Wav2Vec 2.0 and whisper, currently represent the state-of-the-art in ASR. The whisper model is trained on an extensive supervised dataset, including over 680,000 hours of audio data collected from the web,

[4]https://www.assemblyai.com/blog/conformer-1/
[5]https://alphacephei.com/vosk/lm

which has made it more resilient to various accents, background noise, and technical jargon. The whisper model is also capable of transcribing and translating audio in multiple languages, making it a versatile tool. OpenAI has released inference models and code, laying the groundwork for the development of practical applications based on the whisper model.

In contrast to its predecessor, Wav2Vec 2.0 is a self-supervised learning framework that trains models on unlabeled audio data before fine-tuning them on specific datasets. It uses a contrastive predictive coding (CPC) loss function to learn speech representations directly from raw audio data, requiring less labeled data. The model's performance has been impressive, achieving state-of-the-art results on several ASR benchmarks. These advances in unsupervised pre-training techniques and the development of novel ASR frameworks like Whisper and Wav2Vec 2.0 have greatly improved the field of speech recognition, paving the way for new real-world applications. In summary, the Table 6 highlights the varying effectiveness of wav2vec2.0 large and whisper models across different datasets.

5.2 Neural Speech Synthesis

5.2.1 Task Description

Neural speech synthesis is a technology that utilizes artificial intelligence and deep learning techniques to create speech from text or other inputs. Its applications are widespread, including in healthcare, where it can be used to develop assistive technologies for those who are unable to communicate due to neurological impairments. To generate speech, deep neural networks like CNNs, RNNs, transformers, and diffusion models are trained using phonemes and the mel spectrum. The process involves several components, such as text analysis, acoustic models, and vocoders, as shown in Figure 14. Acoustic models convert linguistic features into acoustic features, which are then used by the vocoder to synthesize the final speech signal. Various architectures, including neural vocoders based on GANs like HiFi-GAN [261], are used by the vocoder to generate speech. Neural speech synthesis also enables manipulation of voice, pitch, and speed of speech signals using frameworks such as Fastspeech2 [447] and NANSY/NANSY++ [78, 79]. These frameworks use information bottleneck to disentangle analysis features for controllable synthesis. The research in neural speech synthesis can be classified into two prominent approaches: autoregressive and non-autoregressive models. Autoregressive models generate speech one element at a time, sequentially, while non-autoregressive models generate all the elements simultaneously, in parallel. Table 7 outlines the different architecture proposed under each category.

5.2.2 Datasets

The field of neural speech synthesis is rapidly advancing and relies heavily on high-quality datasets for effective training and evaluation of models. One of the most frequently utilized datasets in this field is the LJ Speech [211], which features about 24 hours of recorded speech from a single female speaker reading passages from the public domain LJ Speech Corpus. This dataset is free and has corresponding transcripts, making it an excellent choice for text-to-speech synthesis tasks. Moreover, it has been used as a benchmark for numerous neural speech synthesis models, including Tacotron [560], WaveNet [394], and DeepVoice [17, 151].

Apart from the LJ Speech dataset, several other datasets are widely used in neural speech synthesis research. The CMU Arctic [260] and L2 Arctic [638] datasets contain recordings of English speakers with diverse accents reading passages designed to capture various phonetic and prosodic aspects of speech. The LibriSpeech [401], VoxCeleb [88], TIMIT Acoustic-Phonetic Continuous Speech Corpus [148], and Common Voice Dataset [16] are other valuable datasets that offer ample opportunities for training and evaluating text-to-speech synthesis models.

Table 7. Exploring the Landscape of TTS and Vocoder Architectures: Autoregressive and Non Autoregressive Models.

Method	Text-To-Speech	Vocoder
Autoregressive Model	Tacotron [560] ,Tacotron2 [476], Deep Voice 1,2,3 Transformer-TTS [302], DurIAN [604], Flowtron [527] RobuTrans [303], DeviceTTS [205],Wave-Tacotron [567] Apple TTS [8]	WaveNet [394], WaveRNN [226], WaveGAN [410] LPCNet [525], GAN-TTS [38], MultiBand-WaveRNN [604] ImporvedLPCNet [524], Bunched LPCNet2 [405]
Non-Autoregressive Model	ParaNet [411], FastSpeech [449], JDI-T [310], EATS [111] FastSpeech2 [447], FastPitch [277], Glow-TTS [244] Flow-TTS [369], SpeedySpeech [523] Parallel Tacotron [122], BVAE-TTS [289] Parallel Tacotron2 [122], Grad-TTS [421], VITS [245] RAD-TTS [478], WaveGrad2 [66], DelightfulTTS [335] PortaSpeech [448], DiffGAN-TTS [333], JETS [311] WavThruVec [487], FastDiff [198], CLONE [336]	Parallel-WaveNet [393], WaveGlow [425], Parallel-WaveGAN [585] MelGAN [268], MultiBand-MelGAN [589], VocGAN [591], WaveGrad [65] DiffWave [262], HiFi-GAN [261], StyleMelGAN [378], Fre-GAN [248] iSTFTNet [235], Avocodo [31]

5.2.3 Models

Neural network-based text-to-speech (TTS) systems have been proposed using neural networks as the basis for speech synthesis, particularly with the emergence of deep learning. In Statistical Parametric Speech Synthesis (SPSS), early neural models replaced HMMs for acoustic modeling. The first modern neural TTS model, WaveNet [394], generated waveforms directly from linguistic features. Other models, such as DeepVoice 1/2 [17, 151], used neural network-based models to follow the three components of statistical parametric synthesis. End-to-end models, including Tacotron 1 & 2 [476, 560], Deep Voice 3, and FastSpeech 1 & 2 [447, 449], simplified text analysis modules and utilized mel-spectrograms to simplify acoustic features with character/phoneme sequences as input. Fully end-to-end TTS systems, such as ClariNet [415], FastSpeech 2 [447], and EATS [110], are capable of directly generating waveforms from text inputs. Compared to concatenative synthesis [6] and statistical parametric synthesis, neural network-based speech synthesis offers several advantages including superior voice quality, naturalness, intelligibility, and reduced reliance on human preprocessing and feature development. Therefore, end-to-end TTS systems represent a promising direction for advancing the field of speech synthesis.

Transformer models have become increasingly popular for generating mel-spectrograms in TTS systems [302, 447]. These models are preferred over RNN structures in end-to-end TTS systems because they improve training and inference efficiency [302, 449]. In a study conducted by Li et al. [302], a multi-head attention mechanism replaced both RNN structures and the vanilla attention mechanism in Tacotron 2 [476]. This approach addressed the long-distance dependency problem and improved pluralization. Phoneme sequences were used as input to generate the mel-spectrogram, and speech samples were synthesized using WaveNet as a vocoder. Results showed that the transformer-based TTS approach was 4.25 times faster than Tacotron 2 and achieved similar MOS (Mean Opinion Score) performance.

Aside from the work mentioned above, there are other studies that are based on the Tacotron architecture. For example, Skerry-Ryan et al. [488] and Wang et al. [561] proposed Tacotron-based models for prosody control. These models use a separate encoder to compute style information from reference audio that is not provided in the text. Another noteworthy work is the Global-style-Token (GST) [561] which improves on style embeddings by adding an attention layer to capture a wider range of acoustic styles.

[6]https://en.wikipedia.org/wiki/Concatenative_synthesis

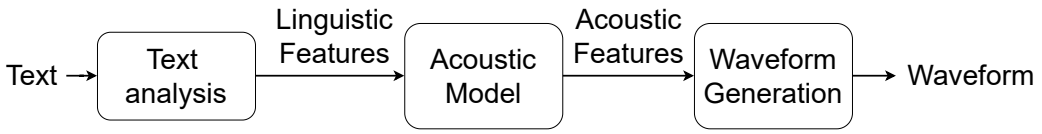

Fig. 14. Neural Text-to-speech (TTS) pipeline: a diagram showing the main modules of a typical TTS system. The system takes text input and processes it through various stages to generate speech output. The text analysis module tokenizes the input text and generates linguistic features such as phonemes and prosody. The acoustic model module then converts these linguistic features into acoustic features, such as mel spectrograms, using a neural network. Finally, the waveform generation module synthesizes the speech waveform from the acoustic features using another neural network.

The FastSpeech [449] algorithm aims to improve the inference speed of TTS systems. To achieve this, it utilizes a feedforward network based on 1D convolution and the self-attention mechanism in transformers to generate Mel-spectrograms in parallel. Additionally, it solves the issue of sequence length mismatch between the Mel-spectrogram sequence and its corresponding phoneme sequence by employing a length regulator based on a duration predictor. The FastSpeech model was evaluated on the LJSpeech dataset and demonstrated significantly faster Mel-spectrogram generation than the autoregressive transformer model while maintaining comparable performance. FastPitch builds on FastSpeech by conditioning the TTS model on fundamental frequency or pitch contour, which improves convergence and eliminates the need for knowledge distillation of Mel-spectrogram targets in FastSpeech.

FastSpeech 2 [447] is another transformer-based TTS system that addresses the issues in FastSpeech and better handles the one-to-many mapping problem in TTS. It leverages more diverse speech information, such as energy, pitch, and more accurate duration, as conditional inputs and trains the system directly on a ground-truth target. Fastspeech 2s is a simplified variant proposed in [61], which eliminates the need for Mel-spectrograms as intermediate output and directly generates speech from the text in inference. Experiments on the LJSpeech dataset showed that FastSpeech 2 and FastSpeech 2s present a simplified training pipeline with fast, robust, and controllable speech synthesis compared to FastSpeech.

Furthermore, in addition to the transformer-based TTS systems like FastSpeech 2 and FastSpeech 2s, researchers have also been exploring the potential of Variational Autoencoder (VAE) based TTS models [158, 190, 245, 289]. These models can learn a latent representation of speech signals from textual input and may be able to produce high-quality speech with less training data and greater control over the generated speech characteristics. For example, authors in [245] used a conditional variational autoencoder (CVAE) to model the acoustic features of speech and an adversarial loss to improve the naturalness of the generated speech. This approach involved conditioning the CVAE on the linguistic features of the input text and using an adversarial loss to match the distribution of the generated speech to that of natural speech. Results from this method have shown promise in generating speech that exhibits natural prosody and intonation.

WaveGrad [65], and DiffWave [262] are notable works that leveraged diffusion models for raw waveform generation, achieving state-of-the-art performance. GradTTS [421] and DiffTTS [212], on the other hand, utilized diffusion models to generate mel features instead of raw waves. DiffVC [422] tackled the challenging one-shot many-to-many voice conversion problem and developed a stochastic differential equation solver. DiffSinger [327] extended sound generation to singing voice synthesis using a shallow diffusion mechanism. Diffsound [588] proposed a sound generation framework conditioned on the text that employed a discrete diffusion model to replace the autoregressive decoder and overcome the unidirectional bias and accumulated errors.

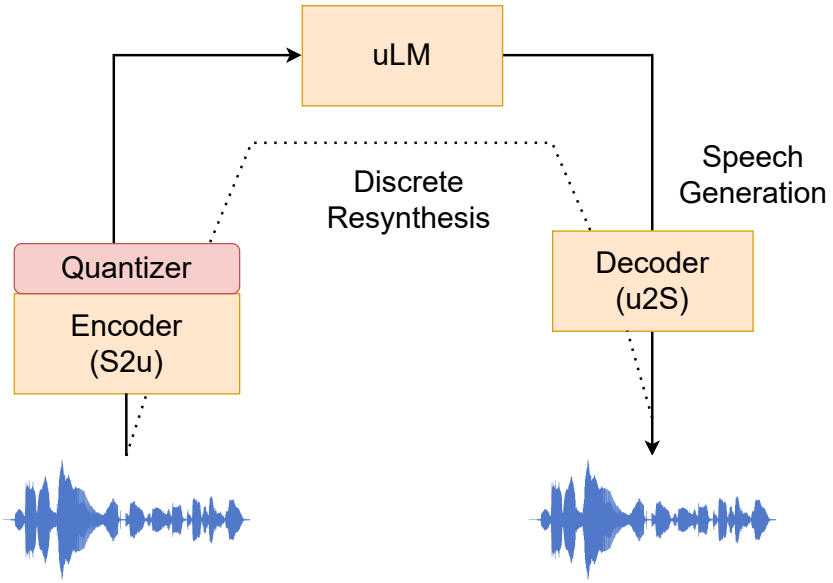

Fig. 15. The architecture of the Generative Spoken Language Model GSLM introduced by Meta in [274]. GSLM model operates through a three-part architecture. Firstly, the encoder takes the speech waveform and transforms it into distinct units represented as S2u. Secondly, the decoder reverses this mapping by converting the units back to the original waveform, represented as u2S. Finally, the language model is unit-based and captures the distribution of unit sequences, which can be viewed as a form of pseudo-text.

EdiTTS [506] is another diffusion-based audio model focusing on the text-to-speech task. During denoising reversal, it induces desired edits through coarse perturbations in the prior space. Guided-TTS [243] and Guided-TTS2 [251] are early text-to-speech models that successfully applied diffusion models in sound generation. [294] combined a voice diffusion model with a spectrogram domain conditioning method to perform text-to-speech with unseen voices during training.

InferGrad [71] improved the diffusion-based text-to-speech model by considering the inference process in training when the number of inference steps is small, enabling fast and high-quality sampling. SpecGrad [257] adapted the time-varying spectral envelope of diffusion noise based on conditioning log-mel spectrogram, incorporating ideas from signal processing. ItoTTS [574] unified text-to-speech and vocoder into a framework based on linear SDE. ProDiff [200] proposed a progressive and fast diffusion model for high-quality text-to-speech. Unlike traditional diffusion models that require hundreds of iterations, ProDiff parameterized the model by predicting clean data and employed a teacher-synthesized mel-spectrogram as a target to reduce data discrepancies and improve prediction sharpness. Finally, Binaural Grad [292] explored the application of diffusion models in binaural audio synthesis, given mono audio, through a two-stage diffusion-based framework.

5.2.4 Alignment

Improving the alignment of text and speech in TTS architecture has been the focus of recent research [21, 28, 34, 61, 219, 244, 309, 368, 370, 421, 448, 475, 478, 623]. Traditional TTS models require external aligners to provide attention alignments of phoneme-to-frame sequences, which can be complex and inefficient. Although autoregressive TTS models use an attention mechanism

to learn these alignments online, these alignments tend to be brittle and often fail to generalize to long utterances and out-of-domain text, resulting in missing or repeating words.

In their study [117], the authors presented a novel text encoder network that includes an additional objective function to explicitly align text and speech encodings. The text encoder architecture is straightforward, consisting of an embedding layer, followed by two bidirectional LSTM layers that maintain the input's resolution. The study utilized the same subword segmentation for the input text as for the ASR output targets. While RNN models with soft attention mechanisms have been proven to be highly effective in various tasks, including speech synthesis, their use in online settings results in quadratic time complexity due to the pass over the entire input sequence for generating each element in the output sequence. In [436], the authors proposed an end-to-end differentiable method for learning monotonic alignments, enabling the computation of attention in linear time. Several enhancements, such as those proposed in [75], have been proposed in recent years to improve alignment in TTS models. Additionally, in [20], the authors introduced a generic alignment learning framework that can be easily extended to various neural TTS models.

The use of normalizing flow has been introduced to address output diversity issues in parallel TTS architectures. This technique is utilized to model the duration of speech, as evidenced by studies conducted in [244, 370, 478]. One such flow-based generative model is Glow-TTS [244], developed specifically for parallel TTS without the need for an external aligner. The model employs the generic Glow architecture previously used in computer vision and vocoder models to produce mel-spectrograms from text inputs, which are then converted to speech audio. Glow-TTS has demonstrated superior synthesis speed over the autoregressive model, Tacotron 2, while maintaining comparable speech quality.

Recently, a new TTS model called EfficientTTS [370] has been introduced. This model outperforms previous models such as Tacotron 2 and Glow-TTS in terms of speech quality, training efficiency, and synthesis speed. The EfficientTTS model uses a multi-head attention mechanism to align input text and speech encodings, enabling it to generate high-quality speech with fewer parameters and faster synthesis speed. Overall, the introduction of normalizing flow and the development of models such as Glow-TTS and EfficientTTS have significantly improved the quality and efficiency of TTS systems.

To resolve output diversity issues in parallel TTS architectures, normalizing flow has been introduced to model the duration of speech [244, 370, 478]. Glow-TTS [244] is a flow-based generative model for parallel TTS that does not require any external aligner12345. It is built on the generic Glow model that is previously used in computer vision and vocoder models3. Glow-TTS is designed to produce mel-spectrograms from text input, which can then be converted to speech audio4. It has been shown to achieve an order-of-magnitude speed-up over the autoregressive model, Tacotron 2, at synthesis with comparable speech quality. EfficientTTS is a recent study that proposed a new TTS model, which significantly outperformed models such as Tacotron 2 [476] and Glow-TTS [244] in terms of speech quality, training efficiency, and synthesis speed. The EfficientTTS [370] model uses a multi-head attention mechanism to align the input text and speech encodings, enabling it to generate high-quality speech with fewer parameters and faster synthesis speed.

5.2.5 Speech Resynthesis

Speech resynthesis is the process of generating speech from a given input signal. The input signal can be in various forms, such as a digital recording, text, or other types of data. The aim of speech resynthesis is to create an output that closely resembles the original signal in terms of sound quality, prosody, and other acoustic characteristics. Speech resynthesis is an important research area with various applications, including speech enhancement [188, 356, 507], and voice conversion [355]. Recent advancements in speech resynthesis have revolutionized the field by incorporating

self-supervised discrete representations to generate disentangled representations of speech content, prosodic information, and speaker identity. These techniques enable the generation of speech in a controlled and precise manner, as seen in [274, 419, 427, 480]. The objective is to generate high-quality speech that maintains or degrades acoustic cues, such as phonotactics, syllabic rhythm, or intonation, from natural speech recordings.

Speech resynthesis is a vital research area with various applications, including speech enhancement and voice conversion, and recent advancements have revolutionized the field by incorporating self-supervised discrete representations. These techniques enable the generation of high-quality speech that maintains or degrades acoustic cues from natural speech recordings, and they have been used in the GSLM [274] architecture for acoustic modeling, speech recognition, and synthesis, as outlined in Figure 15. It comprises a discrete speech encoder, a generative language model, and a speech decoder, all trained without supervision. GSLM is the only prior work addressing the generative aspect of speech pre-training, which builds a text-free language model using discovered units.

5.2.6 Voice Conversion

Modifying a speaker's voice in a provided audio sample to that of another individual is called voice conversion, preserving linguistic content information. TTS and Voice conversion share a common objective of generating natural speech. While models based on RNNs and CNNs have been successfully applied to voice conversion, the use of the transformer has shown promising results. Voice Transformer Network (VTN) [204] is a seq2seq voice conversion (VC) model based on the transformer architecture with TTS pre-training. Seq2seq VC models are attractive as they can convert prosody, and the VTN is a novel approach in this field that has been proven to be effective in converting speech from a source to a target without changing the linguistic content.

ASR and TTS-based voice conversion is a promising approach to voice conversion [511]. It involves using an ASR model to transcribe the source speech into the linguistic representation and then using a TTS model to synthesize the target speech with the desired voice characteristics [420]. However, this approach overlooks the modeling of prosody, which plays an important role in speech naturalness and conversion similarity. To address this issue, researchers have proposed to directly predict prosody from the linguistic representation in a target-speaker-dependent manner [626]. Other researchers have explored using a mix of ASR and TTS features to improve the quality of voice conversion [82, 203, 624, 642].

CycleGAN [232–234], VAE [78, 229, 572], and VAE with the generative adversarial network [185] are other popular VC other popular approaches for non-parallel-voice conversion. CycleGAN-VC [232] uses a cycle-consistent adversarial network to convert the source voice to the target voice and can generate high-quality speech without any extra data, modules, or alignment procedure. Several improvements and modifications are also proposed in recent years [185, 233, 234]. VAE-based voice conversion is a promising approach that can generate high-quality speech with a small amount of training data [78, 229, 572].

5.2.7 Vocoders

The field of audio synthesis has undergone significant advancements in recent years, with various approaches proposed to enhance the quality of synthesized audio. Prior studies have concentrated on improving discriminator architectures or incorporating auxiliary training losses. For instance, MelGAN introduced a multiscale discriminator that uses window-based discriminators at different scales and applies average pooling to downsample the raw waveform. It enforces the correspondence between the input Mel spectrogram and the synthesized waveform using an L1 feature matching loss from the discriminator. In contrast, GAN-TTS [38] utilizes an ensemble of discriminators that

operate on random windows of different sizes and enforce the mapping between the conditioner and the waveform adversarially using conditional discriminators. Another approach, parallel WaveGAN [585], extends the single short-time Fourier transform loss to multi-resolution and employs it as an auxiliary loss for GAN training. Recently, some researchers have improved MelGAN by integrating the multi-resolution short-time Fourier transform loss. HiFi-GAN reuses the multi-scale discriminator from MelGAN and introduces the multi-period discriminator for high-fidelity synthesis. UnivNet employs a multi-resolution discriminator that takes multi-resolution spectrograms as input and can enhance the spectral structure of a synthesized waveform. In contrast, CARGAN integrates partial autoregression into the generator to enhance pitch and periodicity accuracy. The recent generative models for modeling raw audio can be categorized into the following groups.

- Autoregressive models: Although WaveNet is renowned for its exceptional ability to generate high-quality speech, including natural-sounding intonation and prosody, other neural vocoders have emerged as potential alternatives in recent years. For instance, LPCNet [525] employs a combination of linear predictive coding (LPC) and deep neural networks (DNNs) to generate speech of similar quality while being computationally efficient and capable of producing low-bitrate speech. Similarly, SampleRNN [366], an unconditional end-to-end model, has demonstrated potential as it leverages a hierarchical RNN architecture and is trained end-to-end to generate raw speech of high quality.
- Generative Adversarial Network (GAN) vocoders: Numerous vocoders have been created that employ Generative Adversarial Networks (GANs) to generate speech of exceptional quality. These GAN-based vocoders, which include MelGAN MelGAN [268]and HiFIGAN [261], are capable of producing high-fidelity raw audio by conditioning on mel spectrograms. Furthermore, they can synthesize audio at speeds several hundred times faster than real-time on a single GPU, as evidenced by research conducted in [37, 109, 261, 268, 585].
- Diffusion-based models: In recent years, there have been several novel architectures proposed that are based on diffusion. Two prominent examples of these are WaveGrad [64] and DiffWave [262]. The WaveGrad model architecture builds upon prior works from score matching and diffusion probabilistic models, while the DiffWave model uses adaptive noise spectral shaping to adapt the diffusion noise. This adaptation, achieved through time-varying filtering, improves sound quality, particularly in high-frequency bands. Other examples of diffusion-based vocoders include InferGrad [71], SpecGrad [257], and Priorgrad [286]. InfraGrad incorporates the inference process into training to reduce inference iterations while maintaining high quality. SpecGrad adapts the diffusion noise distribution to a given acoustic feature and uses adaptive noise spectral shaping to generate high-fidelity speech waveforms.
- Flow-based models: Parallel WaveNet, WaveGlow, etc. [252, 287, 347, 417, 425] are based on normalizing flows and are capable of generating high-fidelity speech in real-time. While flow-based vocoders generally perform worse than autoregressive vocoders with regard to modeling the density of speech signals, recent research [347] has proposed new techniques to improve their performance.

Universal neural vocoding is a challenging task that has achieved limited success to date. However, recent advances in speech synthesis have shown a promising trend toward improving zero-shot performance by scaling up model sizes. Despite its potential, this approach has yet to be extensively explored. Nonetheless, several approaches have been proposed to address the challenges of universal vocoding. For example, WaveRNN has been utilized in previous studies to achieve universal vocoding (Lorenzo-Trueba et al. [337]; Paul et al. [409]). Another approach Jiao et al. [215] developed involves

constructing a universal vocoder using a flow-based model. Additionally, the GAN vocoder has emerged as a promising candidate for this task, as suggested by You et al. [603].

5.2.8 Controllable Speech Synthesis

Controllable Speech Synthesis [118, 269, 449, 522, 526, 561, 653] is a rapidly evolving research area that focuses on generating natural-sounding speech with the ability to control various aspects of speech, including pitch, speed, and emotion. Controllable Speech Synthesis is positioned in the emerging field of affective computing at the intersection of three disciplines: expressive speech analysis [512], natural language processing, and machine learning. This field aims to develop systems capable of recognizing, interpreting, and generating human-like emotional responses in interactions between humans and machines.

Expressive speech analysis is a critical component of this field. It provides mathematical tools to analyse speech signals and extract various acoustic features, including pitch, loudness, and duration, that convey emotions in speech. Natural language processing is also crucial to this field, as it helps to process the text input and extract the meaning and sentiment of the words. Finally, machine learning techniques are used to model and control the expressive features of the synthesized speech, enabling the systems to produce more expressive and controllable speech [10, 199, 267, 288, 330, 399, 496, 527, 643].

In the last few years, notable advancements have been achieved in this field [159, 242, 439], and several approaches have been proposed to enhance the quality of synthesized speech. For example, some studies propose using deep learning techniques to synthesize expressive speech and conditional generation models to control the prosodic features of speech [242, 439]. Others propose using motion matching-based algorithms to synthesize gestures from speech [159].

5.2.9 Disentangling and Transferring

The importance of disentangled representations for neural speech synthesis cannot be overstated, as it has been widely recognized in the literature that this approach can greatly improve the interpretability and expressiveness of speech synthesis models [189, 353, 426]. Disentangling multiple styles or prosody information during training is crucial to enhance the quality of expressive speech synthesis and control. Various disentangling techniques have been developed using adversarial and collaborative games, the VAE framework, bottleneck reconstructions, and frame-level noise modeling combined with adversarial training.

For instance, Ma et al. [353] have employed adversarial and collaborative games to enhance the disentanglement of content and style, resulting in improved controllability. Hsu et al. [189] have utilized the VAE framework with adversarial training to separate speaker information from noise. Qian et al. [426] have introduced speech flow, which can disentangle rhythm, pitch, content, and timbre through three bottleneck reconstructions. In another work based on, adversarial training, Zhang et al. [619] have proposed a method that disentangles noise from the speaker by modeling the noise at the frame level.

Developing high-quality speech synthesis models that can handle noisy data and generate accurate representations of speech is a challenging task. To tackle this issue, Zhang et al. [627] propose a novel approach involving multi-length adversarial training. This method allows for modeling different noise conditions and improves the accuracy of pitch prediction by incorporating discriminators on the mel-spectrogram. By replacing the traditional pitch predictor model with this approach, the authors demonstrate significant improvements in the fidelity of synthesized speech.

5.2.10 Robustness

Using neural TTS models can present issues with robustness, leading to low-quality audio samples for unseen or atypical text. In response, Li et al. [303] proposed RobuTrans [303], a robust transformer that converts input text to linguistic features before feeding it to the encoder. This model also includes modifications to the attention mechanism and position embedding, resulting in improved MOS scores compared to other TTS models. Another approach to enhancing robustness is the s-Transformer, introduced by Wang et al. [556], which models speech at the segment level, allowing it to capture long-term dependencies and use segment-level encoder-decoder attention. This technique performs similarly to the standard transformer, exhibiting robustness for extra-long sentences. Lastly, Zheng et al. [647] proposed an approach that combines a local recurrent neural network with the transformer to capture sequential and local information in sequences. Evaluation of a 20-hour Mandarin speech corpus demonstrated that this model outperforms the transformer alone in performance.

In their recent paper [587], the authors proposed a novel method for extracting dynamic prosody information from audio recordings, even in noisy environments. Their approach employs probabilistic denoising diffusion models and knowledge distillation to learn speaking style features from a teacher model, resulting in a highly accurate reproduction of prosody and timber. This model shows great potential in applications such as speech synthesis and recognition, where noise-robust prosody information is crucial. Other noteworthy advances in the development of robust TTS systems include the work by [478], which focuses on a robust speech-text alignment module, as well as the use of normalizing flows for diverse speech synthesis.

5.2.11 Low-Resource Neural Speech Synthesis

High-quality paired text and speech data are crucial for building high-quality Text-to-Speech (TTS) systems [142]. Unfortunately, most languages are not supported by popular commercialized speech services due to the lack of sufficient training data [581]. To overcome this challenge, researchers have developed TTS systems under low data resource scenarios using various techniques [123, 142, 517, 581].

Several techniques have been proposed by researchers to enhance the efficiency of low-resource/Zero-shot TTS systems. One of these is the use of semi-supervised speech synthesis methods that utilize unpaired training data to improve data efficiency, as suggested in a study by Liu et al. [321]. Another method involves cascading pre-trained models for ASR, MT, and TTS to increase data size from unlabelled speech, as proposed by Nguyen et al. [386]. In addition, researchers have employed crowdsourced acoustic data collection to develop TTS systems for low-resource languages, as shown in a study by Butryna et al. [49]. Huang et al. [199] introduced a zero-shot style transfer approach for out-of-domain speech synthesis that generates speech samples exhibiting a new and distinctive style, such as speaker identity, emotion, and prosody.

5.3 Speaker recognition

5.3.1 Task Description

Speech signal consists of information on various characteristics of a speaker, such as origin, identity, gender, emotion, etc. This property of speech allows speech-based speaker profiling with a wide range of applications in forensics, recommendation systems, etc. The research on recognizing speakers is extensive and aims to solve two major tasks: speaker identification (what is the identity?) and speaker verification (is the speaker he/she claims to be?). Speaker recognition/verification tasks require extracting a fixed-length vector, called speaker embedding, from unconstrained utterances. These embeddings represent the speakers and can be used for identification or verification tasks.

Recent state-of-the-art speaker-embedding-extractor models are based on DNNs and have shown superior performance on both speaker identification and verification tasks.

- **Speaker Recognition** (SR) relies on speaker identification as a key aspect, where an unknown speaker's speech sample is compared to speech models of known speakers to determine their identity. The primary aim of speaker identification is to distinguish an individual's identity from a group of known speakers. This process involves a detailed analysis of the speaker's voice characteristics such as pitch, tone, accent, and other pertinent features to establish their identity. Recent advancements in deep learning techniques have significantly enhanced speaker identification, leading to the creation of accurate, efficient, and end-to-end models. Various deep learning-based models such as CNNs, RNNs, and their combinations have demonstrated exceptional performance in several subtasks of speaker identification, including verification, identification, diarization, and robust recognition [241, 253, 445].
- **Speaker Verification** (SV) is a process that involves confirming the identity of a speaker through their speech. It differs from speaker identification, which aims to identify unknown speakers by comparing their voices with that of registered speakers in a database. Speaker verification verifies whether a speaker is who they claim to be by comparing their voice with an available speaker template. Deep learning-based speaker verification relies on Speaker Representation based on embeddings, which involves learning low-dimensional vector representations from speech signals that capture speaker characteristics, such as pitch and speaking style, and can be used to compare different speech signals and determine their similarity.

5.3.2 Dataset

The VoxCeleb dataset (1 & 2) is frequently used for speaker recognition research, as cited in [88]. This dataset contains speech data from open-source media utilizing a fully automated pipeline that employs computer vision methods. The pipeline retrieves videos from YouTube, performs active speaker verification via a two-stream synchronization CNN, and then confirms the speaker's identity using CNN-based facial recognition. Another widely used dataset is TIMIT, which includes recordings of phonetically balanced English sentences spoken by hundreds of distinct speakers. TIMIT is typically employed to assess speech recognition and speaker identification systems, as mentioned in [148].

Other noteworthy datasets in the field include the SITW database [364], which provides hand-annotated speech samples for benchmarking text-independent speaker recognition technology, and the RSR2015 database [279], which contains speech recordings acquired in a typical office environment using multiple mobile devices. Additionally, the RedDots project [284] and VOICES corpus [451] offer unique collections of offline voice recordings in furnished rooms with background noise, while the CN-CELEB database [131] focuses on a specific person of interest extracted from bilibili.com using an automated pipeline followed by human verification.

The BookTubeSpeech dataset [414] was also collected using an automated pipeline from BookTube videos, and the Hi-MIA database [428] was designed specifically for far-field scenarios using multiple microphone arrays. The FFSVC20 challenge [429] and DIHARD challenge [458] are speaker verification and diarization research initiatives focusing on far-field and robustness challenges, respectively. Finally, the LibriSpeech dataset [401], originally intended for speech recognition, is also useful for speaker recognition tasks due to its included speaker identity labels.

5.3.3 Models

Speaker identification (SI) and verification (SV) are crucial research topics in the field of speech technology due to their significant importance in various applications such as security [121], forensics [263], biometric authentication [165], and speaker diarization [578]. Speaker recognition has become more popular with technological advancements, including the Internet of Things (IoT), smart devices, voice assistants, smart homes, and humanoids. Therefore, a significant quantity of research has been conducted in this field, and many methods have been developed, making the state-of-the-art in this field quite mature and versatile. However, it has become increasingly challenging to provide an overview of the various methods due to the high number of studies in the field.

A neural network approach for speaker verification was first attempted by Variani et al. [531] in 2014, utilizing four fully connected layers for speaker classification. Their approach has successfully verified speakers with short-duration utterances by obtaining the d-vector by averaging the output of the last hidden layer across frames. Although various attempts have been made to directly learn speaker representation from raw waveforms by other researchers (Jung et al. [220], Ravanelli and Bengio [443]), other well-designed neural networks like CNNs and RNNs have been proposed for speaker verification tasks by Ye and Yang [598]. Nevertheless, the field still requires more powerful deep neural networks for superior extraction of speaker features. To avoid plagiarism, citing all sources used in the text is important.

Speaker verification has seen notable advancements with the advent of more powerful deep neural networks. One such model is the x-vector-based system proposed by Snyder et al. [491], which has gained widespread popularity due to its remarkable performance. Since its introduction, the x-vector system has undergone significant architectural enhancements and optimized training procedures [99]. The widely-used ResNet [170] architecture has been incorporated into the system to improve its performance further. Adding residual connections between frame-level layers has been found to improve the embeddings [147, 611]. This technique has also aided in faster convergence of the back-propagation algorithm and mitigated the vanishing gradient problem [170]. Tang et al. [509] proposed further improvements to the x-vector system. They introduced a hybrid structure based on TDNN and LSTM to generate complementary speaker information at different levels. They also suggested a multi-level pooling strategy to collect the speaker information from global and local perspectives. These advancements have significantly improved speaker verification systems' performance and paved the way for further developments in the field.

Desplanques et al. [104] propose a state-of-the-art architecture for speaker verification utilizing a Time Delay Neural Network (TDNN) called ECAPA-TDNN. The paper presents a range of enhancements to the existing x-vector architecture that leverages recent developments in face verification and computer vision. Specifically, the authors suggest three major improvements. Firstly, they propose restructuring the initial frame layers into 1-dimensional Res2Net modules with impactful skip connections, which can better capture the relationships between different time frames. Secondly, they introduce Squeeze-and-Excitation blocks to the TDNN layers, which help highlight the most informative channels and improve feature discrimination. Lastly, the paper proposes channel attention propagation and aggregation to efficiently propagate attention weights through multiple TDNN layers, further enhancing the model's ability to discriminate between speakers.

Additionally, the paper presents a new approach that utilizes ECAPA-TDNN from the speaker recognition domain as the backbone network for a multiscale channel adaptive module. The proposed method achieves promising results, demonstrating the effectiveness of the proposed architecture in speaker verification. Overall, ECAPA-TDNN offers a comprehensive solution to

speaker verification by introducing several novel contributions that improve the existing x-vector architecture, which has been state-of-the-art in speaker verification for several years. The proposed approach also achieves promising results, suggesting that the proposed architecture can effectively tackle the challenges of speaker verification.

The attention mechanism is a powerful method for obtaining a more discriminative utterance-level feature by explicitly selecting frame-level representations that better represent speaker characteristics. Recently, the Transformer model with a self-attention mechanism has become effective in various application fields, including speaker verification. The Transformer architecture has been extensively explored for speaker verification. TESA [363] is an architecture based on the Transformer's encoder, proposed as a replacement for conventional PLDA-based speaker verification to capture speaker characteristics better. TESA outperforms PLDA on the same dataset by utilizing the next sentence prediction task of BERT [105]. Zhu et al. [652] proposed a method to create fixed-dimensional speaker verification representation using a serialized multi-layer multi-head attention mechanism. Unlike other studies that redesign the inner structure of the attention module, their approach strictly follows the original Transformer, providing simple but effective modifications.

5.4 Speaker Diarization

5.4.1 Task Description

Speaker diarization is a critical component in the analysis of multi-speaker audio data, and it addresses the question of "who spoke when." The term "diarize" refers to the process of making a note or keeping a record of events, as per the English dictionary. A traditional speaker diarization system comprises several crucial components that work together to achieve accurate and efficient speaker diarization. In this section, we will discuss the different components of a speaker diarization system (Figure 16) and their role in achieving accurate speaker diarization.

- *Acoustic Features Extraction*: In the analysis of multi-speaker speech data, one critical component is the extraction of acoustic features [13, 515]. This process involves extracting features such as pitch, energy, and MFCCs from the audio signal. These acoustic features play a crucial role in identifying different speakers by analyzing their unique characteristics.
- *Segmentation*: Segmentation is a crucial component in the analysis of multi-speaker audio data, where the audio signal is divided into smaller segments based on the silence periods between speakers [13, 515]. This process helps in reducing the complexity of the problem and makes it easier to identify different speakers in smaller segments
- *Speaker Embedding Extraction*: This process involves obtaining a low-dimensional representation of each speaker's voice, which is commonly referred to as speaker embedding. This is achieved by passing the acoustic features extracted from the speech signal through a deep neural network, such as a CNN or RNN[490].
- *Clustering*: In this component, the extracted speaker embeddings are clustered based on similarity, and each cluster represents a different speaker [13, 515]. This process commonly uses unsupervised clustering algorithms, such as k-means clustering.
- *Speaker Classification*: In this component, the speaker embeddings are classified into different speaker identities using a supervised classification algorithm, such as SVM or MLP [13, 515].
- *Re-segmentation*: This component is responsible for refining the initial segmentation by adjusting the segment boundaries based on the classification results. It helps in improving the accuracy of speaker diarization by reducing the errors made during the initial segmentation.

Various studies focus on traditional speaker diarization systems [13, 515]. This paper will review the recent efforts toward deep learning-based speaker diarizations techniques.

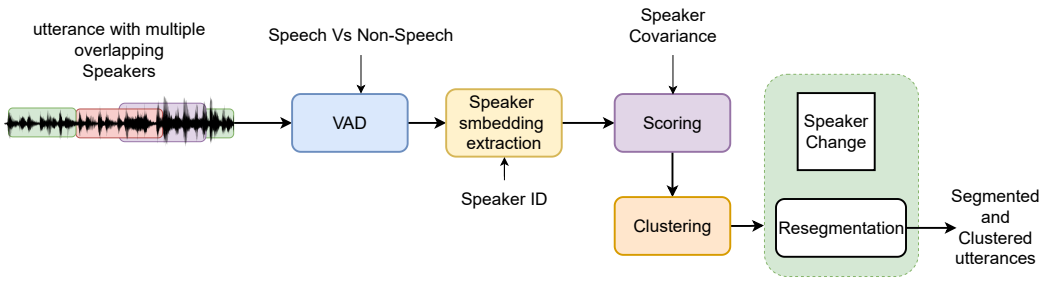

Fig. 16. Speaker diarization system diagram showcasing the process of identifying and differentiating multiple speakers in an audio recording using various techniques such as VAD, segmentation, clustering and re-segmentation.

5.4.2 Dataset

- *NIST SRE 2000 (Disk-8) or CALLHOME dataset*: The NIST SRE 2000 (Disk-8) corpus, also referred to as the CALLHOME dataset, is a frequently utilized resource for speaker diarization in contemporary research papers. Originally released in 2000, this dataset comprises conversational telephone speech (CTS) collected from diverse speakers representing a wide range of ages, genders, and dialects. It includes 500 sessions of multilingual telephonic speech, each containing two to seven speakers, with two primary speakers in each conversation. The dataset covers various topics, including personal and familial relationships, work, education, and leisure activities. The audio recordings were obtained using a single microphone and had a sampling rate of 8 kHz, with 16-bit linear quantization.
- *Directions into Heterogeneous Audio Research (DIHARD) Challenge and dataset*: The DIHARD Challenge, organized by the National Institute of Standards and Technology (NIST), aims to enhance the accuracy of speech recognition and diarization in challenging acoustic environments, such as crowded spaces, distant microphones, and reverberant rooms. The challenge comprises tasks requiring advanced machine-learning techniques, including speaker diarization, recognition, and speech activity detection. The DIHARD dataset used in the challenge comprises over 50 hours of speech from more than 500 speakers, gathered from diverse sources like meetings, broadcast news, and telephone conversations. These recordings feature various acoustic challenges, such as overlapping speech, background noise, and distant or reverberant speech, captured through different microphone setups. To aid in the evaluation process, the dataset has been divided into separate development and evaluation sets. The assessment metrics used to gauge performance include diarization error rate (DER), as well as accuracy in speaker verification, identification, and speech activity detection.
- *Augmented Multi-party Interaction (AMI) database*: The AMI database is a collection of audio and video recordings that capture real-world multi-party conversations in office environments. The database was developed as part of the AMI project, which aimed to develop technology for automatically analyzing multi-party meetings. The database contains over 100 hours of audio and video recordings of meetings involving four to seven participants, totaling 112 meetings. The meetings were held in multiple offices and were designed to reflect the kinds of discussions that take place in typical business meetings. The audio recordings were captured using close-talk microphones placed on each participant and additional microphones placed in the room to capture ambient sound. The video recordings

were captured using multiple cameras placed around the room. In addition to the audio and video recordings, the database also includes annotations that provide additional information about the meetings, including speaker identities, speech transcriptions, and information about the meeting structure (e.g., turn-taking patterns). The AMI database has been used extensively in research on automatic speech recognition, speaker diarization, and other related speech and language processing topics.

- *VoxSRC Challenge and VoxConverse corpus*: The VoxCeleb Speaker Recognition Challenge (VoxSRC) is an annual competition designed to assess the capabilities of speaker recognition systems in identifying speakers from speech recorded in real-world environments. The challenge provides participants with a dataset of audio and visual recordings of interviews, news shows, and talk shows featuring famous individuals. The VoxSRC encompasses several tracks, including speaker diarization, and comprises a development set (20.3 hours, 216 recordings) and a test set (53.5 hours, 310 recordings). Recordings in the dataset may feature between one and 21 speakers, with a diverse range of ambient noises, such as background music and laughter. To facilitate the speaker diarization track of the VoxSRC-21 and VoxSRC-22 competitions, VoxConverse, an audio-visual diarization dataset containing multi-speaker clips of human speech sourced from YouTube videos, is available, and additional details are provided on the project website [7].
- *LibriCSS*: The LibriCSS corpus is a valuable resource for researchers studying speech separation, recognition, and speaker diarization. The corpus comprises 10 hours of multichannel recordings captured using a 7-channel microphone array in a real meeting room. The audio was played from the LibriSpeech corpus, and each of the ten sessions was subdivided into six 10-minute mini-sessions. Each mini-session contained audio from eight speakers and was designed to have different overlap ratios ranging from 0% to 40%. To make research easier, the corpus includes baseline systems for speech separation and Automatic Speech Recognition (ASR) and a baseline system that integrates speech separation, speaker diarization, and ASR. These baseline systems have already been developed and made available to researchers.
- *Rich Transcription Evaluation Series*: The Rich Transcription Evaluation Series dataset is a collection of speech data used for speaker diarization evaluation. The Rich Transcription Fall 2003 Evaluation (RT-03F) was the first evaluation in the series focused on "Who Said What" tasks. The dataset has been used in subsequent evaluations, including the Second DIHARD Diarization Challenge, which used the Jaccard index to compute the JER (Jaccard Error Rate) for each pair of segmentations. The dataset is essential for data-driven spoken language processing methods and calculates speaker diarization accuracy at the utterance level. The dataset includes rules, evaluation methods, and baseline systems to promote reproducible research in the field. The dataset has been used in various speaker diarization systems and their subtasks in the context of broadcast news and CTS data
- *CHiME-5/6 challenge and dataset* The CHiME-5/6 challenge is a speech processing challenge focusing on distant multi-microphone conversational speech diarization and recognition in everyday home environments. The challenge provides a dataset of recordings from everyday home environments, including dinner recordings originally collected for and exposed during the CHiME-5 challenge. The dataset is designed to be representative of natural conversational speech. The challenge features two audio input conditions: single-channel and multichannel. Participants are provided with baseline systems for speech enhancement, speech activity detection (SAD), and diarization, as well as results obtained

[7] https://www.robots.ox.ac.uk/ vgg/data/voxconverse/

with these systems for all tracks. The challenge aims to improve the robustness of diarization systems to variations in recording equipment, noise conditions, and conversational domains.

- *AMI dataset*: The AMI database is a comprehensive collection of 100 hours of recordings sourced from 171 meeting sessions held across various locations. It features two distinct audio sources – one recorded using lapel microphones for individual speakers and the other using omnidirectional microphone arrays placed on the table. It is an ideal dataset for evaluating speaker diarization systems integrated with the ASR module. AMI's value proposition is further enhanced by providing forced alignment data, which captures the timings at the word and phoneme levels and speaker labeling. Finally, it's worth noting that each meeting session involves a small group of three to five speakers.

5.4.3 Models

Speaker diarization has been a subject of research in the field of audio processing, with the goal of separating speakers in an audio recording. In recent years, deep learning has emerged as a powerful technique for speaker diarization, leading to significant advancements in this field. In this article, we will explore some of the recent developments in deep learning architecture for speaker diarization, focusing on different modules of speaker diarization as outlined in Figure 16. Through this discussion, we will highlight major advancements in each module.

- *Segmentation and clustering*: Speaker diarization systems typically use a range of techniques for segmenting speech, such as identifying speaker change, uniform speaker segmentation, ASR-based word segmentation, and supervised speaker turn detection. However, each approach has its own benefits and drawbacks. Uniform speaker segmentation involves dividing speech into segments of equal length, which can be difficult to optimize to capture speaker turn boundaries and include enough speaker information. ASR-based word segmentation identifies word boundaries using automatic speech recognition, but the resulting segments may be too brief to provide adequate speaker information. Supervised speaker turn detection, on the other hand, involves a specialized model that can accurately identify speaker turn timestamps. While this method can achieve high accuracy, it requires labeled data for training. These techniques have been widely discussed in previous research, and choosing the appropriate one depends on the specific requirements of the application.
  - The authors in [94] propose real-time speaker diarization system that combines incremental clustering and local diarization applied to a rolling window of speech data and is designed to handle overlapping speech segments. The proposed pipeline is designed to utilize end-to-end overlap-aware segmentation to detect and separate overlapping speakers.
  - In another related work, authors in [620] introduce a novel speaker diarization system with a generalized neural speaker clustering module as the backbone.
  - In a recent study conducted by Park et al. [406], a new framework for spectral clustering is proposed that allows for automatic parameter tuning of the clustering algorithm in the context of speaker diarization. The proposed technique utilizes normalized maximum eigengap (NME) values to determine the number of clusters and threshold parameters for each row in an affinity matrix during spectral clustering. The authors demonstrated that their method outperformed existing state-of-the-art methods on two different datasets for speaker diarization.
  - Bayesian HMM clustering of x-vector sequences (VBx) diarization approach, which clusters x-vectors using a Bayesian hidden Markov model (BHMM) [278], combined with a ResNet101 (He et al. [170]) *x*-vector extractor achieves superior results on CALLHOME [107], AMI [51] and DIHARD II [459] datasets

- *Speaker Embedding Extraction and Classification*:
  - Attentive Aggregation for Speaker Diarization [271]: This approach uses an attention mechanism to aggregate embeddings from multiple frames and generate speaker embeddings. The speaker embeddings are then used for clustering to identify speaker segments.
  - End-to-End Speaker Diarization with Self-Attention [140]: This method uses a self-attention mechanism to capture the correlations between the input frames and generates embeddings for each frame. The embeddings are then used for clustering to identify speaker segments.
  - Wang et al. [554] present an innovative method for measuring similarity between speaker embeddings in speaker diarization using neural networks. The approach incorporates past and future contexts and uses a segmental pooling strategy. Furthermore, the speaker embedding network and similarity measurement model are jointly trained. The paper extends this framework to target-speaker voice activity detection (TS-VAD) [365]. The proposed method effectively learns the similarity between speaker embeddings by considering both past and future contexts.
  - Time-Depth Separable Convolutions for Speaker Diarization [259]: This approach uses time-depth separable convolutions to generate embeddings for each frame, which are then used for clustering to identify speaker segments. The method is computationally efficient and achieves state-of-the-art performance on several benchmark datasets.
- *Re-segmentation*:
  - Numerous studies in this field centre around developing a re-segmentation strategy for diarization systems that can effectively handle both voice activity and overlapped speech detection. This approach can also be a post-processing step to identify and assign overlapped speech regions accurately. Notable examples of such works include those by Bullock et al. [46] and Bredin and Laurent [44].
- *End-to-End Neural Diarization*: In addition to the above work, end-to-end speaker diarization systems have gained the attention of the research community due to their ability to handle speaker overlaps and their optimization to minimize diarization errors directly. In one such work, the authors propose end-to-end neural speaker diarization that does not rely on clustering and instead uses a self-attention-based neural network to directly output the joint speech activities of all speakers for each segment [140]. Following the trend, several other works propose enhanced architectures based on self-attention [317, 607]

5.5 Speech-to-speech translation

5.5.1 Task Description

Speech-to-text translation (ST) is the process of converting spoken language from one language to another in text form. Traditionally, this has been achieved using a cascaded structure that incorporates automatic speech recognition (ASR) and machine translation (MT) components. However, a more recent end-to-end (E2E) method [14, 59, 161, 463, 503, 616, 646] has gained popularity due to its ability to eliminate issues with error propagation and high latency associated with cascaded methods [60, 497]. The E2E method uses an audio encoder to analyze audio signals and a text decoder to generate translated text.

One notable advantage of ST systems is that they allow for more natural and fluent communication than other language translation methods. By translating speech in real-time, ST systems can capture the subtleties of speech, including tone, intonation, and rhythm, which are essential for effective communication. Developing ST systems is a highly intricate process that involves

integrating various technologies such as speech recognition, natural language processing, and machine translation. One significant obstacle in ST is the variation in accents and dialects across different languages, which can significantly impact the accuracy of the translation.

5.5.2 Dataset

There are numerous datasets available for the end-to-end speech translation task, with some of the most widely used ones being MuST-C [53], IWSLT [466], and CoVoST 2 [542]. These datasets cover a variety of languages, including English, German, Spanish, French, Italian, Dutch, Portuguese, Romanian, Arabic, Chinese, Japanese, Korean, and Russian. For instance, TED-LIUM [455] is a suitable dataset for speech-to-text, text-to-speech, and speech-to-speech translation tasks, as it contains transcriptions and audio recordings of TED talks in English, French, German, Italian, and Spanish. Another open-source dataset is Common Voice, which covers several languages, including English, French, German, Italian, and Spanish. Additionally, VoxForge[8] is designed for acoustic model training and includes speech recordings and transcriptions in several languages, including English, French, German, Italian, and Spanish. LibriSpeech [401] is a dataset of spoken English specifically designed for speech recognition and speech-to-text translation tasks. Lastly, How2 [120] is a multimodal machine translation dataset that includes speech recordings, text transcriptions, and video and image data, covering English, German, Italian, and Spanish. These datasets have been instrumental in training state-of-the-art speech-to-speech translation models and will continue to play a crucial role in further advancing the field.

5.5.3 Models

End-to-end speech translation models are a promising approach to direct the speech translation field. These models use a single sequence-to-sequence model for speech-to-text translation and then text-to-speech translation. In 2017, researchers demonstrated that end-to-end models outperform cascade models[3]. One study published in 2019 provides an overview of different end-to-end architectures and the usage of an additional connectionist temporal classification (CTC) loss for better convergence [26]. The study compares different end-to-end architectures for speech-to-text translation. In 2019, Google introduced Translatotron [213], an end-to-end speech-to-speech translation system. Translatotron uses a single sequence-to-sequence model for speech-to-text translation and then text-to-speech translation. No transcripts or other intermediate text representations are used during inference. The system was validated by measuring the BLEU score, computed with text transcribed by a speech recognition system. Though the results lag behind a conventional cascade system, the feasibility of the end-to-end direct speech-to-speech translation was demonstrated [213].

Another study published in 2020 describes an end-to-end speech translation system that combines pre-trained models (Wav2Vec 2.0 and mBART) with coupling modules between the encoder and decoder and uses an efficient fine-tuning technique, which trains only 20% of its total parameters [599]. The system was submitted to the IWSLT 2021 offline speech translation task by the UPC Machine Translation group. The task involved building a system capable of translating English audio recordings from TED talks into German text.

E2E ST is often improved by pretraining the encoder and/or decoder with transcripts from speech recognition or text translation tasks [106, 541, 580, 616]. Consequently, it has become the standard approach used in various toolkits [208, 541, 637, 646]. However, transcripts are not always available, and the significance of pretraining for E2E ST is rarely studied. Zhang et al. [615] explored the effectiveness of E2E ST trained solely on speech-translation pairs and proposed an algorithm for training from scratch. The proposed system outperforms previous studies in four

[8]http://www.voxforge.org/

benchmarks covering 23 languages without pretraining. The paper also discusses neural acoustic feature modeling, which extracts acoustic features directly from raw speech signals to simplify inductive biases and enhance speech description.

5.6 Speech enhancement

5.6.1 Task Description

In situations where there is ambient noise present, speech recognition systems can encounter difficulty in correctly interpreting spoken language signals, resulting in reduced performance [119]. One possible solution to address this issue is the development of speech enhancement systems that can eliminate noise and other types of signal distortion from spoken language, thereby improving signal quality. These systems are frequently implemented as a preprocessing step to enhance the accuracy of speech recognition and can serve as an effective approach for enhancing the performance of ASR systems in noisy environments. This section will delve into the significance of speech enhancement technology in boosting the accuracy of speech recognition.

5.6.2 Dataset

One popular dataset for speech enhancement tasks is AISHELL-4, which comprises authentic Mandarin speech recordings captured during conferences using an 8-channel circular microphone array. In accordance with [139], AISHELL-4 is composed of 211 meeting sessions, each featuring 4 to 8 speakers, for a total of 120 hours of content. This dataset is of great value for research into multi-speaker processing owing to its realistic acoustics and various speech qualities, including speaker diarization and speech recognition

Another popular dataset used for speech enhancement is the dataset from Deep Noise Suppression (DNS) challenge [446], a large-scale dataset of noisy speech signals and their corresponding clean speech signals. The DNS dataset contains over 10, 000 hours of noisy speech signals and over 1, 000 hours of clean speech signals, making it useful for training deep learning models for speech enhancement. The Voice Bank Corpus (VCTK) is another dataset containing speech recordings from 109 speakers, each recording approximately 400 sentences. The dataset contains clean and noisy speech recordings, making it useful for training speech enhancement models. These datasets provide realistic acoustics, rich natural speech characteristics, and large-scale noisy and clean speech signals, making them useful for training deep learning models.

5.6.3 Models

Several Classical algorithms have been reported in the literature for speech enhancement, including spectral subtraction [40], Wiener and Kalman filtering [312, 465], MMSE estimation [124], comb filtering [216], subspace methods [166]. Phase spectrum compensation [398]. However, classical algorithms such as spectral subtraction and Wiener filtering approach the problem in the spectral domain and are restricted to stationary or quasi-stationary noise. Neural network-based approaches inspired from other areas such as computer vision [9, 141, 182] and generative adversarial networks [137, 314, 456, 573] or developed for general audio processing tasks [152, 565] have outperformed the classical approaches. Various neural network models based on different architectures, including fully connected neural networks [583], deep denoising autoencoder [339], CNN [138], LSTM [74], and Transformer [256] have effectively handled diverse noisy conditions.

Diffusion-based models have also shown promising results for speech enhancement [291, 342, 600] and have led to the development of novel speech enhancement algorithms called Conditional Diffusion Probabilistic Model (CDiffuSE) that incorporates characteristics of the observed noisy speech signal into the diffusion and reverse processing [342]. CDiffuSE is a generalized formulation of the diffusion probabilistic model that can adapt to non-Gaussian real noises in the estimated

Table 8. Table summarizing the performance of different speech enhancement algorithms on the Deep Noise Suppression (DNS) Challenge dataset, showcasing the achieved improvements in terms of PESQ-WB, PESQ-NB, SI-SDR-WB, and SI-SDR-NB metrics, and identifying the top-performing methods in each category

Model	Metric				Architecture
	PESQ-WB	PESQ-NB	SI-SDR-WB	SI-SDR-NB	
FRCRN [641]	3.23	-	-	-	U-Net + CRN
Sudo rm -rf [520]	2.95	-	19.7	-	UConvBlock + CNN
DCTCRN-P [304]	2.82	-	-	-	CNN
PoCoNet [210]	2.7885	-	-	-	
FullSubNet [167]	2.777	3.305	17.29	-	LSTM
RNN-Modulation [537]	2.75	-	-	-	GRU
Conv-TasNet-SNR [264]	2.73	-	-	-	CNN
Sudo rm-rf [519]	2.69	-	18.6	-	UConvBlock + CNN
RemixIT [520]	2.34	-	16.0	-	UConvBlock
SN-Net [645]	-	3.39	-	19.52	CNN
DCCRN-E-Aug [196]	-	3.214	-	-	CNN + LSTM
DTLN [569]	-	3.04	16.34	-	LSTM
DCCRN-E [196]	-	3.04	-	-	CNN + LSTM

speech signal. Another diffusion-based model for speech enhancement is StoRM [291], which stands for Stochastic Regeneration Model. It uses a predictive model to remove vocalizing and breathing artifacts while producing high-quality samples using a diffusion process, even in adverse conditions. StoRM has shown great ability at bridging the performance gap between predictive and generative approaches for speech enhancement. Furthermore, authors in [600] propose cold diffusion process is an advanced iterative version of the diffusion process to recover clean speech from noisy speech. According to the authors, it can be utilized to restore high-quality samples from arbitrary degradations. Table 8 summarizing the performance of different speech enhancement algorithms on the Deep Noise Suppression (DNS) Challenge dataset using different metrics.

5.7 Audio Super Resolution

5.7.1 Task Description

Audio super-resolution is a technique that involves predicting the missing high-resolution components of low-resolution audio signals. Achieving this task can be difficult due to the continuous nature of audio signals. Current methods typically approach super-resolution by treating audio as discrete data and focusing on fixed scale factors. In order to accomplish audio super-resolution, deep neural networks are trained using pairs of low and high-quality audio examples. During testing, the model predicts missing samples within a low-resolution signal. Some recent deep network approaches have shown promise by framing the problem as a regression issue either in the time or frequency domain [313]. These methods have been able to achieve impressive results.

5.7.2 Datasets

This section provides an overview of the diverse datasets utilized in Audio Super Resolution literature. One of the most frequently used datasets is the MUSDB18, specifically designed for music source separation and enhancement. This dataset encompasses more than 150 songs with distinct tracks for individual instruments. Another prominent dataset is UrbanSound8K, which comprises over, 8000 environmental sound files collected from 10 different categories, making it ideal for evaluating Audio Super Resolution algorithms in noisy environments. Furthermore, the VoiceBank dataset is another essential resource for evaluating Audio Super Resolution systems, comprising over 10,000 speech recordings from five distinct speakers. This dataset offers a rich source of information for assessing speech processing systems, including Audio Super Resolution. Another dataset, LibriSpeech, features more than 1000 hours of spoken words from several books and speakers, making it valuable for evaluating Audio Super Resolution algorithms to enhance the quality of spoken words. Finally, the TED-LIUM dataset, which includes over 140 hours of speech recordings from various speakers giving TED talks, provides a real-world setting for evaluating Audio Super Resolution algorithms for speech enhancement. By using these datasets, researchers can evaluate Audio Super Resolution systems for a wide range of audio signals and improve the generalizability of these algorithms for real-world scenarios.

5.7.3 Models

Audio super-resolution has been extensively explored using deep learning architectures [7, 39, 163, 247, 283, 313, 326, 384, 442, 601]. One notable paper by Rakotonirina [442] proposes a novel network architecture that integrates convolution and self-attention mechanisms for audio super-resolution. Specifically, they use Attention-based Feature-Wise Linear Modulation (AFiLM) [442] to modulate the activations of the convolutional model. In another recent work by Yoneyama et al. [601], the super-resolution task is decomposed into domain adaptation and resampling processes to handle acoustic mismatch in unpaired low- and high-resolution signals. To address this, they jointly optimize the two processes within the CycleGAN framework.

Moreover, the Time-Frequency Network (TFNet) [313] proposed a deep network that achieves promising results by modeling the task as a regression problem in either time or frequency domain. To further enhance audio super-resolution, the paper proposes a time-frequency network that combines time and frequency domain information. Finally, recent advancements in diffusion models have introduced new approaches to neural audio upsampling. Specifically, Lee and Han [283], and Han and Lee [163] propose NU-Wave 1 and 2 diffusion probabilistic models, respectively, which can produce high-quality waveforms with a sampling rate of 48kHz from coarse 16kHz or 24kHz inputs. These models are a promising direction for improving audio super-resolution.

5.8 Voice Activity Detection (VAD)

5.8.1 Task Description

Due to the increasing sophistication of mobile devices like smartphones, speech-controlled applications have become incredibly popular. These apps offer a hands-free method for controlling home devices, facilitating telephony, and allowing drivers to safely use their vehicle's infotainment systems while on the go. However, accurately distinguishing between noise and human speech is critical for these applications to work without interruption. To overcome this issue, Voice Activity Detection (VAD) systems have been created to recognize speech presence or absence, thus ensuring consistent and effective operation.

5.8.2 Datasets

Voice activity detection models can be trained and evaluated using various datasets, each with unique features. The TIMIT dataset is popular, providing, 6300 phonetically transcribed utterances from 630 speakers. On the other hand, CHiME-5 is designed for speech separation and recognition in real-world environments and includes multichannel recordings of 20 speakers in locations such as cafés, buses, and pedestrian areas. Despite its primary purpose, CHiME-5 is widely used for voice activity detection. AURORA-4 is specifically designed to evaluate the robustness of ASR systems and contains over 10, 000 in noisy speech utterances recorded in environments like car noise, babble noise, and street noise. It is also extended to VAD for evaluating challenging scenarios. DEMAND is a suitable dataset for evaluating VAD algorithms as it includes over 1200 artificially created noise signals with various noise types like white noise, pink noise, and café noise. Finally, VoxCeleb contains over 100,000 utterances from more than 6,000 speakers, primarily designed for speaker recognition systems evaluation, but it can also be used for voice activity detection.

5.8.3 Models

Recent advances in deep learning have greatly improved the performance of voice activity detection (VAD), particularly in noisy environments [373, 450]. To further improve VAD accuracy, researchers have explored various deep learning architectures, including NAS-VAD [450] and self-attentive VAD [217]. NAS-VAD employs neural architecture search to reduce the need for human effort in network design and has demonstrated superior performance in terms of AUC and F1-score compared to other models. Similarly, self-attentive VAD uses a self-attention mechanism to capture long-term dependencies in input signals and has also outperformed other models on the TIMIT dataset. Additionally, a deep neural network (DNN) system has been proposed for automatic speech detection in audio signals [373]. This system uses MLPs, RNNs, and CNNs, with CNNs delivering the best performance. Furthermore, a hybrid acoustic-lexical deep learning approach has been proposed for deception detection, combining both acoustic and lexical features.

5.9 Speech Quality Assessment

5.9.1 Task Description

Speech quality assessment is a crucial process that involves the objective evaluation of speech signals using various metrics and measures. The primary aim of this assessment is to determine the level of intelligibility and comprehensibility of speech to a human listener. Although human evaluation is considered the gold standard for assessing speech quality, it can be time-consuming, expensive, and not scalable. Mean opinion score (MOS) is the most commonly used and reliable method of obtaining human judgments for speech quality estimation. Accurate speech quality assessment is essential in the development and design of real-world applications such as ASR, Speech Enhancement, and VoIP.

5.9.2 Datasets

The speech quality assessment algorithms are evaluated using several datasets, each with unique characteristics. The TIMIT Acoustic-Phonetic Continuous Speech Corpus [148] has clean speech recordings and artificially generated degraded versions for speech synthesis and quality assessment research. The NOIZEUS dataset [197] is designed for evaluating noise reduction and speech quality assessment algorithms, with clean speech and artificially degraded versions containing various types of noise and distortion. The ETSI Aurora databases [354] are used for evaluating speech enhancement techniques and quality assessment algorithms, containing speech recordings with different types of distortions like acoustic echo and background noise. Furthermore, for training

and validation, the clean speech recordings from the DNS Challenge [446] can be used along with the noise dataset such as FSDK50 [134] for additive noise degradation.

5.9.3 Models

Current objective methods such as Perceptual Evaluation of Speech Quality (PESQ) [453] and Perceptual Objective Listening Quality Assessment (POLQA) [35] for evaluating the quality of speech mostly rely on the availability of the corresponding clean reference. These methods fail in real-world scenarios where the ground truth clean reference is unavailable. In recent years, several attempts to automatically estimate the MOS using neural networks for performing quality assessment and predicting ratings or scores have attracted much attention [52, 54, 114, 115, 395, 495]. These approaches outperform traditional approaches without the need for a clean reference. However, they lack robustness and generalization capabilities, limiting their use in real-world applications. The authors in [395] explore Deep machine listening for Estimating Speech Quality (DESQ) for predicting the perceived speech quality based on phoneme posterior probabilities obtained using a deep neural network.

In recent years, there have been several quality assessment frameworks developed to estimate speech quality, such as NORESQA [362] based on non-matching reference (NMR). NORESQA takes inspiration from the human ability to assess speech quality even when the content is non-matching. Additionally, NORESQA introduces two new metrics - NORESQA-score, which is based on SI-SDR for speech, and NORESQA-MOS, which evaluates the Mean Opinion Score (MOS) of a speech recording using non-matching references. A recent extension to NORESQA, known as NORESQA-MOS, has been proposed in [361]. The primary difference between these frameworks is that while NORESQA estimates speech quality using non-matching references through NORESQA-score and NORESQA-MOS, NORESQA-MOS is specifically designed to assess the MOS of a given speech recording using NMRs.

5.10 Speech Separation

5.10.1 Task Description

Speech separation refers to separating a mixed audio signal into its sources, including speech, music, and background noise. The problem is often referred to as the cocktail party problem [169], as it mimics the difficulty of listening to a conversation in a noisy room with multiple speakers. This problem is particularly relevant in real-world scenarios such as phone conversations, meetings, and live events, where various extraneous sounds may contaminate speech. Traditionally, speech separation has been studied as a signal-processing problem, where researchers have focused on developing algorithms to separate sources based on their spectral characteristics [535, 612]. However, recent advances in machine learning have led to a new approach that formulates speech separation as a supervised learning problem [175, 345, 564]. This approach has seen a significant improvement in performance with the advent of deep neural networks, which can learn complex relationships between input features and output sources.

5.10.2 Datasets

The WSJ0-2mix dataset comprises mixtures of two Wall Street Journal corpus (WSJ) speakers. It consists of a training set of 30,000 mixtures and a test set of 5000 mixtures, and it has been widely used to evaluate speech separation algorithms. CHiME-4 is a dataset that contains recordings of multiple speakers in real-world environments, such as a living room, a kitchen, and a café and is designed to test algorithms in challenging acoustic environments. TIMIT-2mix is a dataset based on the TIMIT corpus, consisting of mixtures of two speakers, and includes a training set of 462 mixtures and a test set of 400 mixtures. The dataset provides a more controlled environment than

Table 9. Table comparing the performance of different speech separation methods using SI-SDRi metrics on various speech separation benchmarks.

Model	Architecture	WSJ0-2mix	WSJ0-3mix	WSJ0-5mix	Libri2Mix	Libri5Mix	Libri10Mix	Libri20Mix	WHAM
Separate And Diffuse [350]	Diffusion	23.9	20.9	-	21.5	14.2	9	5.2	-
MossFormer (L) [640]	Transformer	22.8	21.2	-	-	-	-	-	-
MossFormer (M) [640]	Transformer	22.5	20.8	-	-	-	-	-	17.3
SepFormer [499]	Transformer	22.3	19.5	-	-	-	-	-	-
Sandglasset [276]	Transformer + LSTM	21.0	19.5	-	-	-	-	-	-
Hungarian PIT [116]		-	-	13.22	-	12.72	7.78	4.26	-
TDANet (L) [301]		-	-	-	17.4	-	-	-	15.2
TDANet [301]		-	-	-	16.9	-	-	-	14.8
Sepit [349]		22.4	20.1	-	-	13.7	8.2	-	-
Gated DualPathRNN [379]	CNN + LSTM	20.12	16.85	10.56	-	-	-	-	-
Dual-path RNN [344]	LSTM	18.8	-	-	-	-	-	-	-
Conv-Tasnet [346]	CNN	15.3	-	-	-	-	-	-	-

CHiME-4 to test speech separation algorithms. LibriMix is derived from the LibriSpeech corpus and includes mixtures of up to four speakers, with a training set of 100,000 mixtures and a test set of 1,000 mixtures, providing a more realistic and challenging environment than WSJ0-2mix. Lastly, the MUSDB18 dataset contains mixtures of music tracks separated into individual stems, including vocals, drums, bass, and other instruments. It consists of a training set of 100 songs and a test set of 50 songs. Despite not being specifically designed for that purpose, it has been used as a benchmark for evaluating speech separation algorithms.

5.10.3 Models

Deep Clustering++ [175], first proposed in 2015, employs deep neural networks to extract features from the input signal and cluster similar feature vectors in a latent space to separate different speakers. The model's performance is improved using spectral masking and a permutation invariant training method. The advantage of this model is its ability to handle multiple speakers, but it also has a high computational cost. Chimera++ [564] is another effective model that combines deep clustering with mask-inference networks in a multi-objective training scheme. The model is trained using a multitask learning approach, optimizing speech enhancement and speaker identification. Chimera++ can perform speech enhancement and speaker identification but has a relatively long training time.

TasNet v2 [345] employs a convolutional neural network (CNN) to process the input signal and generate a time-frequency mask for each source. The model is trained using an invariant permutation training (PIT) method [258], which enables it to separate multiple sources accurately. TasNet v2 achieves state-of-the-art performance in various speech separation tasks with high separation accuracy, but its disadvantage is its relatively high computational cost. The variant of TasNet based on CNNs is proposed in [346]. The model is called Conv-TasNet and can generate a time-frequency mask for each source to obtain the separated source's signal. Compared to previous models, Conv-TasNet has faster processing time but lower accuracy.

In recent research, encoder-decoder architectures have been explored for effectively separating source signals. One promising approach is the Hybrid Tasnet architecture [590], which utilizes an encoder to extract features from the input signal and a decoder to generate the independent sources. This hybrid architecture captures both short-term and long-term dependencies in the input signal,

leading to improved separation performance. However, it should be noted that this model's higher computational cost should be considered when selecting an appropriate separation method.

Dual-path RNN [344] uses RNN architecture to perform speech separation. The model uses a dual-path structure [344] to capture low-frequency and high-frequency information in the input signal. Dual-path RNN achieves impressive performance in various speech separation tasks. The advantage of this model is its ability to capture low-frequency and high-frequency information, but its disadvantage is its high computational cost. Gated DualPathRNN [379] is a variant of Dual-path RNN that employs gated recurrent units (GRUs) to improve the model's performance. The model uses a gating mechanism to control the flow of information in the recurrent network, allowing it to capture long-term dependencies in the input signal. Gated DualPathRNN achieves state-of-the-art performance in various speech separation tasks. The advantage of this model is its ability to capture long-term dependencies, but its disadvantage is its higher computational cost than other models.

Wavesplit [610] employs a Wave-U-Net [498] architecture to perform speech separation. The model uses a fully convolutional neural network to extract features from the input signal and generate a time-frequency mask for each source. Wavesplit achieves impressive performance in various speech separation tasks. The advantage of this model is its high separation accuracy and relatively fast processing time, but its disadvantage is its relatively high memory usage.

Numerous studies have investigated the application of Transformer architecture in the context of speech separation. One such study is SepFormer [499], which has yielded encouraging outcomes on the WSJ0-2mix and WSJ0-3mix datasets, as evidenced by the data presented in Table 9. Additionally, MossFormer [640] is another cutting-edge architecture that has successfully pushed the boundaries of monaural speech separation across multiple speech separation benchmarks. It is worth noting that although both models employ attention mechanisms, MossFormer integrates a blend of convolutional modules to further amplify its performance.

Diffusion models have been proven to be highly effective in various machine learning tasks related to computer vision, as well as speech-processing tasks. The recent development of DiffSep [467] for speech separation, which is based on score-matching of a stochastic differential equation, has shown competitive performance on the VoiceBank-DEMAND dataset. Additionally, Separate And Diffuse [350], another diffusion-based model that utilizes a pretrained diffusion model, currently represents the state-of-the-art performance in various speech separation benchmarks (refer to Table 9). These advancements demonstrate the significant potential of diffusion models in advancing the field of machine learning and speech processing.

5.11 Spoken Language Understanding

5.11.1 Task Description

Spoken Language Understanding (SLU) is a rapidly developing field that brings together speech processing and natural language processing to help machines comprehend human speech and respond appropriately. The ultimate goal of SLU is to bridge the gap between human and machine understanding. Typically, SLU tasks involve identifying the domain or topic of a spoken utterance, determining the speaker's intent or goal in making the utterance, and filling in any relevant slots or variables associated with that intent. For example, consider the spoken utterance, "*What is the weather like in San Francisco today*?" An SLU system would need to identify the domain (weather), the intent (obtaining current weather information), and the specific slot to be filled (location-San Francisco) to generate an appropriate response. By improving SLU capabilities, we can enable more effective communication between humans and machines, making interactions more natural and efficient.

Data-driven methods are frequently utilized to achieve these tasks, employing large datasets to train models capable of accurately recognizing and interpreting spoken language. Among these methods, machine learning techniques, such as deep neural networks, are widely employed, given their exceptional ability to handle complex and ambiguous speech data. The SLU task may be subdivided into the following categories for greater clarity.

- *Keyword Spotting*: Keyword Spotting (KS) is a technique used in speech processing to identify specific words or phrases within spoken language. It involves analysing audio recordings and detecting instances of pre-defined keywords or phrases. This technique is commonly used in applications such as voice assistants, where the system needs to recognize specific commands or questions from the user.
- *Intent Classification*: Intent Classification (IC) is a spoken language understanding task that involves identifying the intent behind a spoken sentence. It is usually implemented as a pipeline process, with a speech recognition module followed by text processing that classifies the intents. However, end-to-end intent classification using speech has numerous advantages compared to the conventional pipeline approach using AST followed by NLP modules.
- *Slot Filling*: Slot Filling (SF) is a widely used technique in Speech Language Understanding (SLU) that enables the extraction of important information, such as names, dates, and locations, from a user's speech. The process involves identifying the specific pieces of information that are relevant to the user's request and placing them into pre-defined slots. For instance, if a user asks for the weather in a particular city, the system will identify the city name and fill it into the appropriate slot, thereby providing an accurate and relevant response.

5.11.2 Dataset

- Keyword Spotting Datasets:
  - *Coucke et al. [96]*: This dataset is a speech command recognition dataset that consists of 105,000 spoken commands in English, with each command being one of 35 keywords. The dataset is designed to be highly varied and challenging, with a diverse set of speakers and background noise conditions.
  - *Leroy et al. [293]*: This dataset is a federated learning-based keyword spotting dataset, it is composed of data from multiple sources that are trained together without sharing the raw data. The dataset consists of audio recordings from multiple devices and environments, with the goal of improving the robustness of KS across different devices and settings
  - *Auto-KWS* [547]: This dataset is automatically generated using TTS approach. The dataset consists of 1000 keywords spoken by 100 different synthetic voices, with variations in accent, gender, and age.
  - *Speech Commands* [566]: This data is a large-scale dataset for KS task that consists of over 100, 000 spoken commands in English, with each command belonging to 35 different keywords. The dataset is specifically designed to be highly varied and challenging, with a diverse set of speakers and background noises. It is commonly used as a benchmark dataset for KS research.
- Intent Classification and Slot Filling
  - *ATIS* [173]: The Airline Travel Information System (ATIS) dataset is a collection of spoken queries and responses related to airline travel, such as flight reservations, flight status, and airport information. The dataset is annotated with both intent labels (e.g. "flight booking", "flight status inquiry") and slot labels (e.g. depart city, arrival city,

date). The ATIS dataset has been used extensively as a benchmark for natural language understanding models.
- *SNIPS* [97]: SNIPS is a dataset of voice commands designed for building a natural language understanding system. It consists of thousands of examples of spoken requests, each annotated with the intent of the request (e.g. "play music", "set an alarm", etc.). The dataset is widely used for training IC and SF models.
- *Fluent Speech Commands* [343]: It is a dataset of voice commands for controlling smart home devices, such as lights, thermostats, and locks. The dataset consists of over 1,5000 spoken commands, each labeled with the intended devices and action (e.g. "turn on the living room lights", "set the thermostat to 72 degrees"). The dataset is designed to have variations in speaker accent, background noise, and device placement.
- *MIT-Restaurant and MIT-Movie* [328]: These are two datasets created by researchers at MIT for training natural language understanding models from restaurant and movie information requests. The dataset contains spoken and text-based queries, each labeled with the intent of the request (e.g. "find a nearby Italian restaurant"," get information about the movie Inception") and relevant slot information (e.g. restaurant type, movie name, etc). The datasets are widely used for benchmarking natural language understanding models.

5.11.3 Models

- *Keyword Spotting*: The state-of-the-art techniques for keyword spotting in speech involve deep learning models, such as CNNs [454] and transformers [36]. Wav2Keyword is one of the popular model based on Wav2Vec2.0 architecture [471] and have achieved SOTA results on Speech Commands data V1 and V21. Another model that achieves SOTA classification accuracy on the Google Speech commands dataset is Keyword Transformer (KWT) [471]. KWT uses a transformer model and achieves 98.6% and 97.7% accuracy on the 12 and 35-word tasks, respectively. KWT also has low latency and can be used on mobile devices.
- The DIET architecture, as introduced in [47], is a transformer-based multitask model that addresses intent classification and entity recognition simultaneously. DIET allows for seamless integration of various pre-trained embeddings such as BERT, GloVe, and ConveRT. Results from experiments show that DIET outperforms fine-tuned BERT and has the added benefit of being six times faster to train.
- Chang et al. [56] investigated the effectiveness of prompt tuning on the GSLM architecture and showcased its competitiveness on various SLU tasks, such as KS, IC, and SF. Impressively, this approach achieves comparable results with fewer trainable parameters than full fine-tuning. Despite being a popular and effective technique in numerous NLP tasks, prompt tuning has not received much attention in the speech community. Additionally, other researchers have pursued a different path by utilizing pre-trained wav2vec2.0 and different adapters [308] to attain state-of-the-art outcomes.

Despite the remarkable progress made in SLU, accurately comprehending human speech in real-life situations remains a formidable challenge, particularly due to the existence of diverse accents, dialects, and linguistic variations. In this regard, ongoing research endeavors strive to surmount these obstacles and improve the precision and efficacy of SLU systems.

Recent studies, including the comprehensive analysis of the performance of different models and techniques for Keyword Spotting (KS) and Slot Filling (SF) tasks on Google Speech Commands and ATIS benchmark datasets (Table 10), have furnished valuable insights into the strengths and limitations of such approaches in SLU. Capitalizing on these findings and leveraging the latest

Table 10. Comprehensive performance analysis of various models for Keyword Spotting (KS) and Slot Filling (SF) tasks, evaluated on two benchmark datasets: Google Speech Commands for KS and ATIS for SF.

Keyword Spotting on Google Speech Commands (Accuracy % ↑)					Slot Filling on ATIS (F1 ↑)		
Model	Reference	Google Speech Commands V1 12	Google Speech Commands V2 12	Google Speech Commands V2 35	Model	Reference	ATIS
TripletLoss-res15	[538]	98.56	98.37	97.0	CTRAN	[438]	0.9846
Wav2KWS	[471]	97.9	98.5	97.8	Bi-model with a decoder	[558]	0.9689
KWT-3	[36]	97.49 ±0.15	98.56 ±0.07	97.69 ±0.09	Joint BERT	[67]	0.961
KWT-1	[36]	97.27 ±0.08	98.43±0.08	97.74 ±0.03	Joint BERT + CRF	[67]	0.96
KWT-2	[36]	97.26±0.18	98.08±0.10	96.95±0.14	SF-ID	[389]	0.958
Attention RNN	[460]	95.6	96.9	93.9	Capsule-NLU	[618]	0.952

advances in deep learning and speech recognition could help us continue to expand the frontiers of spoken language understanding and drive further innovation in this domain.

5.12 Audio/visual multimodal speech processing

The process of speech perception in humans is intricate and involves multiple sensory modalities, including auditory and visual cues. The generation of speech sounds involves articulators such as the tongue, lips, and teeth, with their movements being critical for producing different speech sounds and also visible to others. The importance of visual cues becomes more pronounced for individuals with hearing impairments who depend on lip-reading to comprehend spoken language, while individuals with normal hearing can also benefit from visual cues in noisy environments.

When investigating language comprehension and communication, it is essential to consider both auditory and visual information, as studies have demonstrated that visual information can assist in distinguishing between acoustically similar sounds that differ in articulatory characteristics. A comprehensive understanding of the interaction between these sensory modalities can lead to the development of assistive technologies for individuals with hearing impairments and enhance communication strategies in challenging listening environments.

5.12.1 Task Description

The tasks under audiovisual multimodal processing can be subdivided into the following categories.

- *Lip-reading*: Lip-reading is a remarkable ability that allows us to comprehend spoken language from silent videos. However, it is a challenging task even for humans. Recent advancements in deep learning technology have enabled the development of neural network-based lip-reading models to accomplish this task with high accuracy. These models take silent facial videos as input and produce the corresponding speech audio or characters as output. The potential applications of automatic lip-reading models are vast and diverse, including enabling videoconferencing in noisy environments, using surveillance videos as long-range listening devices, and facilitating conversations in noisy social settings. Developing these models could significantly improve our daily lives.
- *Audiovisual speech separation*: Audiovisual speech separation has been a subject of interest in recent years due to the human ability to focus auditory attention on a single sound source amidst the noise, known as the "cocktail party effect." This phenomenon has become a typical problem in computer speech recognition, and automatic speech separation is required to separate an input audio signal into individual speech sources. Ephrat et al. [126] proposed that audiovisual speech separation can perform better than audio-only speech separation, as viewing a speaker's face enhances the model's capacity to resolve perceived ambiguity. Automatic speech separation has numerous applications, including assistive

technology for the hearing-impaired and head-mounted assistive devices for noisy meeting scenarios.

- *Talking face generation*: Generating a natural talking face of a target character saying a given speech with lip synchronization while ensuring a smooth transition of facial images is known as talking face generation. This task has gained significant interest and is a challenging problem due to the continuously changing facial region, which depends not only on visual information (given face image) but also on acoustic information (given speech audio) to achieve lip-speech synchronization. Despite the challenges, talking face generation has many potential applications, such as teleconferencing, generating virtual characters with specific facial movements, and enhancing speech comprehension. Significant work related to the talking face generation task has been done in recent years [62, 129, 130, 494, 648].

5.12.2 Datasets

Several datasets are widely used for audiovisual multimodal research, including VoxCeleb, TCD-TIMID [168] , etc. We briefly discuss some of them in the following section.

- *TCD-TIMID [168]:* This is an extensive and diverse audiovisual dataset that encompasses both audio and video recordings of 600 distinct sentences spoken by 60 participants. The dataset features a wide range of speakers with different genders, accents, and backgrounds, making it highly suitable for talker-independent speech recognition research. The audio recordings are of exceptional quality, captured using high-fidelity microphones with a sampling rate of 48kHz. Meanwhile, the video footage is of 720p resolution and includes depth information for every frame
- *LipReading in the Wild (LRW) [89]:* The LRW is a comprehensive audiovisual dataset that encompasses 500 distinct words spoken by more than 1000 speakers. This dataset has been segmented into distinct training, evaluation, and test sets to facilitate efficient research. Additionally, the LRW-1000 dataset [594] represents a subset of LRW, featuring a 1000-word vocabulary. Researchers can benefit from pre-trained weights included with this dataset, simplifying the evaluation process. Overall, these datasets are highly regarded in the scientific community for their size and versatility in supporting research related to speech recognition and natural language processing
- *LRS2 and LRS3* [9]: The LRS2 and LRS3 datasets are additional examples of audiovisual speech recognition datasets that have been gathered from videos captured in real-world settings. Each of these datasets has its own distinct train/test split and includes cropped face tracks as well as corresponding audio clips sourced from British television. Both datasets are considered to be of significant value to researchers in the field of speech recognition, particularly those focused on audiovisual analysis.
- *GRID [93]:* This dataset comprises high-fidelity audio and video recordings of more than 1000 sentences spoken by 34 distinct speakers, including 18 males and 16 females. The sentences were gathered using the prompt "put red at G9 now" and are widely employed in research related to audio-visual speech separation and talking face synthesis. The dataset is considered to be of exceptional quality and is highly sought after in the scientific community.

5.12.3 Models

In recent years, there has been a significant surge in the development of algorithms designed to tackle multimodal tasks. Specifically, researchers have devoted significant attention to developing neural networks tailored for tasks such as Text-to-Speech (TTS) [245, 447–449]. Multimodal

[9]https://www.robots.ox.ac.uk/vgg/data/lip_reading/lrs2.html

processing, which integrates information from both visual and auditory modalities, has been critical in advancing tasks related to our daily lives. For instance, lip-reading, a task that involves transcribing videos of lip movements to characters, has undergone significant advancements in recent years, with or without accompanying audio. One notable work [493] proposes a hybrid model based on CNN, LSTM, and an attention mechanism for generating characters by modeling the correlations between the lip video and audio. The authors also propose a new dataset, LRS, to further aid in developing lip-reading models. Another model for lip-reading, LiRA [352], is trained using a self-supervised architecture. LiRA projects the lip image sequence and audio waveform into high-level representations by calculating the loss between them in the pre-training stage, thereby achieving word-level and sentence-level lip-reading. Ephrat et al. [125] has also proposed a model for capturing human emotions expressed through the acoustic signal. They argue that viewing the task as an acoustic regression problem is advantageous instead of focusing on visual-to-text modeling. Vid2Speech [127] is another CNN-based model that inputs facial image sequences and outputs the corresponding speech raw audio waveform. It uses a two-tower CNN-based model that inputs facial greyscale images and optical flow computed between facial image frames. Finally, other models based on mutual information maximization [644] and spatiotemporal fusion [630] have also been proposed for the lip-reading task.

In an early attempt to develop algorithms for audiovisual speech separation, the authors of [126] proposed a CNN-based architecture that encodes facial images and speech spectrograms to compute a complex mask for speech separation. Additionally, they introduced the AVspeech dataset in this work. AV-CVAE [385] utilizes a conditional VAE to detect the lip movements of the speaker and predict separated speech. In a deviation from speech signals, [377] focuses on audiovisual singing separation and employs a two-stream CNN architecture, Y-Net [367], to process audio and video separately. This work introduces a large dataset of solo singing videos for audiovisual singing separation. The VisualSpeech [146] architecture takes a face image sequence and mixed audio of lip movement as input and predicts a complex mask. It also proposes a cross-modal embedding space to facilitate the correlation of audio and visual modalities. Finally, FaceFilter [90] uses still images as visual information, and other methods for the audiovisual speech separation task are proposed in [9, 141, 372].

The rise of Deepfake videos on the internet has led to a surge in demand for creating realistic talking faces for various applications, such as video production, marketing, and entertainment. Previously, the conventional approach involved manipulating 3D meshes to create specific faces, which was time-consuming and limited to certain identities. However, recent advancements in deep generative models have made significant progress. For example, DAVS [648] introduced an end-to-end trainable deep neural network capable of learning a joint audiovisual representation, which uses adversarial training to disentangle the latent space. Another architecture proposed by ATVGnet [62] consists of an audio transformation network (AT-net) and a visual generation network (VG-net) for processing acoustic and visual information, respectively. This method introduced a regression-based discriminator, a dynamically adjustable pixel-wise loss, and an attention mechanism. In [651], a novel framework for talking face generation was presented, which discovers audiovisual coherence through an asymmetrical mutual information estimator. Furthermore, the authors in [129] proposed an end-to-end approach based on generative adversarial networks that use noisy speech for talking face generation. In addition, alternative methods based on conditional recurrent adversarial networks and speech-driven talking face generation were introduced in [130, 494].

6 Advanced Transfer Learning Techniques for Speech Processing

6.1 Domain Adaptation

6.1.1 Task Description

Domain adaptation is a field that deals with adapting a model trained on a labeled dataset from a source domain to a target domain, where the source domain differs from the target domain. The goal of domain adaptation is to reduce the performance gap between the source and target domains by minimizing the difference between their distributions. In speech processing, domain adaptation has various applications such as speech recognition [43, 83, 194, 285, 387], speaker verification [73, 178, 555, 577, 622], and speech synthesis [579, 608]. This section explores the use of domain adaptation in these tasks by reviewing recent literature on the subject. Specifically, we discuss the techniques used in domain adaptation, their effectiveness, and the challenges that arise when applying them to speech processing.

6.1.2 Models

Various techniques have been proposed to adapt a deep learning model for speech processing tasks. An example of a technique is reconstruction-based domain adaptation, which leverages an additional reconstruction task to generate a communal representation for all the domains. The Deep Reconstruction Classification Network (DRCN) [149] is an illustration of such an approach, as it endeavors to address both tasks concurrently: (i) classification of the source data and (ii) reconstruction of the input data. Another technique used in domain adaptation is the domain-adversarial neural network architecture, which aims to learn domain-invariant features using a gradient reversal layer [50, 551, 631].

Different domain adaptation techniques are successfully applied to different speech processing tasks, such as speaker recognition [43, 194, 306, 387] and verification [72, 73, 299, 622, 650], where the goal is to verify the identity of a speaker using their voice. One approach for domain adaptation in speaker verification is to use adversarial domain training to learn speaker-independent features insensitive to variations in the recording environment [72].

Domain adaptation has also been applied to speech recognition [207, 360, 500, 608] to improve speech recognition accuracy in a target domain. One recent approach for domain adaptation in ASR is prompt-tuning [108], which involves fine-tuning the ASR system on a small amount of data from the new domain. Another approach is to use adapter modules for transducer-based speech recognition systems [357, 464], which can balance the recognition accuracy of general speech and improve recognition on adaptation domains. The Machine Speech Chain integrates both end-to-end (E2E) ASR and neural text-to-speech (TTS) into one circle [608]. This integration can be used for domain adaptation by fine-tuning the E2E ASR on a small amount of data from the new domain and then using the TTS to generate synthetic speech in the new domain for further training.

In addition to domain adaptation techniques used in speech recognition, there has been growing interest in adapting text-to-speech (TTS) models to specific speakers or domains. This research direction is critical, especially in low-resource settings where collecting sufficient training data can be challenging. Several recent works have proposed different approaches for speaker and domain adaptation in TTS, such as AdaSpeech [63, 576, 586].

6.2 Meta Learning

6.2.1 Task Description

Meta-learning is a branch of machine learning that focuses on improving the learning algorithms used for tasks such as parameter initialization, optimization strategies, network architecture, and

distance metrics. This approach has been demonstrated to facilitate faster fine-tuning, better performance convergence, and the ability to train models from scratch, which is especially advantageous for speech-processing tasks. Meta-learning techniques have been employed in various speech-processing tasks, such as low-resource ASR [186, 209], SV [621], TTS [202] and domain generalization for speaker recognition [236].

Meta-learning has the potential to improve speech processing tasks by learning better learning algorithms that can adapt to new tasks and data more efficiently. Meta-learning can also reduce the cost of model training and fine-tuning, which is particularly useful for low-resource speech processing tasks. Further investigation is required to delve into the full potential of meta-learning in speech processing and to develop more effective meta-learning algorithms for different speech-processing tasks.

6.2.2 Models

In low-resource ASR, meta-learning is used to quickly adapt unseen target languages by formulating ASR for different languages as different tasks and meta-learning the initialization parameters from many pretraining languages [186, 486]. The proposed approach, MetaASR [186], significantly outperforms the state-of-the-art multitask pretraining approach on all target languages with different combinations of pretraining languages. In speaker verification, meta-learning is used to improve the meta-learning training for SV by introducing two methods to improve the backbone embedding network [70]. The proposed methods can obtain consistent improvements over the existing meta-learning training framework [272].

Meta-learning has proven to be a promising approach in various speech-related tasks, including low-resource ASR and speaker verification. In addition to these tasks, meta-learning has also been applied to few-shot speaker adaptive TTS and language-agnostic TTS, demonstrating its potential to improve performance across different speech technologies. Meta-TTS [202] is an example of a meta-learning model used for a few-shot speaker adaptive TTS. It can synthesize high-speaker-similarity speech from a few enrolment samples with fewer adaptation steps. Similarly, a language-agnostic meta-learning approach is proposed in [351] for low-resource TTS.

6.3 Parameter-Efficient Transfer Learning

Transfer learning has played a significant role in the recent progress of speech processing. Fine-tuning pre-trained large models, such as those trained on LibriSpeech [401] or Common Voice [16], has been widely used for transfer learning in speech processing. However, fine-tuning all parameters for each downstream task can be computationally expensive. To overcome this challenge, researchers have been exploring parameter-efficient transfer learning techniques that optimize only a fraction of the model parameters, aiming to improve training efficiency. This article investigates these parameter-efficient transfer learning techniques in speech processing, evaluates their effectiveness in improving training efficiency without sacrificing performance, and discusses the challenges and opportunities associated with these techniques, highlighting their potential to advance the field of speech processing.

6.3.1 Adapters

In recent years, retrofitting adapter modules with a few parameters to pre-trained models has emerged as an effective approach in speech processing. This involves optimizing the adapter modules while keeping the pre-trained parameters frozen for downstream tasks. Recent studies (Li et al., 2023; Liu et al., 2021) [308, 592] have shown that adapters often outperform fine-tuning while using only a fraction of the total parameters. Different adapter architectures are available, such as bottleneck adapters (Houlsby et al., 2019)[184], tiny attention adapters (Zhao et al., 2022)[639],

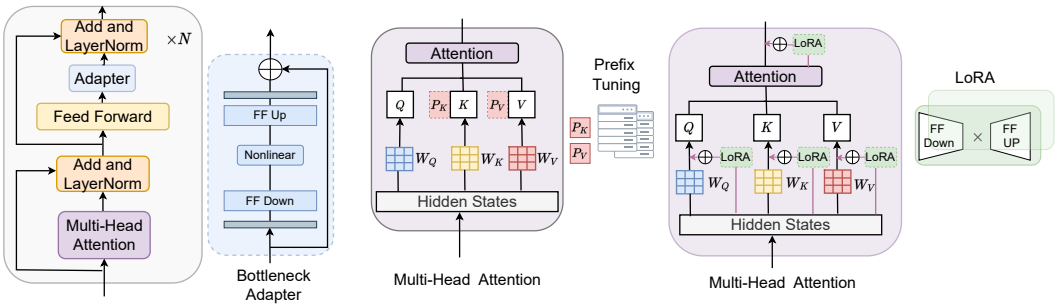

Fig. 17. Transformer architecture and Adapter, Prefix Tuning, and LoRA.

prefix-tuning adapters (Li and Liang, 2021)[307], and LoRA adapters (Hu et al., 2022)[193], among others Next, we will review the different approaches for parameter-efficient transfer learning. The different approaches are illustrated in Figure 17 and Figure 18

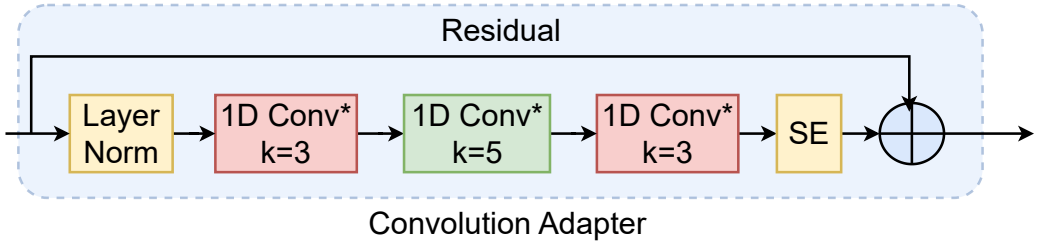

Fig. 18. The architecture of 1D convolution layer-based lightweight adapter. k is the kernel size of 1D convolution. $*$ denotes depth-wise convolution.

Adapter Tuning. Adapters are a type of neural module that can be retrofitted onto a pre-trained language model, with significantly fewer parameters than the original model. One such type is the bottleneck or standard adapter (Houlsby et al., 2019; Pfeiffer et al., 2020) [183, 413]. The adapter takes an input vector $h \in \mathbf{R}^d$ and down-projects it to a lower-dimensional space with dimensionality m (where $m < d$), applies a non-linear function $g(\cdot)$, and then up-projects the result back to the original d-dimensional space. Finally, the output is obtained by adding a residual connection.

$$\boldsymbol{h} \leftarrow \boldsymbol{h} + g(\boldsymbol{h}\boldsymbol{W}_{\text{down}})\boldsymbol{W}_{\text{up}} \tag{30}$$

where matrices $\boldsymbol{W}_{\text{down}}$ and $\boldsymbol{W}_{\text{up}}$ are used as down and up projection matrices, respectively, with $\boldsymbol{W}$down having dimensions $\mathbb{R}^{d\times m}$ and $\boldsymbol{W}_{\text{up}}$ having dimensions $\mathbb{R}^{m\times d}$. Previous studies have empirically shown that a two-layer feedforward neural network with a bottleneck is effective. In this work, we follow the experimental settings outlined in [413] for the adapter, which is inserted after the feedforward layer of every transformer module, as depicted in Figure 17.

Prefix tuning. Recent studies have suggested modifying the attention module of the Transformer model to improve its performance in natural language processing tasks. This approach involves adding learnable vectors to the pre-trained multi-head attention keys and values at every layer, as depicted in Figure 17. Specifically, two sets of learnable prefix vectors, $\boldsymbol{P_K}$ and $\boldsymbol{P_V}$, are concatenated with the original key and value matrices $\boldsymbol{K}$ and $\boldsymbol{V}$, while the query matrix $\boldsymbol{Q}$ remains unchanged.

The resulting matrices are then used for multi-head attention, where each head of the attention mechanism is computed as follows:

$$\text{head}_i = \text{Attn}(\boldsymbol{Q}\boldsymbol{W}_Q^{(i)}, [\boldsymbol{P}_K^{(i)}, \boldsymbol{K}\boldsymbol{W}_Q^{(i)}], [\boldsymbol{P}_V^{(i)}, \boldsymbol{V}\boldsymbol{W}_Q^{(i)}]) \tag{31}$$

where Attn(·) is scaled dot-product attention given by:

$$\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{32}$$

The attention heads in each layer are modified by prefix tuning, with only the prefix vectors $\boldsymbol{P}K$ and $\boldsymbol{P}V$ being updated during training. This approach provides greater control over the transmission of acoustic information between layers and effectively activates the pre-trained model's knowledge.

LoRA. LoRA is a novel approach proposed by Hu et al. (2021) [192], which aims to approximate weight updates in the Transformer by injecting trainable low-rank matrices into its layers. In this method, a pre-trained weight matrix $W \in \mathbb{R}^{d\times k}$ is updated by a low-rank decomposition $\boldsymbol{W} + \Delta\boldsymbol{W} = \boldsymbol{W} + \boldsymbol{W}\text{down}\boldsymbol{W}\text{up}$, where $\boldsymbol{W}\text{down} \in \mathbb{R}^{d\times r}$, $\boldsymbol{W}\text{up} \in \mathbb{R}^{r\times k}$ are tunable parameters and r represents the rank of the decomposition matrices, with $r < d$. Specifically, for a given input $\boldsymbol{x}$ to the linear projection in the multi-headed attention layer, LoRA modifies the projection output $\boldsymbol{h}$ as follows:

$$\boldsymbol{h} \leftarrow \boldsymbol{h} + s \cdot \boldsymbol{x}\boldsymbol{W}_{\text{down}}\boldsymbol{W}_{\text{up}} \tag{33}$$

In this work, LoRA is integrated into four locations of the multi-head attention layer, as illustrated in Figure 17. Thanks to its lightweight nature, the pre-trained model can accommodate many small modules for different tasks, allowing for efficient task switching by replacing the modules. Additionally, LoRA incurs no inference latency and achieves a convergence rate that is comparable to that of training the original model, unlike fully fine-tuned models [192].

Convolutional Adapter. CNNs have become increasingly popular in the field of speech processing due to their ability to learn task-specific information and combine channel-wise information within local receptive fields. To further improve the efficiency of CNNs for speech processing tasks, Li et al. (2023) [308] proposed a lightweight adapter, called the ConvAdapter, which uses three 1D convolutional layers, layer normalization, and a squeeze-and-excite module (Zhang et al., 2017) [195], as shown in Figure 18. By utilizing depth-wise convolution, which requires fewer parameters and is more computationally efficient, the authors were able to achieve better performance while using fewer resources. In this approach, the ConvAdapter is added to the same location as the Bottleneck Adapter (Figure 17).

Table 11, Table 12, and Table 13 present the results of various speech processing tasks in the SURE benchmark. The findings demonstrate that the adapter-based methods perform comparably well in fine-tuning. However, there is no significant advantage of any particular adapter type over others for these benchmark tasks and datasets.

6.3.2 Knowledge Distillation (KD)

Knowledge distillation involves training a smaller model to mimic the behavior of a larger and more complex model. This can be done by training the smaller model to predict the outputs of the larger model or, by using, the larger model's hidden representations as input to the smaller model. Knowledge distillation is effective in reducing the computational cost of training and inference.

Cho et al. [77] conducted knowledge distillation (KD) by directly applying it to the downstream task. One way to improve this approach is to use KD as pre-training for various downstream tasks, thus allowing for knowledge reuse. A noteworthy result achieved by Denisov and Vu [103] was using KD in pretraining. However, they achieved this by initializing an utterance encoder with a

Table 11. The study evaluated various parameter-efficient training methods on pre-trained Word2Vec 2.0, including full fine-tuning, on the SURE benchmark. The fraction of trainable parameters were represented by percentages, with the number of KS task's trainable parameters given. Results are reported using weighted-f1 as the metric (w-f1) on MELD, with the best performance in bold and the second best underlined. To avoid data imbalance, the researchers opted for using weighted-f1 as the metric. The study cites Li et al. (2023) [308] as a reference.

Method	#Parameters	SER (acc % / w-f1) ↑		SR (acc %) ↑		ASR (wer) ↓			KS (acc %) ↑
		ESD	MELD	ESD	VCTK	ESD	FLEURS	LS	Speech Command
Fine Tuning	315,703,947	**96.53**	42.93	99.00	92.36	0.2295	**0.135**	**0.0903**	99.08
Adapter	25,467,915 (8.08%)	94.07	41.58	98.87	96.32	0.2290	0.214	0.2425	99.19
Prefix Tuning	1,739,787 (**0.55%**)	90.00	44.21	**99.73**	**98.49**	**0.2255**	0.166	0.1022	98.86
LoRA	3,804,171 (1.20%)	90.00	**47.05**	99.00	97.61	0.2428	0.149	0.1014	98.28
ConvAdapter	2,952,539 (0.94%)	91.87	46.30	99.60	97.61	0.2456	0.2062	0.2958	**98.99**

Table 12. Results on SURE benchmark for full fine-tuning and other parameter-efficient training methods on pre-trained Wav2Vec 2.0 for IC and PR tasks on **FS**: Fluent Speech [343] and **LS**: LibriSpeech [401] datasets, respectively.

Method	IC		PR		SF		
	#Parameters	FS	#Parameters	LS	#Parameters	SNIPS	
		ACC% ↑		PER ↓		F1 % ↑	CER ↓
Fine-Tuning	315707288	99.60	311304394	0.0577	311375119	93.89	0.1411
Adapter	25471256 (8.06%)	99.39	25278538 (8.01%)	0.1571	25349263 (8.14%)	92.60	0.1666
Prefix Tuning	1743128 (0.55%)	93.43	1550410 (0.49%)	0.1598	1621135 (0.50%)	62.32	0.6041
LoRA	3807512 (1.20%)	99.68	3614794 (1.16%)	0.1053	3685519 (1.18%)	90.61	0.2016
ConvAdapter	3672344 (1.16%)	95.60	3479626 (1.11%)	0.1532	3550351 (1.14%)	59.27	0.6405

trained ASR model's backbone, followed by a trained NLU backbone. Knowledge distillation can be applied directly into a wav2vec 2.0 encoder without ASR training and a trained NLU module to enhance this method. Kim et al. [250] implemented a more complex architecture, utilizing KD in both the pretraining and fine-tuning stages.

6.3.3 Model Compression

Researchers have also explored various architectural modifications to existing models to make them more parameter-efficient. One such approach is *pruning* [136, 563], where motivated by lottery-ticket hypothesis (LTH) [135], the task-irrelevant parameters are masked based on some threshold

Table 13. Results on the SURE benchmark for the TTS task. MCD and WER are the metrics used to compare fine-tuning and other parameter-efficient approaches.

Method	Parameters (%)	LTS		L2ARCTIC	
		MCD ↓	WER ↓	MCD ↓	WER ↓
Fine-tuning	35802977	6.2038	0.2655	6.71469	0.2141
Adapter	659200	6.1634	0.3143	6.544	0.2504
Prefix	153600	6.2523	0.3334	7.4264	0.3244
LoRA	81920	6.8319	0.3786	7.0698	0.3291
Convadapter	108800	6.9202	0.3365	6.9712	0.3227

defined by importance score, such as some parameter norm. Another form of compression could be *low-rank factorization* [191], where the parameter matrices are factorized into lower-rank matrices with much fewer parameters. Finally, *quantization* is a popular approach to reduce the model size and improve energy efficiency with a minimal performance penalty. It involves transforming 32-bit floating point model weights into integers with fewer bit-counts [596]—8-bit, 4-bit, 2-bit, and even 1-bit—through scaling and shifting. At the same time, the quantization of the activation is also handled based on the input.

Lai et al. [273] iteratively prune and subsequently fine-tune wav2vec2.0 on downstream tasks to obtained improved results over fine-tuned wav2vec2.0. Winata et al. [570] employ low-rank transformers to excise the model size by half and increase the inference speed by 1.35 times. Peng et al. [412] employ KD and quantization to make wav2vec2.0 twice as fast, twice as energy efficient, and 4.8 times smaller at the cost of a 7% increase in WER. Without the KD step, the model is 3.6 times smaller with mere 0.1% WER degradation.

7 Conclusion and Future Research Directions

The rapid advancements in deep learning techniques have revolutionized speech processing tasks, enabling significant progress in speech recognition, speaker recognition, and speech synthesis. This paper provides a comprehensive review of the latest developments in deep learning techniques for speech-processing tasks. We begin by examining the early developments in speech processing, including representation learning and HMM-based modeling, before presenting a concise summary of fundamental deep learning techniques and their applications in speech processing. Furthermore, we discuss key speech-processing tasks, highlight the datasets used in these tasks, and present the latest and most relevant research works utilizing deep learning techniques.

We envisage several lines of development in speech processing:

(1) *Large Speech Models:* Beyond wav2vec2.0, larger and richer speech-to-text and text-to-speech (TTS) models with larger datasets. TTS models with more natural and human-like prosody could be built using adversarial training, with a discriminator to tell machine-generated speech apart from the reference speech.

(2) *Multilingual Models:* Self-supervised learning has revolutionized speech recognition, especially for low-resource languages where labeled datasets are either limited or unavailable. The latest self-supervised speech recognition model, XLS-R, is the largest to date with over 2 billion parameters and has been trained on 128 languages, which is more than twice the number of its predecessor. Importantly, scaling up larger multilingual models like XLS-R leads to significant performance gains, making them a more favorable option than single-language models in the future
(3) *Multimodal Speech Models:* Most speech and text models work with only one modality in their input and output. However, with the ever-increasing scale of the large generative models, mixed modality input and output is a natural evolutionary step. Recently revealed LLMs, like GPT-4 [396] and Kosmos-I [201], can jointly work with images and text as their inputs. Thus, large multimodal models for speech and other modalities are likely soon to emerge.
(4) *In-Context Learning:* Based on such mixed-modality models, in-context learning could be developed for a myriad of speech tasks where the tasks could be defined in the input, along with examples. Such works have been shown to work well in LLMs, such as InstructGPT [397], FLAN-T5 [86], LLaMA [514], etc.
(5) *Controllable Speech Generation:* A special case of the above point could be controllable TTS where the various attributes—such as tone, accent, age, gender, etc—of the generated speech can be controlled with in-context text guidance.
(6) *Parameter-efficient Learning:* As the growing size of the LLMs and speech models, there is an urgent need to adapt these models with as few parameter updates as possible. Thus, there is a need for specialized adapters that can update these emerging mixed-modality large models. On the other hand, model compression [273, 412, 570] is another practical way to deal with these large models, as many of these are shown to be very sparse, especially for particular tasks.
(7) *Explainability:* Explainability remains elusive to these large networks as they grow. Researchers are steadfast in explaining these networks' functioning and learning dynamics. Recently, much work has been done to learn the fine-tuning and in-context learning dynamics of these large models for text under the neural-tangent-kernel (NTK) asymptotic framework [359]. Such exploration is yet to be done in the speech domain. More yet, explainability could be built-in as inductive bias in architecture. To this end, brain-inspired architectures [374] are being developed, which may shed more light on this aspect of large models.
(8) *Neuroscience-inspired Architectures:* Recently, much study has been devoted to comparing speech-processing architectures with the functioning of human brains [374]. These works show a high correlation between the layers of speech models and functional hierarchy in human brains. This has inspired many neuroscience-based speech models that perform comparably to the SOTA models [374].
(9) *Text-to-Audio Models for Text-to-Speech:* Lately, transformer- and diffusion-based text-to-audio (TTA) model development is turning into an exciting area of research. Until recently, most of these models [150, 265, 325, 557, 588] overlooked speech in favour of general audio. In the future, however, the models will likely strive to be equally performant in both audio and speech. To that end, current TTS methods will likely be an integral part of those models. Recently, Suno-AI [504] have aimed at striking a good balance between general audio and speech, although their implementation is not public, nor have they provided any detailed paper.

References

[1] 2022. Conformer-1. *AssemblyAI* (2022). https://www.assemblyai.com/blog/conformer-1/

[2] 2022. Speech Recognition With Conformer. *Nvidia* (2022). https://docs.nvidia.com/tao/tao-toolkit/text/asr/speech_recognition_with_conformer.html

[3] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing* 22, 10 (2014), 1533–1545.

[4] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 10 (2014), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

[5] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. 2012. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4277–4280. https://doi.org/10.1109/ICASSP.2012.6288864

[6] Osama Abdeljaber, Onur Avci, Serkan Kiranyaz, Moncef Gabbouj, and Daniel J Inman. 2017. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *Journal of Sound and Vibration* 388 (2017), 154–170.

[7] Sherif Abdulatif, Ruizhe Cao, and Bin Yang. 2022. CMGAN: Conformer-Based Metric-GAN for Monaural Speech Enhancement. *arXiv preprint arXiv:2209.11112* (2022).

[8] Sivanand Achanta, Albert Antony, Ladan Golipour, Jiangchuan Li, Tuomo Raitio, Ramya Rasipuram, Francesco Rossi, Jennifer Shi, Jaimin Upadhyay, David Winarsky, et al. 2021. On-device neural speech synthesis. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1155–1161.

[9] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121* (2018).

[10] Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote. 2020. Using Vaes and Normalizing Flows for One-Shot Text-To-Speech Synthesis of Expressive Speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6179–6183. https://doi.org/10.1109/ICASSP40776.2020.9053678

[11] Waleed Alsabhan. 2023. Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention. *Sensors* 23, 3 (2023), 1386.

[12] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR, 173–182.

[13] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing* 20, 2 (2012), 356–370.

[14] Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, et al. 2020. Findings of the IWSLT 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*. 1–34.

[15] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205* (2021).

[16] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[17] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*. PMLR, 195–204.

[18] Kartik Audhkhasi, George Saon, Zoltán Tüske, Brian Kingsbury, and Michael Picheny. 2019. Forget a Bit to Learn Better: Soft Forgetting for CTC-Based Automatic Speech Recognition.. In *Interspeech*. 2618–2622.

[19] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296* (2021).

[20] Rohan Badlani, Adrian Łańcucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2022. One TTS alignment to rule them all. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6092–6096.

[21] Rohan Badlani, Adrian Łańcucki, Kevin J. Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2022. One TTS Alignment to Rule Them All. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP). 6092–6096. https://doi.org/10.1109/ICASSP43922.2022.9747707

[22] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912* (2019).

[23] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*. PMLR, 1298–1312.

[24] Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453* (2019).

[25] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020).

[26] Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 792–799.

[27] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[28] He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. 2022. A3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing. In *International Conference on Machine Learning*. PMLR, 1399–1411.

[29] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[30] Zhongxin Bai and Xiao-Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* 140 (2021), 65–99.

[31] Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, and Young-Sun Joo. 2022. Avocodo: Generative adversarial network for artifact-free vocoder. *arXiv preprint arXiv:2206.13404* (2022).

[32] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth'CHiME'speech separation and recognition challenge: dataset, task and baselines. *arXiv preprint arXiv:1803.10609* (2018).

[33] Murali Karthick Baskar, Shinji Watanabe, Ramon Astudillo, Takaaki Hori, Lukáš Burget, and Jan Černockỳ. 2019. Semi-supervised sequence-to-sequence ASR using unpaired speech and text. *arXiv preprint arXiv:1905.01152* (2019).

[34] Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. 2020. Location-relative attention mechanisms for robust long-form speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6194–6198.

[35] John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. 2013. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society* 61, 6 (2013), 366–384.

[36] Axel Berg, Mark O'Connor, and Miguel Tairum Cruz. 2021. Keyword transformer: A self-attention model for keyword spotting. *arXiv preprint arXiv:2104.00769* (2021).

[37] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. [n. d.]. High Fidelity Speech Synthesis with Adversarial Networks. In *International Conference on Learning Representations*.

[38] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. 2019. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646* (2019).

[39] Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei W Koh, and Stefano Ermon. 2019. Temporal FiLM: Capturing Long-Range Sequence Dependencies with Feature-Wise Modulations. *Advances in Neural Information Processing Systems* 32 (2019).

[40] Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* 27, 2 (1979), 113–120.

[41] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[42] Herve A Bourlard and Nelson Morgan. 1994. *Connectionist speech recognition: a hybrid approach*. Vol. 247. Springer Science & Business Media.

[43] Pierre-Michel Bousquet and Mickael Rouvier. 2019. On robustness of unsupervised domain adaptation for speaker recognition. In *Interspeech*.

[44] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. *arXiv preprint arXiv:2104.04045* (2021).

[45] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[46] Latané Bullock, Hervé Bredin, and Leibny Paola Garcia-Perera. 2020. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7114–7118.

[47] Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936* (2020).

[48] Maxime Burchi and Radu Timofte. 2023. Audio-Visual Efficient Conformer for Robust Speech Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2258–2267.

[49] Alena Butryna, Shan-Hui Cathy Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, et al. 2020. Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: an overview. *arXiv preprint arXiv:2010.06778* (2020).

[50] Anoop C S, Prathosh A P, and A G Ramakrishnan. 2021. Unsupervised Domain Adaptation Schemes for Building ASR in Low-Resource Languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 342–349. https://doi.org/10.1109/ASRU51503.2021.9688269

[51] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*. Springer, 28–39.

[52] Andrew A Catellier and Stephen D Voran. 2020. Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 331–335.

[53] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language* 66 (2021), 101155.

[54] Benjamin Cauchi, Kai Siedenburg, Joao F Santos, Tiago H Falk, Simon Doclo, and Stefan Goetze. 2019. Non-intrusive speech quality prediction using modulation energies and lstm-network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 7 (2019), 1151–1163.

[55] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. 2021. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133* (2021).

[56] Kai-Wei Chang et al. 2022. An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks. *arXiv preprint arXiv:2203.16773* (2022).

[57] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 5 (2021), 1–32.

[58] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909* (2021).

[59] Junkun Chen, Mingbo Ma, Renjie Zheng, and Liang Huang. 2020. Mam: Masked acoustic modeling for end-to-end speech-to-text translation. *arXiv preprint arXiv:2010.11445* (2020).

[60] Junkun Chen, Mingbo Ma, Renjie Zheng, and Liang Huang. 2021. SpecRec: An Alternative Solution for Improving End-to-End Speech-to-Text Translation via Spectrogram Reconstruction.. In *Interspeech*. 2232–2236.

[61] Jiawei Chen, Xu Tan, Yichong Leng, Jin Xu, Guihua Wen, Tao Qin, and Tie-Yan Liu. 2021. Speech-t: Transducer for text to speech and beyond. *Advances in Neural Information Processing Systems* 34 (2021), 6621–6633.

[62] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7832–7841.

[63] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2021. Adaspeech: Adaptive text to speech for custom voice. *arXiv preprint arXiv:2103.00993* (2021).

[64] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. [n. d.]. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*.

[65] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. WaveGrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713* (2020).

[66] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. 2021. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660* (2021).

[67] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909* (2019).

[68] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[69] Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, et al. 2022. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6152–6156.

[70] Yafeng Chen, Wu Guo, and Bin Gu. 2021. Improved meta-learning training for speaker verification. *arXiv preprint arXiv:2103.15421* (2021).

[71] Zehua Chen, Xu Tan, Ke Wang, Shifeng Pan, Danilo Mandic, Lei He, and Sheng Zhao. 2022. Infergrad: Improving Diffusion Models for Vocoder by Considering Inference in Training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8432–8436.

[72] Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2020. Adversarial Domain Adaptation for Speaker Verification Using Partially Shared Network.. In *Interspeech*. 3017–3021.

[73] Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2021. Self-supervised learning based domain adaptation for robust speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5834–5838.

[74] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey. 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[75] Chung-Cheng Chiu and Colin Raffel. 2017. Monotonic chunkwise attention. *arXiv preprint arXiv:1712.05382* (2017).

[76] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia* 17, 11 (2015), 1875–1886.

[77] Won Ik Cho, Donghyun Kwak, J. Yoon, and Nam Soo Kim. 2020. Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation. In *Interspeech*.

[78] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems* 34 (2021), 16251–16265.

[79] Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim. 2022. NANSY++: Unified Voice Synthesis with Neural Analysis and Synthesis. *arXiv preprint arXiv:2211.09407* (2022).

[80] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron Van Den Oord. 2019. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing* 27, 12 (2019), 2041–2053.

[81] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems* 28 (2015).

[82] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742* (2019).

[83] Anurag Chowdhury, Austin Cozzo, and Arun Ross. 2022. Domain Adaptation for Speaker Recognition in Singing and Spoken Voice. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7192–7196.

[84] Hoon Chung, Hyeong-Bae Jeon, and Jeon Gue Park. 2020. Semi-supervised training for sequence-to-sequence speech recognition using reinforcement learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.

[85] Hoon Chung, Hyeong-Bae Jeon, and Jeon Gue Park. 2020. Semi-supervised Training for Sequence-to-Sequence Speech Recognition Using Reinforcement Learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–6. https://doi.org/10.1109/IJCNN48605.2020.9207023

[86] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *ArXiv* abs/2210.11416 (2022).

[87] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. 2020. In defence of metric learning for speaker recognition. *arXiv preprint arXiv:2003.11982* (2020).

[88] J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

[89] Joon Son Chung and Andrew Zisserman. 2017. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer, 87–103.

[90] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. 2020. Facefilter: Audio-visual speech separation using still images. *arXiv preprint arXiv:2005.07074* (2020).

[91] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240* (2019).

[92] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 244–250.

[93] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.

[94] Juan M Coria, Hervé Bredin, Sahar Ghannay, and Sophie Rosset. 2021. Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1139–1146.

[95] Marvin Coto-Jiménez. 2019. Improving post-filtering of artificial speech using pre-trained LSTM neural networks. *Biomimetics* 4, 2 (2019), 39.

[96] Alice Coucke, Mohammed Chlieh, Thibault Gisselbrecht, David Leroy, Mathieu Poumeyrol, and Thibaut Lavril. 2019. Efficient keyword spotting using dilated convolutions and gating. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6351–6355.

[97] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190* (2018).

[98] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*. PMLR, 933–941.

[99] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[100] Keqi Deng, Songjun Cao, Yike Zhang, and Long Ma. 2021. Improving Hybrid CTC/Attention End-to-End Speech Recognition with Pretrained Acoustic and Language Models. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 76–82. https://doi.org/10.1109/ASRU51503.2021.9688009

[101] Keqi Deng, Songjun Cao, Yike Zhang, Long Ma, Gaofeng Cheng, Ji Xu, and Pengyuan Zhang. 2022. Improving CTC-Based Speech Recognition Via Knowledge Transferring from Pre-Trained Language Models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8517–8521. https://doi.org/10.1109/ICASSP43922.2022.9747887

[102] Keqi Deng, Songjun Cao, Yike Zhang, Long Ma, Gaofeng Cheng, Ji Xu, and Pengyuan Zhang. 2022. Improving CTC-based speech recognition via knowledge transferring from pre-trained language models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8517–8521.

[103] Pavel Denisov and Ngoc Thang Vu. 2020. Pretrained Semantic Speech Embeddings for End-to-End Spoken Language Understanding via Cross-Modal Teacher-Student Learning. In *Interspeech*.

[104] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143* (2020).

[105] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[106] Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting transformer to end-to-end spoken language translation. In *Proceedings of INTERSPEECH 2019*. International Speech Communication Association (ISCA), 1133–1137.

[107] Mireia Diez, Lukáš Burget, Federico Landini, Shuai Wang, and Honza Černockỳ. 2020. Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6519–6523.

[108] Saket Dingliwal, Ashish Shenoy, Sravan Bodapati, Ankur Gandhe, Ravi Teja Gadde, and Katrin Kirchhoff. 2021. Prompt-tuning in ASR systems for efficient domain-adaptation. *arXiv preprint arXiv:2110.06502* (2021).

[109] Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208* (2018).

[110] Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen, and Karen Simonyan. [n. d.]. End-to-end Adversarial Text-to-Speech. In *International Conference on Learning Representations*.

[111] Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. 2020. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575* (2020).

[112] Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5884–5888. https://doi.org/10.1109/ICASSP.2018.8462506

[113] Peiyan Dong, Siyue Wang, Wei Niu, Chengming Zhang, Sheng Lin, Zhengang Li, Yifan Gong, Bin Ren, Xue Lin, and Dingwen Tao. 2020. Rtmobile: Beyond real-time mobile acceleration of rnns for speech recognition. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[114] Xuan Dong and Donald S Williamson. 2020. An attention enhanced multi-task model for objective speech assessment in real-world environments. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 911–915.

[115] Xuan Dong and Donald S Williamson. 2020. A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals. *arXiv preprint arXiv:2007.15797* (2020).

[116] Shaked Dovrat, Eliya Nachmani, and Lior Wolf. 2021. Many-speakers single channel speech separation with optimal permutation training. *arXiv preprint arXiv:2104.08955* (2021).

[117] Jennifer Drexler and James Glass. 2019. Explicit alignment of text and speech encodings for attention-based end-to-end speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 913–919.

[118] Chenpeng Du and Kai Yu. 2022. Phone-Level Prosody Modelling With GMM-Based MDN for Diverse and Controllable Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 190–201. https://doi.org/10.1109/TASLP.2021.3133205

[119] Jun Du, Qing Wang, Tian Gao, Yong Xu, Li-Rong Dai, and Chin-Hui Lee. 2014. Robust speech recognition with speech enhanced deep neural networks. In *Fifteenth annual conference of the international speech communication association*.

[120] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2735–2744.

[121] Jide S Edu, Jose M Such, and Guillermo Suarez-Tangil. 2020. Smart home personal assistants: a security and privacy review. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–36.

[122] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J Weiss, and Yonghui Wu. 2021. Parallel tacotron: Non-autoregressive and controllable tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5709–5713.

[123] Ashraf Elneima and Mikołaj Bińkowski. 2022. Adversarial Text-to-Speech for low-resource languages. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*. 76–84.

[124] Yariv Ephraim. 1992. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Transactions on Signal Processing* 40, 4 (1992), 725–735.

[125] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 455–462.

[126] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619* (2018).

[127] Ariel Ephrat and Shmuel Peleg. 2017. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5095–5099.

[128] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine* 39, 3 (2022), 42–62.

[129] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2020. End-to-end generation of talking faces from noisy speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 1948–1952.

[130] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. 2021. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia* 24 (2021), 3480–3490.

[131] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang. 2020. Cn-celeb: a challenging chinese speaker recognition dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7604–7608.

[132] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Fifteenth annual conference of the international speech communication association*.

[133] Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking synthetic data for speech recognition. *arXiv preprint arXiv:2106.07803* (2021).

[134] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 829–852.

[135] Jonathan Frankle and Michael Carbin. 2018. The Lottery Ticket Hypothesis: Training Pruned Neural Networks. *CoRR* abs/1803.03635 (2018). arXiv:1803.03635 http://arxiv.org/abs/1803.03635

[136] Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot. *ArXiv* abs/2301.00774 (2023).

[137] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*. PMLR, 2031–2041.

[138] Szu-Wei Fu, Yu Tsao, and Xugang Lu. 2016. SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement.. In *Interspeech*. 3768–3772.

[139] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. 2021. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603* (2021).

[140] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 296–303.

[141] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. 2017. Visual speech enhancement. *arXiv preprint arXiv:1711.08789* (2017).

[142] Adam Gabryś, Goeric Huybrechts, Manuel Sam Ribeiro, Chung-Ming Chien, Julian Roth, Giulia Comini, Roberto Barra-Chicote, Bartek Perz, and Jaime Lorenzo-Trueba. 2022. Voice Filter: Few-shot text-to-speech speaker adaptation using voice conversion as a post-processing module. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7902–7906.

[143] Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE transactions on neural networks and learning systems* 32, 10 (2020), 4291–4308.

[144] Mark Gales, Steve Young, et al. 2008. The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing* 1, 3 (2008), 195–304.

[145] Chenyang Gao, Yue Gu, Francesco Caliva, and Yuzong Liu. 2023. Self-supervised speech representation learning for keyword-spotting with light-weight transformers. *arXiv preprint arXiv:2303.04255* (2023).

[146] Ruohan Gao and Kristen Grauman. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 15490–15500.

[147] Daniel Garcia-Romero, Alan McCree, David Snyder, and Gregory Sell. 2020. JHU-HLTCOE system for the VoxSRC speaker recognition challenge. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7559–7563.

[148] John S Garofolo. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993* (1993).

[149] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 597–613.

[150] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model. arXiv:2304.13731 [eess.AS]

[151] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems* 30 (2017).

[152] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. 2019. Attention wave-u-net for speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 249–253.

[153] Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711* (2012).

[154] Alex Graves and Alex Graves. 2012. Connectionist temporal classification. *Supervised sequence labelling with recurrent neural networks* (2012), 61–93.

[155] Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*. PMLR, 1764–1772.

[156] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 6645–6649.

[157] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).

[158] Haohan Guo, Fenglong Xie, Frank K Soong, Xixin Wu, and Helen Meng. 2022. A Multi-Stage Multi-Codebook VQ-VAE Approach to High-Performance Neural TTS. *arXiv preprint arXiv:2209.10887* (2022).

[159] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

[160] Mohamed Farouk Abdel Hady and Friedhelm Schwenker. 2013. Semi-supervised learning. *Handbook on Neural Information Processing* (2013), 215–239.

[161] Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. *arXiv preprint arXiv:2105.03095* (2021).
[162] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 87–110.
[163] Seungu Han and Junhyeok Lee. 2022. NU-Wave 2: A general neural audio upsampling model for various sampling rates. *arXiv preprint arXiv:2206.08545* (2022).
[164] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191* (2020).
[165] Rafizah Mohd Hanifa, Khalid Isa, and Shamsul Mohamad. 2021. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering* 90 (2021), 107005.
[166] Peter SK Hansen. 1997. *Signal subspace methods for speech enhancement.* Ph. D. Dissertation. Citeseer.
[167] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. 2021. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6633–6637.
[168] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615.
[169] Simon Haykin and Zhe Chen. 2005. The cocktail party problem. *Neural computation* 17, 9 (2005), 1875–1902.
[170] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[171] Yanzhang He, Rohit Prabhavalkar, Kanishka Rao, Wei Li, Anton Bakhtin, and Ian McGraw. 2017. Streaming small-footprint keyword spotting using sequence-to-sequence models. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 474–481.
[172] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. 2019. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6381–6385.
[173] Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.
[174] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJz6tiCqYm
[175] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 31–35.
[176] Yosuke Higuchi, Brian Yan, Siddhant Arora, Tetsuji Ogawa, Tetsunori Kobayashi, and Shinji Watanabe. 2022. BERT Meets CTC: New Formulation of End-to-End Speech Recognition with Pre-trained Masked Language Model. *arXiv preprint arXiv:2210.16663* (2022).
[177] Bertrand Higy and Peter Bell. 2018. Few-shot learning with attention-based sequence-to-sequence models. *arXiv preprint arXiv:1811.03519* (2018).
[178] Ivan Himawan, Fernando Villavicencio, Sridha Sridharan, and Clinton Fookes. 2019. Deep domain adaptation for anti-spoofing in speaker verification systems. *Computer Speech & Language* 58 (2019), 377–402.
[179] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.
[180] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.
[181] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[182] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. 2018. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence* 2, 2 (2018), 117–128.
[183] Neil Houlsby, , et al. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. 2790–2799.
[184] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2790–2799. https://proceedings.mlr.press/v97/houlsby19a.html

[185] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. 2017. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849* (2017).

[186] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. Meta learning for end-to-end low-resource speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7844–7848.

[187] Wei-Ning Hsu et al. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[188] Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. 2022. ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Enhancement. *arXiv preprint arXiv:2212.11377* (2022).

[189] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. 2019. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5901–5905.

[190] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al. [n. d.]. Hierarchical Generative Modeling for Controllable Speech Synthesis. In *International Conference on Learning Representations*.

[191] Yen-Chang Hsu, Ting Hua, Sung-En Chang, Qiang Lou, Yilin Shen, and Hongxia Jin. 2022. Language model compression with weighted low-rank factorization. *ArXiv* abs/2207.00112 (2022).

[192] Edward J Hu et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[193] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[194] Hang-Rui Hu, Yan Song, Ying Liu, Li-Rong Dai, Ian McLoughlin, and Lin Liu. 2022. Domain Robust Deep Embedding Learning for Speaker Recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7182–7186.

[195] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2017. Squeeze-and-Excitation Networks. https://doi.org/10.48550/ARXIV.1709.01507

[196] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264* (2020).

[197] Yi Hu and Philipos C Loizou. 2007. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing* 16, 1 (2007), 229–238.

[198] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934* (2022).

[199] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Advances in Neural Information Processing Systems* 35 (2022), 10970–10983.

[200] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2595–2605.

[201] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *ArXiv* abs/2302.14045 (2023).

[202] Sung-Feng Huang, Chyi-Jiunn Lin, Da-Rong Liu, Yi-Chen Chen, and Hung-yi Lee. 2022. Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1558–1571.

[203] Wen-Chin Huang, Tomoki Hayashi, Xinjian Li, Shinji Watanabe, and Tomoki Toda. 2021. On prosody modeling for ASR+ TTS based voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 642–649.

[204] Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, and Tomoki Toda. 2019. Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. *arXiv preprint arXiv:1912.06813* (2019).

[205] Zhiying Huang, Hao Li, and Ming Lei. 2020. Deviceits: A small-footprint, fast, stable network for on-device text-to-speech. *arXiv preprint arXiv:2010.15311* (2020).

[206] Yun-Ning Hung, Chih-Wei Wu, Iroro Orife, Aaron Hipple, William Wolcott, and Alexander Lerch. 2022. A large TV dataset for speech and music activity detection. *EURASIP Journal on Audio, Speech, and Music Processing* 2022, 1

(2022), 21.
[207] Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, and Yanzhang He. 2022. Large-scale asr domain adaptation using self-and semi-supervised learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6627–6631.
[208] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234* (2020).
[209] Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7904–7908.
[210] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy. 2020. Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss. *arXiv preprint arXiv:2008.04470* (2020).
[211] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.
[212] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409* (2021).
[213] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*. PMLR, 10120–10134.
[214] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. 2020. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991* (2020).
[215] Yunlong Jiao, Adam Gabryś, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa, and Viacheslav Klimkov. 2021. Universal neural vocoding with parallel wavenet. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6044–6048.
[216] Wen Jin, Xin Liu, Michael S Scordilis, and Lu Han. 2009. Speech enhancement using harmonic emphasis and adaptive comb filtering. *IEEE transactions on audio, speech, and language processing* 18, 2 (2009), 356–368.
[217] Yong Rae Jo, Young Ki Moon, Won Ik Cho, and Geun Sik Jo. 2021. Self-attentive vad: Context-aware detection of voice from noise. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6808–6812.
[218] Anubhav Johri, Ashish Tripathi, et al. 2019. Parkinson disease detection using deep neural networks. In *2019 twelfth international conference on contemporary computing (IC3)*. IEEE, 1–4.
[219] Yooncheol Ju, Ilhwan Kim, Hongsun Yang, Ji-Hoon Kim, Byeongyeol Kim, Soumi Maiti, and Shinji Watanabe. 2022. TriniTTS: Pitch-controllable End-to-end TTS without External Aligner. In *Proc. Interspeech*. 16–20.
[220] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu. 2019. Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. *arXiv preprint arXiv:1904.08104* (2019).
[221] Jee-weon Jung, Hee-Soo Heo, Ha-Jin Yu, and Joon Son Chung. 2021. Graph attention networks for speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6149–6153.
[222] Jee-weon Jung, Hee-Soo Heo, Ha-Jin Yu, and Joon Son Chung. 2021. Graph Attention Networks for Speaker Verification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6149–6153. https://doi.org/10.1109/ICASSP39728.2021.9414057
[223] Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7084–7088.
[224] Samuel Kakuba, Alwin Poulose, and Dong Seog Han. 2022. Deep Learning-Based Speech Emotion Recognition Using Multi-Level Fusion of Concurrent Features. *IEEE Access* 10 (2022), 125538–125551.
[225] Taku Kala and Takahiro Shinozaki. 2018. Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 5759–5763.
[226] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *International Conference on Machine Learning*. PMLR, 2410–2419.
[227] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099* (2016).
[228] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* (2014).

[229] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2019. ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 9 (2019), 1432–1443.

[230] Naoyuki Kanda, Jian Wu, Yu Wu, Xiong Xiao, Zhong Meng, Xiaofei Wang, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Takuya Yoshioka. 2022. Streaming Speaker-Attributed ASR with Token-Level Speaker Embeddings. *arXiv preprint arXiv:2203.16685* (2022).

[231] Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2022. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8082–8086.

[232] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6820–6824.

[233] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2020. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. *arXiv preprint arXiv:2010.11672* (2020).

[234] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2021. Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5919–5923.

[235] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. 2022. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6207–6211.

[236] Jiawen Kang, Ruiqi Liu, Lantian Li, Yunqi Cai, Dong Wang, and Thomas Fang Zheng. 2020. Domain-invariant speaker vector projection by model-agnostic meta-learning. *arXiv preprint arXiv:2005.11900* (2020).

[237] Ioannis Kansizoglou, Loukas Bampis, and Antonios Gasteratos. 2019. An active learning paradigm for online audio-visual emotion recognition. *IEEE Transactions on Affective Computing* 13, 2 (2019), 756–768.

[238] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 449–456.

[239] Shigeki Karita, Nelson Yalta, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *INTERSPEECH*.

[240] Kazuya Kawakami. 2008. *Supervised sequence labelling with recurrent neural networks*. Ph. D. Dissertation. Technical University of Munich.

[241] Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. *arXiv preprint arXiv:2001.11128* (2020).

[242] Tom Kenter, Vincent Wan, Chun-An Chan, Rob Clark, and Jakub Vit. 2019. CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. In *International Conference on Machine Learning*. PMLR, 3331–3340.

[243] Heeseung Kim, Sungwon Kim, and Sungroh Yoon. 2022. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*. PMLR, 11119–11133.

[244] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems* 33 (2020), 8067–8077.

[245] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[246] Juntae Kim and Jeehye Lee. 2021. Generalizing RNN-transducer to out-domain audio via sparse self-attention layers. *arXiv preprint arXiv:2108.10752* (2021).

[247] Jaechang Kim, Yunjoo Lee, Seunghoon Hong, and Jungseul Ok. 2022. Learning continuous representation of audio for arbitrary scale super resolution. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3703–3707.

[248] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee. 2021. Fre-gan: Adversarial frequency-consistent audio synthesis. *arXiv preprint arXiv:2106.02297* (2021).

[249] Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An efficient transformer for automatic speech recognition. *arXiv preprint arXiv:2206.00888* (2022).

[250] Seongbin Kim, Gyuwan Kim, Seongjin Shin, and Sangmin Lee. 2021. Two-Stage Textual Knowledge Distillation for End-to-End Spoken Language Understanding. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7463–7467. https://doi.org/10.1109/ICASSP39728.2021.9414619

[251] Sungwon Kim, Heeseung Kim, and Sungroh Yoon. 2022. Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data. *arXiv preprint arXiv:2205.15370* (2022).
[252] Sungwon Kim, Sang-gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon. 2018. FloWaveNet: A generative flow for raw audio. *arXiv preprint arXiv:1811.02155* (2018).
[253] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. 2020. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7009–7013.
[254] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. 2021. 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing* 151 (2021), 107398. https://doi.org/10.1016/j.ymssp.2020.107398
[255] Serkan Kiranyaz, Turker Ince, Ridha Hamila, and Moncef Gabbouj. 2015. Convolutional neural networks for patient-specific ECG classification. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2608–2611.
[256] Yuma Koizumi, Kohei Yatabe, Marc Delcroix, Yoshiki Masuyama, and Daiki Takeuchi. 2020. Speech enhancement using self-adaptation and multi-head self-attention. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 181–185.
[257] Yuma Koizumi, Heiga Zen, Kohei Yatabe, Nanxin Chen, and Michiel Bacchiani. 2022. SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping. *arXiv preprint arXiv:2203.16749* (2022).
[258] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 10 (2017), 1901–1913.
[259] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8102–8106.
[260] John Kominek and Alan W Black. 2004. The CMU Arctic speech databases. In *Fifth ISCA workshop on speech synthesis*.
[261] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 17022–17033.
[262] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).
[263] Sergey Koval and Sergey Krynov. 2020. Practice of usage of spectral analysis for forensic speaker identification. In *RLA2C 1998-Speaker Recognition and its Commercial and Forensic Applications*. 136–140.
[264] Yuichiro Koyama, Tyler Vuong, Stefan Uhlich, and Bhiksha Raj. 2020. Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks. *arXiv preprint arXiv:2005.11611* (2020).
[265] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre D'efossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. AudioGen: Textually Guided Audio Generation. *ArXiv* abs/2209.15352 (2022).
[266] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6124–6128.
[267] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet. 2020. Transfer learning of the expressivity using FLOW metric learning in multispeaker text-to-speech synthesis. In *INTERSPEECH 2020*.
[268] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019).
[269] Ohsung Kwon, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. 2019. An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis. *IEEE Signal Processing Letters* 26, 9 (2019), 1383–1387. https://doi.org/10.1109/LSP.2019.2931673
[270] Youngki Kwon, Hee-Soo Heo, Jee-weon Jung, You Jin Kim, Bong-Jin Lee, and Joon Son Chung. 2022. Multi-scale speaker embedding-based graph attention networks for speaker diarisation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8367–8371.
[271] Youngki Kwon, Jee-weon Jung, Hee-Soo Heo, You Jin Kim, Bong-Jin Lee, and Joon Son Chung. 2021. Adapting speaker embeddings for speaker diarisation. *arXiv preprint arXiv:2104.02879* (2021).
[272] Seong Min Kye, Youngmoon Jung, Hae Beom Lee, Sung Ju Hwang, and Hoirin Kim. 2020. Meta-learning for short utterance speaker recognition with imbalance length pairs. *arXiv preprint arXiv:2004.02863* (2020).
[273] Cheng-I Jeff Lai, Yang Zhang, Alexander H. Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David D. Cox, and James R. Glass. 2021. PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition. *CoRR* abs/2106.05933 (2021). arXiv:2106.05933 https://arxiv.org/abs/2106.05933

[274] Kushal Lakhotia et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics* 9 (2021).

[275] Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. Emorl: continuous acoustic emotion classification using deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4445–4450.

[276] Max WY Lam, Jun Wang, Dan Su, and Dong Yu. 2021. Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5759–5763.

[277] Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6588–6592.

[278] Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language* 71 (2022), 101254.

[279] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li. 2012. The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases. In *Annual Conference of the International Speech Communication Association (Interspeech)*.

[280] Anthony Larcher, Ambuj Mehrish, Marie Tahon, Sylvain Meignier, Jean Carrive, David Doukhan, Olivier Galibert, and Nicholas Evans. 2021. Speaker embeddings for diarization of broadcast data in the allies challenge. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5799–5803.

[281] Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2020. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering* 14 (2020), 342–356.

[282] Hung-yi Lee, Abdelrahman Mohamed, Shinji Watanabe, Tara Sainath, Karen Livescu, Shang-Wen Li, Shu-wen Yang, and Katrin Kirchhoff. 2022. Self-supervised Representation Learning for Speech Processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*. 8–13.

[283] Junhyeok Lee and Seungu Han. 2021. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321* (2021).

[284] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David Van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al. 2015. The RedDots data collection for speaker recognition. In *Interspeech 2015*.

[285] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka. 2019. The CORAL+ algorithm for unsupervised domain adaptation of PLDA. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5821–5825.

[286] Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. [n. d.]. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *International Conference on Learning Representations*.

[287] Sang-gil Lee, Sungwon Kim, and Sungroh Yoon. 2020. Nanoflow: Scalable normalizing flows with sublinear parameter complexity. *Advances in Neural Information Processing Systems* 33 (2020), 14058–14067.

[288] Sang-Hoon Lee, Seung-Bin Kim, Ji-Hyun Lee, Eunwoo Song, Min-Jae Hwang, and Seong-Whan Lee. 2022. HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis. *Advances in Neural Information Processing Systems* 35 (2022), 16624–16636.

[289] Yoonhyung Lee, Joongbo Shin, and Kyomin Jung. 2021. Bidirectional variational inference for non-autoregressive text-to-speech. In *International Conference on Learning Representations*.

[290] Quentin Lemaire and Andre Holzapfel. 2019. Temporal convolutional networks for speech and music detection in radio broadcast. In *20th International Society for Music Information Retrieval Conference, ISMIR 2019, 4-8 November 2019*. International Society for Music Information Retrieval.

[291] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. 2022. StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation. *arXiv preprint arXiv:2212.11851* (2022).

[292] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiang-Yang Li, Tao Qin, et al. 2022. BinauralGrad: A Two-Stage Conditional Diffusion Probabilistic Model for Binaural Audio Synthesis. *arXiv preprint arXiv:2205.14807* (2022).

[293] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6341–6345.

[294] Alon Levkovitch, Eliya Nachmani, and Lior Wolf. 2022. Zero-Shot Voice Conditioning for Denoising Diffusion TTS Models. *arXiv preprint arXiv:2206.02246* (2022).

[295] Bo Li, Shuo-yiin Chang, Tara N Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, and Yonghui Wu. 2020. Towards fast and accurate streaming end-to-end ASR. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6069–6073.

[296] Bo Li, Anmol Gulati, Jiahui Yu, Tara N Sainath, Chung-Cheng Chiu, Arun Narayanan, Shuo-Yiin Chang, Ruoming Pang, Yanzhang He, James Qin, et al. 2021. A better and faster end-to-end model for streaming asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5634–5638.

[297] Bo Li, Tara N Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean K Chin, et al. 2017. Acoustic Modeling for Google Home.. In *Interspeech*. 399–403.

[298] Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing* 11, 1 (2022).

[299] Jingyu Li, Wei Liu, and Tan Lee. 2022. EDITnet: A Lightweight Network for Unsupervised Domain Adaptation in Speaker Verification. *arXiv preprint arXiv:2206.07548* (2022).

[300] Jingdong Li, Hui Zhang, Xueliang Zhang, and Changliang Li. 2019. Single channel speech enhancement using temporal convolutional recurrent neural networks. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 896–900.

[301] Kai Li, Runxuan Yang, and Xiaolin Hu. 2022. An efficient encoder-decoder architecture with top-down attention for speech separation. *arXiv preprint arXiv:2209.15200* (2022).

[302] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6706–6713.

[303] Naihan Li, Yanqing Liu, Yu Wu, Shujie Liu, Sheng Zhao, and Ming Liu. 2020. Robutrans: A robust transformer-based text-to-speech model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8228–8235.

[304] Qinglong Li, Fei Gao, Haixin Guan, and Kaichi Ma. 2021. Real-time monaural speech enhancement with short-time discrete cosine transform. *arXiv preprint arXiv:2102.04629* (2021).

[305] Qiujia Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C Woodland, Liangliang Cao, and Trevor Strohman. 2021. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6388–6392.

[306] Rongjin Li, Weibin Zhang, and Dongpeng Chen. 2022. The coral++ algorithm for unsupervised domain adaptation of speaker recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7172–7176.

[307] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353

[308] Yingting Li, Ambuj Mehrish, Shuai Zhao, Rishabh Bhardwaj, Amir Zadeh, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2023. Evaluating Parameter-Efficient Transfer Learning Approaches on SURE Benchmark for Speech Understanding. *arXiv preprint arXiv:2303.03267* (2023).

[309] Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2022. StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis. *arXiv preprint arXiv:2205.15439* (2022).

[310] Dan Lim, Won Jang, Heayoung Park, Bongwan Kim, Jaesam Yoon, et al. 2020. Jdi-t: Jointly trained duration informed transformer for text-to-speech without explicit alignment. *arXiv preprint arXiv:2005.07799* (2020).

[311] Dan Lim, Sunghee Jung, and Eesung Kim. 2022. JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech. *arXiv preprint arXiv:2203.16852* (2022).

[312] Jae Lim and Alan Oppenheim. 1978. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, 3 (1978), 197–210.

[313] Teck Yian Lim, Raymond A. Yeh, Yijia Xu, Minh N. Do, and Mark Hasegawa-Johnson. 2018. Time-Frequency Networks for Audio Super-Resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 646–650. https://doi.org/10.1109/ICASSP.2018.8462049

[314] Ju Lin, Sufeng Niu, Zice Wei, Xiang Lan, Adriaan J Wijngaarden, Melissa C Smith, and Kuang-Ching Wang. 2019. Speech enhancement using forked generative adversarial networks with spectral subtraction. *Proceedings of Interspeech 2019* (2019).

[315] Ju Lin, Adriaan J. de Lind van Wijngaarden, Kuang-Ching Wang, and Melissa C. Smith. 2021. Speech Enhancement Using Multi-Stage Self-Attentive Temporal Convolutional Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3440–3450. https://doi.org/10.1109/TASLP.2021.3125143

[316] Jheng-hao Lin, Yist Y Lin, Chung-Ming Chien, and Hung-yi Lee. 2021. S2vc: a framework for any-to-any voice conversion with self-supervised pretrained representations. *arXiv preprint arXiv:2104.02901* (2021).

[317] Qingjian Lin, Yu Hou, and Ming Li. 2020. Self-Attentive Similarity Measurement Strategies in Speaker Diarization.. In *INTERSPEECH*. 284–288.

[318] Wei-Wei Lin and Man-Wai Mak. 2020. Wav2Spk: A Simple DNN Architecture for Learning Speaker Embeddings from Waveforms.. In *INTERSPEECH*. 3211–3215.
[319] Shaoshi Ling and Yuzong Liu. 2020. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659* (2020).
[320] Alexander H Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. 2023. Towards end-to-end unsupervised speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 221–228.
[321] Alexander H Liu, Cheng-I Jeff Lai, Wei-Ning Hsu, Michael Auli, Alexei Baevskiv, and James Glass. 2022. Simple and effective unsupervised speech synthesis. *arXiv preprint arXiv:2204.02524* (2022).
[322] Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2351–2366.
[323] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6419–6423.
[324] Chunxi Liu, Frank Zhang, Duc Le, Suyoun Kim, Yatharth Saraf, and Geoffrey Zweig. 2021. Improving RNN transducer based ASR with auxiliary tasks. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 172–179.
[325] Haohe Liu, Zehua Chen, Yiitan Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and MarkD . Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *ArXiv* abs/2301.12503 (2023).
[326] Haohe Liu, Woosung Choi, Xubo Liu, Qiuqiang Kong, Qiao Tian, and DeLiang Wang. 2022. Neural vocoder is all you need for speech super-resolution. *arXiv preprint arXiv:2203.14941* (2022).
[327] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11020–11028.
[328] Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8386–8390.
[329] Mengzhuo Liu and Yangjie Wei. 2022. An Improvement to Conformer-Based Model for High-Accuracy Speech Feature Extraction and Learning. *Entropy* 24, 7 (2022), 866.
[330] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. 2021. Expressive TTS Training With Frame and Style Reconstruction Loss. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1806–1818. https://doi.org/10.1109/TASLP.2021.3076369
[331] Rui Liu, Berrak Sisman, and Haizhou Li. 2021. Graphspeech: Syntax-aware graph attention network for neural speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6059–6063.
[332] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. 2021. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1717–1728.
[333] Songxiang Liu, Dan Su, and Dong Yu. 2022. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972* (2022).
[334] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2019. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2019), 1862–1878.
[335] Yanqing Liu, Zhihang Xu, Gang Wang, Kuan Chen, Bohan Li, Xu Tan, Jinzhu Li, Lei He, and Sheng Zhao. 2021. Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021. *arXiv preprint arXiv:2110.12612* (2021).
[336] Zhengxi Liu, Qiao Tian, Chenxu Hu, Xudong Liu, Menglin Wu, Yuping Wang, Hang Zhao, and Yuxuan Wang. 2022. Controllable and Lossless Non-Autoregressive End-to-End Text-to-Speech. *arXiv preprint arXiv:2207.06088* (2022).
[337] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal. 2018. Towards achieving robust universal neural vocoding. *arXiv preprint arXiv:1811.06292* (2018).
[338] Xugang Lu, Sheng Li, and Masakiyo Fujimoto. 2020. Automatic speech recognition. *Speech-to-Speech Translation* (2020), 21–38.
[339] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2013. Speech enhancement based on deep denoising autoencoder.. In *Interspeech*, Vol. 2013. 436–440.
[340] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe. 2021. A Study on Speech Enhancement Based on Diffusion Probabilistic Model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 659–666.
[341] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. 2022. Conditional Diffusion Probabilistic Model for Speech Enhancement. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7402–7406. https://doi.org/10.1109/ICASSP43922.2022.9746901

[342] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. 2022. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7402–7406.

[343] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670* (2019).

[344] Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 46–50.

[345] Yi Luo and Nima Mesgarani. 2018. Real-time single-channel dereverberation and separation with time-domain audio separation network.. In *Interspeech*. 342–346.

[346] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27, 8 (2019), 1256–1266.

[347] Manh Luong and Viet Anh Tran. 2021. FlowVocoder: A small Footprint Neural Vocoder based Normalizing flow for Speech Synthesis. *arXiv preprint arXiv:2109.13675* (2021).

[348] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[349] Shahar Lutati, Eliya Nachmani, and Lior Wolf. 2022. Sepit approaching a single channel speech separation bound. *arXiv preprint arXiv:2205.11801* (2022).

[350] Shahar Lutati, Eliya Nachmani, and Lior Wolf. 2023. Separate And Diffuse: Using a Pretrained Diffusion Model for Improving Source Separation. *arXiv preprint arXiv:2301.10752* (2023).

[351] Florian Lux and Ngoc Thang Vu. 2022. Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features. *arXiv preprint arXiv:2203.03191* (2022).

[352] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2021. Lira: Learning visual speech representations from audio through self-supervision. *arXiv preprint arXiv:2106.09171* (2021).

[353] Shuang Ma, Daniel Mcduff, and Yale Song. 2019. Neural TTS stylization with adversarial and collaborative games. In *International Conference on Learning Representations*.

[354] Duncan Macho, Laurent Mauuary, Bernhard Noé, Yan Ming Cheng, Doug Ealey, Denis Jouvet, Holly Kelleher, David Pearce, and Fabien Saadoun. 2002. Evaluation of a noise-robust DSR front-end on Aurora databases. In *Seventh International Conference on Spoken Language Processing*.

[355] Gallil Maimon and Yossi Adi. 2022. Speaking Style Conversion With Discrete Self-Supervised Units. *arXiv preprint arXiv:2212.09730* (2022).

[356] Soumi Maiti and Michael I Mandel. 2020. Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 206–210.

[357] Somshubra Majumdar, Shantanu Acharya, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Damage Control During Domain Adaptation for Transducer Based Automatic Speech Recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 130–135.

[358] Somshubra Majumdar, Jagadeesh Balam, Oleksii Hrinchuk, Vitaly Lavrukhin, Vahid Noroozi, and Boris Ginsburg. 2021. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv preprint arXiv:2104.01721* (2021).

[359] Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. 2022. A Kernel-Based View of Language Model Fine-Tuning. *ArXiv* abs/2210.05643 (2022).

[360] Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6344–6348.

[361] Pranay Manocha and Anurag Kumar. 2022. Speech quality assessment through MOS using non-matching references. *arXiv preprint arXiv:2206.12285* (2022).

[362] Pranay Manocha, Buye Xu, and Anurag Kumar. 2021. NORESQA: A framework for speech quality assessment using non-matching references. *Advances in Neural Information Processing Systems* 34 (2021), 22363–22378.

[363] Narla John Metilda Sagaya Mary, Srinivasan Umesh, and Sandesh Varadaraju Katta. 2021. S-vectors and TESA: Speaker embeddings and a speaker authenticator based on transformer encoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 404–413.

[364] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. 2016. The speakers in the wild (SITW) speaker recognition database.. In *Interspeech*. 818–822.

[365] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, et al. 2020. Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario. *arXiv preprint arXiv:2005.07272*

(2020).
[366] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. SampleRNN: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837* (2016).
[367] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G Elmore, and Linda Shapiro. 2018. Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 893–901.
[368] Shivam Mehta, Éva Székely, Jonas Beskow, and Gustav Eje Henter. 2022. Neural HMMS Are All You Need (For High-Quality Attention-Free TTS). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7457–7461. https://doi.org/10.1109/ICASSP43922.2022.9746686
[369] Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Flow-tts: A non-autoregressive network for text to speech based on flow. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7209–7213.
[370] Chenfeng Miao, Liang Shuang, Zhengchen Liu, Chen Minchuan, Jun Ma, Shaojun Wang, and Jing Xiao. 2021. Efficienttts: An efficient and high-quality text-to-speech architecture. In *International Conference on Machine Learning*. PMLR, 7700–7709.
[371] Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. 2020. Transformer-Based Online CTC/Attention End-To-End Speech Recognition Architecture. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6084–6088. https://doi.org/10.1109/ICASSP40776.2020.9053165
[372] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1368–1396.
[373] Serban Mihalache and Dragos Burileanu. 2022. Using Voice Activity Detection and Deep Neural Networks with Hybrid Speech Feature Extraction for Deceptive Speech Detection. *Sensors* 22, 3 (2022), 1228.
[374] Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems* 35 (2022), 33428–33443.
[375] Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. 2019. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660* (2019).
[376] Joao Monteiro, Md Jahangir Alam, and Tiago H Falk. 2019. Combining Speaker Recognition and Metric Learning for Speaker-Dependent Representation Learning.. In *INTERSPEECH*. 4015–4019.
[377] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. 2021. A cappella: Audio-visual singing voice separation. *arXiv preprint arXiv:2104.09946* (2021).
[378] Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs. 2021. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6034–6038.
[379] Eliya Nachmani, Yossi Adi, and Lior Wolf. 2020. Voice separation with an unknown number of multiple speakers. In *International Conference on Machine Learning*. PMLR, 7164–7175.
[380] Tomohiro Nakatani. 2019. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proc. Interspeech*, Vol. 2019.
[381] Yoshihiko Nankaku, Kenta Sumiya, Takenori Yoshimura, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda. 2021. Neural sequence-to-sequence speech synthesis using a hidden semi-Markov model based structured attention mechanism. *arXiv preprint arXiv:2108.13985* (2021).
[382] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access* 7 (2019), 19143–19165.
[383] Huu Binh Nguyen, Duong Van Hai, Tien Dat Bui, Hoang Ngoc Chau, and Quoc Cuong Nguyen. 2022. Multi-Channel Speech Enhancement using a Minimum Variance Distortionless Response Beamformer based on Graph Convolutional Network. *International Journal of Advanced Computer Science and Applications* 13, 10 (2022).
[384] Viet-Anh Nguyen, Anh HT Nguyen, and Andy WH Khong. 2022. Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 161–165.
[385] Viet-Nhat Nguyen, Mostafa Sadeghi, Elisa Ricci, and Xavier Alameda-Pineda. 2021. Deep variational generative models for audio-visual speech separation. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.
[386] Xuan-Phi Nguyen, Sravya Popuri, Changhan Wang, Yun Tang, Ilia Kulikov, and Hongyu Gong. 2022. Improving Speech-to-Speech Translation Through Unlabeled Text. *arXiv preprint arXiv:2210.14514* (2022).

[387] Phani Sankar Nidadavolu, Jesús Villalba, and Najim Dehak. 2019. Cycle-gans for domain adaptation of acoustic features for speaker recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6206–6210.

[388] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences* 9, 19 (2019), 4050.

[389] Peiqing Niu, Zhongfu Chen, Meina Song, et al. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. *arXiv preprint arXiv:1907.00390* (2019).

[390] Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai. 2019. Real-Time Neural Text-to-Speech with Sequence-to-Sequence Acoustic Model and WaveGlow or Single Gaussian WaveRNN Vocoders.. In *INTERSPEECH*. 1308–1312.

[391] Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai. 2019. Tacotron-Based Acoustic Model Using Phoneme Alignment for Practical Neural Text-to-Speech Systems. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 214–221. https://doi.org/10.1109/ASRU46091.2019.9003956

[392] Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai. 2020. Transformer-Based Text-to-Speech with Weighted Forced Attention. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6729–6733. https://doi.org/10.1109/ICASSP40776.2020.9053915

[393] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*. PMLR, 3918–3926.

[394] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[395] Jasper Ooster and Bernd T Meyer. 2019. Improving deep models of speech quality prediction through voice activity detection and entropy-based measures. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 636–640.

[396] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).

[397] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv* abs/2203.02155 (2022).

[398] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. 2011. The importance of phase in speech enhancement. *speech communication* 53, 4 (2011), 465–494.

[399] Giridhar Pamisetty and K Sri Rama Murty. 2023. Prosody-TTS: An end-to-end speech synthesis system with prosody control. *Circuits, Systems, and Signal Processing* 42, 1 (2023), 361–384.

[400] Jing Pan, Tao Lei, Kwangyoun Kim, Kyu J. Han, and Shinji Watanabe. 2022. SRU++: Pioneering Fast Recurrence with Attention for Speech Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7872–7876. https://doi.org/10.1109/ICASSP43922.2022.9746187

[401] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[402] Ashutosh Pandey and DeLiang Wang. 2019. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6875–6879.

[403] Ilias Papastratis. 2021. Speech Recognition: a review of the different deep learning approaches. *Accessed on* 2 (2021), 2021.

[404] Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le. 2020. Improved noisy student training for automatic speech recognition. *arXiv preprint arXiv:2005.09629* (2020).

[405] Sangjun Park, Kihyun Choo, Joohyung Lee, Anton V Porov, Konstantin Osipov, and June Sig Sung. 2022. Bunched LPCNet2: Efficient Neural Vocoders Covering Devices from Cloud to Edge. *arXiv preprint arXiv:2203.14416* (2022).

[406] Tae Jin Park, Kyu J Han, Manoj Kumar, and Shrikanth Narayanan. 2019. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters* 27 (2019), 381–385.

[407] Vishal Passricha and Rajesh Kumar Aggarwal. 2019. A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. *Journal of Intelligent Systems* 29, 1 (2019), 1261–1274.

[408] Dipjyoti Paul, Sankar Mukherjee, Yannis Pantazis, and Yannis Stylianou. 2021. A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization.. In *Interspeech*. 3625–3629.

[409] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. 2020. Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions. *arXiv preprint arXiv:2008.05289* (2020).
[410] Blanca Pena and Luofeng Huang. 2021. Wave-GAN: a deep learning approach for the prediction of nonlinear regular wave loads and run-up on a fixed cylinder. *Coastal Engineering* 167 (2021), 103902.
[411] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. 2020. Non-autoregressive neural text-to-speech. In *International conference on machine learning*. PMLR, 7586–7598.
[412] Zilun Peng, Akshay Budhkar, Ilana Tuil, Jason Levy, Parinaz Sobhani, Raphael Cohen, and Jumana Nassour. 2021. Shrinking Bigfoot: Reducing wav2vec 2.0 footprint. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Virtual, 134–141. https://doi.org/10.18653/v1/2021.sustainlp-1.14
[413] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterFusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247* (2020).
[414] Minh Pham, Zeqian Li, and Jacob Whitehill. 2020. Toward better speaker embeddings: Automated collection of speech samples from unknown distinct speakers. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7089–7093.
[415] Wei Ping, Kainan Peng, and Jitong Chen. 2018. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281* (2018).
[416] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654* (2017).
[417] Wei Ping, Kainan Peng, Kexin Zhao, and Zhao Song. 2020. Waveflow: A compact flow-based model for raw audio. In *International Conference on Machine Learning*. PMLR, 7706–7716.
[418] Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-Training for End-to-End Speech Translation. *Proc. Interspeech 2020* (2020), 1476–1480.
[419] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355* (2021).
[420] Adam Polyak, Lior Wolf, and Yaniv Taigman. 2019. TTS skins: Speaker conversion via ASR. *arXiv preprint arXiv:1904.08983* (2019).
[421] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*. PMLR, 8599–8608.
[422] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. 2021. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821* (2021).
[423] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI.. In *Interspeech*. 2751–2755.
[424] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A Comparison of sequence-to-sequence models for speech recognition.. In *Interspeech*. 939–943.
[425] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3617–3621.
[426] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. 2020. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*. PMLR, 7836–7846.
[427] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. 2022. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International Conference on Machine Learning*. PMLR, 18003–18017.
[428] Xiaoyi Qin, Hui Bu, and Ming Li. 2020. Hi-mia: A far-field text-dependent speaker verification database and the baselines. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7609–7613.
[429] Xiaoyi Qin, Ming Li, Hui Bu, Rohan Kumar Das, Wei Rao, Shrikanth Narayanan, and Haizhou Li. 2020. The ffsvc 2020 evaluation plan. *arXiv preprint arXiv:2002.00387* (2020).
[430] Zhibin Qiu, Mengfan Fu, Yinfeng Yu, LiLi Yin, Fuchun Sun, and Hao Huang. 2022. SRTNet: Time Domain Speech Enhancement Via Stochastic Refinement. *arXiv preprint arXiv:2210.16805* (2022).
[431] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.
[432] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[433] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
[434] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[435] Kacper Radzikowski, Robert Nowak, Le Wang, and Osamu Yoshie. 2019. Dual supervised learning for non-native speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing* 2019 (2019), 1–10.
[436] Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *International conference on machine learning*. PMLR, 2837–2846.
[437] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[438] Mehrdad Rafiepour and Javad Salimi Sartakhti. 2023. CTRAN: CNN-Transformer-based Network for Natural Language Understanding. *arXiv preprint arXiv:2303.10606* (2023).
[439] Tuomo Raitio, Ramya Rasipuram, and Dan Castellani. 2020. Controllable neural text-to-speech synthesis using intuitive prosodic features. *arXiv preprint arXiv:2009.06775* (2020).
[440] Thejan Rajapakshe, Siddique Latif, Rajib Rana, Sara Khalifa, and Björn W Schuller. 2020. Deep reinforcement learning with pre-training for time-efficient training of automatic speech recognition. *arXiv preprint arXiv:2005.11172* (2020).
[441] Thejan Rajapakshe, Rajib Rana, Sara Khalifa, Jiajun Liu, and Bjorn Schuller. 2022. A novel policy for pre-trained deep reinforcement learning for speech emotion recognition. In *Australasian Computer Science Week 2022*. 96–105.
[442] Nathanaël Carraz Rakotonirina. 2021. Self-attention for audio super-resolution. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.
[443] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 1021–1028.
[444] Mirco Ravanelli, Titouan Parcollet, and Yoshua Bengio. 2019. The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6465–6469.
[445] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6989–6993.
[446] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. 2020. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981* (2020).
[447] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (2020).
[448] Yi Ren, Jinglin Liu, and Zhou Zhao. 2021. Portaspeech: Portable and high-quality generative text-to-speech. *Advances in Neural Information Processing Systems* 34 (2021), 13963–13974.
[449] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems* 32 (2019).
[450] Daniel Rho, Jinhyeok Park, and Jong Hwan Ko. 2022. Nas-vad: Neural architecture search for voice activity detection. *arXiv preprint arXiv:2201.09032* (2022).
[451] Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al. 2018. Voices obscured in complex environmental settings (voices) corpus. *arXiv preprint arXiv:1804.05053* (2018).
[452] Julius Richter, Guillaume Carbajal, and Timo Gerkmann. 2020. Speech Enhancement with Stochastic Temporal Convolutional Networks.. In *Interspeech*. 4516–4520.
[453] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, 749–752.
[454] Amir Mohammad Rostami, Ali Karimi, and Mohammad Ali Akhaee. 2022. Keyword spotting in continuous speech using convolutional neural network. *Speech Communication* 142 (2022), 15–21.
[455] Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus.. In *LREC*. 125–129.
[456] Sidheswar Routray and Qirong Mao. 2022. Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network. *Computer Speech & Language* 71 (2022), 101270.
[457] Mickael Rouvier, Richard Dufour, and Pierre-Michel Bousquet. 2021. Review of different robust x-vector extractors for speaker verification. In *2020 28th European Signal Processing Conference (EUSIPCO)*. 1–5. https://doi.org/10.23919/

Eusipco47968.2020.9287426
[458] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2018. First DIHARD challenge evaluation plan. *2018, tech. Rep.* (2018).
[459] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2019. The second dihard diarization challenge: Dataset, task, and baselines. *arXiv preprint arXiv:1906.07839* (2019).
[460] Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirkó Visontai, and Stella Laurenzo. 2020. Streaming keyword spotting on mobile devices. *arXiv preprint arXiv:2005.06720* (2020).
[461] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2021. Perceptual-similarity-aware deep speaker representation learning for multi-speaker generative modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1033–1048.
[462] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. 2017. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078* (2017).
[463] Elizabeth Salesky, Matthias Sperber, and Alan W Black. 2019. Exploring phoneme-level speech representations for end-to-end speech translation. *arXiv preprint arXiv:1906.01199* (2019).
[464] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8537–8541.
[465] Pascal Scalart et al. 1996. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 2. IEEE, 629–632.
[466] Carolina Scarton, Mikel L Forcada, Miquel Espla-Gomis, and Lucia Specia. 2019. Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality. *arXiv preprint arXiv:1910.06204* (2019).
[467] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi. 2022. Diffusion-based Generative Speech Source Separation. *arXiv preprint arXiv:2210.17327* (2022).
[468] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).
[469] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
[470] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.
[471] Deokjin Seo, Heung-Seon Oh, and Yuchul Jung. 2021. Wav2kws: Transfer learning from speech representations for keyword spotting. *IEEE Access* 9 (2021), 80682–80691.
[472] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. 2022. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065* (2022).
[473] Joan Serrà, Jordi Pons, and Santiago Pascual. 2021. SESQA: semi-supervised learning for speech quality assessment. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 381–385.
[474] Benjamin Sertolli, Zhao Ren, Björn W Schuller, and Nicholas Cummins. 2021. Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. *Computer Speech & Language* 68 (2021), 101204.
[475] Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. 2020. Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling. *arXiv preprint arXiv:2010.04301* (2020).
[476] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.
[477] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Christian Fuegen, Frank Zhang, Duc Le, Ching-Feng Yeh, and Michael L Seltzer. 2020. Weak-attention suppression for transformer based speech recognition. *arXiv preprint arXiv:2005.09137* (2020).
[478] Kevin J Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro. 2021. RAD-TTS: Parallel flow-based TTS with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.
[479] Hye-Jin Shim, Jungwoo Heo, Jae-Han Park, Ga-Hui Lee, and Ha-Jin Yu. 2022. Graph Attentive Feature Aggregation for Text-Independent Speaker Verification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7972–7976. https://doi.org/10.1109/ICASSP43922.2022.9746257
[480] Amitay Sicherman and Yossi Adi. 2023. Analysing Discrete Self Supervised Speech Representation for Spoken Language Modeling. *arXiv preprint arXiv:2301.00591* (2023).

[481] Nikola Simić, Siniša Suzić, Tijana Nosek, Mia Vujović, Zoran Perić, Milan Savić, and Vlado Delić. 2022. Speaker recognition using constrained convolutional neural networks in emotional speech. *Entropy* 24, 3 (2022), 414.
[482] Ruby Melody Simply, Eliran Dafna, and Yaniv Zigel. 2019. Diagnosis of Obstructive Sleep Apnea using Speech Signals from Awake Subjects. *IEEE Journal of Selected Topics in Signal Processing* 14, 2 (2019), 251–260.
[483] Gundeep Singh, Sahil Sharma, Vijay Kumar, Manjit Kaur, Mohammed Baz, and Mehedi Masud. 2021. Spoken language identification using deep learning. *Computational Intelligence and Neuroscience* 2021 (2021).
[484] Prachi Singh and Sriram Ganapathy. 2021. Self-Supervised Metric Learning With Graph Clustering For Speaker Diarization. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 90–97. https://doi.org/10.1109/ASRU51503.2021.9688271
[485] Prachi Singh, Amrit Kaul, and Sriram Ganapathy. 2023. Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization. *arXiv preprint arXiv:2302.12716* (2023).
[486] Satwinder Singh, Ruili Wang, and Feng Hou. 2022. Improved Meta Learning for Low Resource Speech Recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4798–4802.
[487] Hubert Siuzdak, Piotr Dura, Pol van Rijn, and Nori Jacoby. 2022. WavThruVec: Latent speech representation as intermediate features for neural speech synthesis. *arXiv preprint arXiv:2203.16930* (2022).
[488] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*. PMLR, 4693–4702.
[489] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).
[490] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep neural network embeddings for text-independent speaker verification.. In *Interspeech*, Vol. 2017. 999–1003.
[491] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5329–5333.
[492] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.
[493] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6447–6456.
[494] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786* (2018).
[495] Meet H Soni and Hemant A Patil. 2016. Novel deep autoencoder features for non-intrusive speech quality assessment. In *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2315–2319.
[496] Alexander Sorin, Slava Shechtman, and Ron Hoory. 2020. Principal Style Components: Expressive Style Control and Cross-Speaker Transfer in Neural TTS.. In *INTERSPEECH*. 3411–3415.
[497] Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. *arXiv preprint arXiv:2004.06358* (2020).
[498] Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185* (2018).
[499] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 21–25.
[500] Vrunda N Sukhadia and S Umesh. 2023. Domain Adaptation of low-resource Target-Domain models using well-trained ASR Conformer Models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 295–301.
[501] Aolan Sun, Jianzong Wang, Ning Cheng, Huayi Peng, Zhen Zeng, Lingwei Kong, and Jing Xiao. 2021. Graphpb: Graphical representations of prosody boundary in speech synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 438–445.
[502] Aolan Sun, Jianzong Wang, Ning Cheng, Huayi Peng, Zhen Zeng, and Jing Xiao. 2020. GraphTTS: Graph-to-Sequence Modelling in Neural Text-to-Speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6719–6723. https://doi.org/10.1109/ICASSP40776.2020.9053355
[503] Tzu-Wei Sung, Jun-You Liu, Hung-yi Lee, and Lin-shan Lee. 2019. Towards end-to-end speech-to-text translation with two-pass decoding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7175–7179.
[504] Suno-AI. 2023. *Bark*. https://github.com/suno-ai/bark
[505] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. [n. d.]. End-to-End ASR: from Supervised to Semi-Supervised Learning with Modern Architectures. In *ICML 2020 Workshop on Self-supervision in Audio and Speech*.

[506] Jaesung Tae, Hyeongju Kim, and Taesu Kim. 2021. EdiTTS: Score-based Editing for Controllable Text-to-Speech. *CoRR* abs/2110.02584 (2021). arXiv:2110.02584 https://arxiv.org/abs/2110.02584

[507] Ke Tan and DeLiang Wang. 2019. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), 380–390.

[508] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo. 2019. ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6805–6809. https://doi.org/10.1109/ICASSP.2019.8683282

[509] Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Deep speaker embedding learning with multi-level pooling for text-independent speaker verification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6116–6120.

[510] Fei Tao and Carlos Busso. 2019. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Speech Communication* 113 (2019), 25–35.

[511] Xiaohai Tian, Eng Siong Chng, and Haizhou Li. 2019. A vocoder-free WaveNet voice conversion with non-parallel data. *arXiv preprint arXiv:1902.03705* (2019).

[512] Noé Tits, Fengna Wang, Kevin El Haddad, Vincent Pagel, and Thierry Dutoit. 2019. Visualization and Interpretation of Latent Spaces for Controlling Expressive Speech Synthesis Through Audio Analysis. *Proc. Interspeech 2019* (2019), 4475–4479.

[513] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2018. Sequence-to-sequence ASR optimization via reinforcement learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5829–5833.

[514] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv* abs/2302.13971 (2023).

[515] Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing* 14, 5 (2006), 1557–1565.

[516] Emiru Tsunoo, Yosuke Kashiwagi, Toshiyuki Kumakura, and Shinji Watanabe. 2019. Transformer ASR with contextual block processing. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 427–433.

[517] Tao Tu, Yuan-Jui Chen, Cheng-chieh Yeh, and Hung-Yi Lee. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508* (2019).

[518] Zoltán Tüske, Kartik Audhkhasi, and George Saon. 2019. Advancing Sequence-to-Sequence Based Speech Recognition.. In *Interspeech*. 3780–3784.

[519] Efthymios Tzinis, Yossi Adi, Vamsi K Ithapu, Buye Xu, and Anurag Kumar. 2022. Continual self-training with bootstrapped remixing for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6947–6951.

[520] Efthymios Tzinis, Yossi Adi, Vamsi K Ithapu, Buye Xu, Paris Smaragdis, and Anurag Kumar. 2022. RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1329–1341.

[521] Panagiotis Tzirakis, Anurag Kumar, and Jacob Donley. 2021. Multi-channel speech enhancement using graph neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3415–3419.

[522] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. 2020. Emotional Speech Synthesis with Rich and Granularized Control. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7254–7258. https://doi.org/10.1109/ICASSP40776.2020.9053732

[523] Jan Vainer and Ondřej Dušek. 2020. Speedyspeech: Efficient neural speech synthesis. *arXiv preprint arXiv:2008.03802* (2020).

[524] Jean-Marc Valin, Umut Isik, Paris Smaragdis, and Arvindh Krishnaswamy. 2022. Neural speech synthesis on a shoestring: Improving the efficiency of lpcnet. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8437–8441.

[525] Jean-Marc Valin and Jan Skoglund. 2019. LPCNet: Improving neural speech synthesis through linear prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5891–5895.

[526] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6189–6193. https://doi.org/10.1109/ICASSP40776.2020.9054556

[527] Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957* (2020).

[528] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems* 29 (2016).
[529] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).
[530] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4052–4056.
[531] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4052–4056. https://doi.org/10.1109/ICASSP.2014.6854363
[532] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[533] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat* 1050, 20 (2017), 10–48550.
[534] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341* (2018).
[535] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. 2018. *Audio source separation and speech enhancement*. John Wiley & Sons.
[536] Thilo von Neumann, Keisuke Kinoshita, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2021. Graph-PIT: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers. *arXiv preprint arXiv:2107.14446* (2021).
[537] Tyler Vuong, Yangyang Xia, and Richard M Stern. 2021. A modulation-domain loss for neural-network-based real-time speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6643–6647.
[538] Roman Vygon and Nikolay Mikhaylovskiy. 2021. Learning efficient representations for keyword spotting with triplet loss. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*. Springer, 773–785.
[539] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4879–4883.
[540] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111* (2023).
[541] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171* (2020).
[542] Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310* (2020).
[543] Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*. PMLR, 10937–10947.
[544] Feng Wang and David MJ Tax. 2016. Survey on the attention based RNN model and its applications in computer vision. *arXiv preprint arXiv:1601.06823* (2016).
[545] Gary Wang. 2019. Deep text-to-speech system with seq2seq model. *arXiv preprint arXiv:1903.07398* (2019).
[546] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5265–5274.
[547] Jingsong Wang, Yuxuan He, Chunyu Zhao, Qijie Shao, Wei-Wei Tu, Tom Ko, Hung-yi Lee, and Lei Xie. 2021. Auto-KWS 2021 Challenge: Task, datasets, and baselines. *arXiv preprint arXiv:2104.00513* (2021).
[548] Jixuan Wang, Xiong Xiao, Jian Wu, Ranjani Ramamurthy, Frank Rudzicz, and Michael Brudno. 2020. Speaker Diarization with Session-Level Speaker Embedding Refinement Using Graph Neural Networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7109–7113. https://doi.org/10.1109/ICASSP40776.2020.9054176
[549] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker diarization with LSTM. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5239–5243.
[550] Qing Wang, Pengcheng Guo, Sining Sun, Lei Xie, and John HL Hansen. 2019. Adversarial Regularization for End-to-End Robust Speaker Verification.. In *Interspeech*. 4010–4014.
[551] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li. 2018. Unsupervised Domain Adaptation via Domain Adversarial Training for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP). 4889–4893. https://doi.org/10.1109/ICASSP.2018.8461423

[552] Tianzi Wang, Jiajun Deng, Mengzhe Geng, Zi Ye, Shoukang Hu, Yi Wang, Mingyu Cui, Zengrui Jin, Xunying Liu, and Helen Meng. 2022. Conformer Based Elderly Speech Recognition System for Alzheimer's Disease Detection. *arXiv preprint arXiv:2206.13232* (2022).

[553] Tingting Wang, Zexu Pan, Meng Ge, Zhen Yang, and Haizhou Li. 2023. Time-Domain Speech Separation Networks With Graph Encoding Auxiliary. *IEEE Signal Processing Letters* 30 (2023), 110–114.

[554] Weiqing Wang, Qingjian Lin, Danwei Cai, and Ming Li. 2022. Similarity measurement of segment-level speaker embeddings in speaker diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2645–2658.

[555] Xueyi Wang, Lantian Li, and Dong Wang. 2019. VAE-based domain adaptation for speaker verification. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 535–539.

[556] Xi Wang, Huaiping Ming, Lei He, and Frank K Soong. 2020. s-transformer: Segment-transformer for robust neural speech synthesis. *arXiv preprint arXiv:2011.08480* (2020).

[557] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. 2023. AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models. arXiv:2304.00830 [cs.SD]

[558] Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235* (2018).

[559] Yongqiang Wang, Yangyang Shi, Frank Zhang, Chunyang Wu, Julian Chan, Ching-Feng Yeh, and Alex Xiao. 2021. Transformer in Action: A Comparative Study of Transformer-Based Acoustic Models for Large Scale Speech Recognition Applications. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6778–6782. https://doi.org/10.1109/ICASSP39728.2021.9414087

[560] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[561] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*. PMLR, 5180–5189.

[562] Yi Wang, Shiqi Zhang, and Joohyung Lee. 2019. Bridging commonsense reasoning and probabilistic planning via a probabilistic action language. *Theory and Practice of Logic Programming* 19, 5-6 (2019), 1090–1106.

[563] Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2019. Structured Pruning of Large Language Models. *CoRR* abs/1910.04732 (2019). arXiv:1910.04732 http://arxiv.org/abs/1910.04732

[564] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey. 2018. Alternative objective functions for deep clustering. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 686–690.

[565] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. 2020. Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. *IEEE/ACM transactions on audio, speech, and language processing* 28 (2020), 1778–1787.

[566] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

[567] Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P Kingma. 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5679–5683.

[568] Chao Weng, Jia Cui, Guangsen Wang, Jun Wang, Chengzhu Yu, Dan Su, and Dong Yu. 2018. Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition.. In *Interspeech*. 761–765.

[569] Nils L Westhausen and Bernd T Meyer. 2020. Dual-signal transformation lstm network for real-time noise suppression. *arXiv preprint arXiv:2005.07551* (2020).

[570] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. 2020. Lightweight and Efficient End-To-End Speech Recognition Using Low-Rank Transformer. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6144–6148. https://doi.org/10.1109/ICASSP40776.2020.9053878

[571] Da-Yi Wu and Hung-yi Lee. 2020. One-shot voice conversion by vector quantization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7734–7738.

[572] Da-Yi Wu and Hung-yi Lee. 2020. One-Shot Voice Conversion by Vector Quantization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7734–7738. https://doi.org/10.1109/ICASSP40776.2020.9053854

[573] Jianfeng Wu, Yongzhu Hua, Shengying Yang, Hongshuai Qin, and Huibin Qin. 2019. Speech enhancement using generative adversarial network by distilling knowledge from statistical method. *Applied Sciences* 9, 16 (2019), 3396.

[574] Shoule Wu and Ziqiang Shi. 2021. ItoTTS and ItoWave: Linear Stochastic Differential Equation Is All You Need For Audio Generation. *arXiv preprint arXiv:2105.07583* (2021).
[575] Xianchao Wu. 2022. Deep Sparse Conformer for Speech Recognition. *arXiv preprint arXiv:2209.00260* (2022).
[576] Yihan Wu, Xu Tan, Bohan Li, Lei He, Sheng Zhao, Ruihua Song, Tao Qin, and Tie-Yan Liu. 2022. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. *arXiv preprint arXiv:2204.00436* (2022).
[577] Wei Xia, Jing Huang, and John HL Hansen. 2019. Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5816–5820.
[578] Xiong Xiao, Naoyuki Kanda, Zhuo Chen, Tianyan Zhou, Takuya Yoshioka, Sanyuan Chen, Yong Zhao, Gang Liu, Yu Wu, Jian Wu, et al. 2021. Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5824–5828.
[579] Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari. 2020. Cross-Lingual Text-To-Speech Synthesis via Domain Adaptation and Perceptual Similarity Regression in Speaker Space.. In *Interspeech*. 2947–2951.
[580] Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. *arXiv preprint arXiv:2105.05752* (2021).
[581] Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2802–2812.
[582] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3030–3034.
[583] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2014. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 1 (2014), 7–19.
[584] Jinlong Xue, Yayue Deng, Yichen Han, Ya Li, Jianqing Sun, and Jiaen Liang. 2022. ECAPA-TDNN for Multi-speaker Text-to-speech Synthesis. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 230–234.
[585] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6199–6203.
[586] Yuzi Yan, Xu Tan, Bohan Li, Tao Qin, Sheng Zhao, Yuan Shen, and Tie-Yan Liu. 2021. Adaspeech 2: Adaptive text to speech with untranscribed data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6613–6617.
[587] Dongchao Yang, Songxiang Liu, Jianwei Yu, Helin Wang, Chao Weng, and Yuexian Zou. 2022. NoreSpeech: Knowledge Distillation based Conditional Diffusion Model for Noise-robust Expressive TTS. *arXiv preprint arXiv:2211.02448* (2022).
[588] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983* (2022).
[589] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2021. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 492–498.
[590] Gene-Ping Yang, Chao-I Tuan, Hung-Yi Lee, and Lin-shan Lee. 2019. Improved speech separation with time-and-frequency cross-domain joint embedding and clustering. *arXiv preprint arXiv:1904.07845* (2019).
[591] Jinhyeok Yang, Junmo Lee, Youngik Kim, Hoonyoung Cho, and Injung Kim. 2020. VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network. *arXiv preprint arXiv:2007.15256* (2020).
[592] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing Universal PERformance Benchmark. *CoRR* abs/2105.01051 (2021). arXiv:2105.01051 https://arxiv.org/abs/2105.01051
[593] Shiqing Yang and Min Liu. 2022. Data augmentation for speaker verification. In *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*. 1247–1251.
[594] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. 2019. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*. 1–8. https://doi.org/10.1109/FG.2019.8756582
[595] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[596] Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W. Mahoney, and Kurt Keutzer. 2020. HAWQV3: Dyadic Neural Network Quantization. *CoRR* abs/2011.10680 (2020). arXiv:2011.10680 https://arxiv.org/abs/2011.10680

[597] Yusuke Yasuda, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2019. Investigation of Enhanced Tacotron Text-to-speech Synthesis Systems with Self-attention for Pitch Accent Language. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6905–6909. https://doi.org/10.1109/ICASSP.2019.8682353

[598] Feng Ye and Jun Yang. 2021. A deep neural network model for speaker identification. *Applied Sciences* 11, 8 (2021), 3603.

[599] Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380* (2021).

[600] Hao Yen, François G Germain, Gordon Wichern, and Jonathan Le Roux. 2022. Cold Diffusion for Speech Enhancement. *arXiv preprint arXiv:2211.02527* (2022).

[601] Reo Yoneyama, Ryuichi Yamamoto, and Kentaro Tachibana. 2022. Nonparallel High-Quality Audio Super Resolution with Domain Adaptation and Resampling CycleGANs. *arXiv preprint arXiv:2210.15887* (2022).

[602] Ji Won Yoon, Beom Jun Woo, and Nam Soo Kim. 2022. Hubert-ee: Early exiting hubert for efficient speech recognition. *arXiv preprint arXiv:2204.06328* (2022).

[603] Jaeseong You, Dalhyun Kim, Gyuhyeon Nam, Geumbyeol Hwang, and Gyeongsu Chae. 2021. Gan vocoder: Multi-resolution discriminator is all you need. *arXiv preprint arXiv:2103.05236* (2021).

[604] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al. 2019. Durian: Duration informed attention network for multimodal synthesis. *arXiv preprint arXiv:1909.01700* (2019).

[605] Dong Yu and Li Deng. 2016. *Automatic speech recognition*. Vol. 1. Springer.

[606] Fisher Yu and Vladlen Koltun. 2015. Multi-Scale Context Aggregation by Dilated Convolutions. *CoRR* abs/1511.07122 (2015).

[607] Yechan Yu, Dongkeon Park, and Hong Kook Kim. 2022. Auxiliary loss of transformer with residual connection for end-to-end speaker diarization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8377–8381.

[608] Fengpeng Yue, Yan Deng, Lei He, Tom Ko, and Yu Zhang. 2022. Exploring machine speech chain for domain adaptation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6757–6761.

[609] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. *Advances in neural information processing systems* 32 (2019).

[610] Neil Zeghidour and David Grangier. 2021. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2840–2849.

[611] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. 2019. But system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592* (2019).

[612] Jihen Zeremdini, Mohamed Anouar Ben Messaoud, and Aicha Bouzid. 2015. A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation. *Brain informatics* 2 (2015), 155–166.

[613] Albert Zeyer, André Merboldt, Wilfried Michel, Ralf Schlüter, and Hermann Ney. 2021. Librispeech transducer model with internal language model prior correction. *arXiv preprint arXiv:2104.03006* (2021).

[614] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. 2019. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6301–6305.

[615] Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In *International Conference on Machine Learning*. PMLR, 26193–26205.

[616] Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020. Adaptive feature selection for end-to-end speech translation. *arXiv preprint arXiv:2010.08518* (2020).

[617] Chunlei Zhang and Kazuhito Koishida. 2017. End-to-end text-independent speaker verification with triplet loss on short utterances.. In *Interspeech*. 1487–1491.

[618] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2018. Joint slot filling and intent detection via capsule neural networks. *arXiv preprint arXiv:1812.09471* (2018).

[619] Chen Zhang, Yi Ren, Xu Tan, Jinglin Liu, Kejun Zhang, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2021. Denoispeech: Denoising text to speech with frame-level noise modeling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7063–7067.

[620] Chunlei Zhang, Jiatong Shi, Chao Weng, Meng Yu, and Dong Yu. 2022. Towards end-to-end speaker diarization with generalized neural speaker clustering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8372–8376.

[621] Hanyi Zhang, Longbiao Wang, Kong Aik Lee, Meng Liu, Jianwu Dang, and Hui Chen. 2021. Meta-learning for cross-channel speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5839–5843.

[622] Hanyi Zhang, Longbiao Wang, Kong Aik Lee, Meng Liu, Jianwu Dang, and Helen Meng. 2023. Meta-Generalization for Domain-Invariant Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1024–1036.

[623] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. 2018. Forward attention in sequence-to-sequence acoustic modeling for speech synthesis. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4789–4793.

[624] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. 2019. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), 540–552.

[625] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. 2019. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 3 (2019), 631–644.

[626] Jing-Xuan Zhang, Li-Juan Liu, Yan-Nian Chen, Ya-Jun Hu, Yuan Jiang, Zhen-Hua Ling, and Li-Rong Dai. 2020. Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer. *arXiv preprint arXiv:2009.01475* (2020).

[627] Lichao Zhang, Yi Ren, Liqun Deng, and Zhou Zhao. 2022. Hifidenoise: High-fidelity denoising text to speech with adversarial networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7232–7236.

[628] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7829–7833.

[629] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6, 1 (2019), 1–23.

[630] Xingxuan Zhang, Feng Cheng, and Shilin Wang. 2019. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 713–722.

[631] Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2022. TDASS: Target Domain Adaptation Speech Synthesis Framework for Multi-speaker Low-Resource TTS. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN55064.2022.9892596

[632] Ying Zhang, Hao Che, and Xiaorui Wang. 2021. Non-parallel Sequence-to-Sequence Voice Conversion for Arbitrary Speakers. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362095

[633] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *arXiv preprint arXiv:2303.01037* (2023).

[634] Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou, Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang, Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran, Tara N. Sainath, Françoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu, Ruoming Pang, and Yonghui Wu. 2022. BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1519–1532. https://doi.org/10.1109/JSTSP.2022.3182537

[635] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504* (2020).

[636] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926* (2023).

[637] Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2020. NeurST: Neural speech translation toolkit. *arXiv preprint arXiv:2012.10018* (2020).

[638] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A non-native English speech corpus.. In *Interspeech*. 2783–2787.

[639] Hongyu Zhao, Hao Tan, and Hongyuan Mei. 2022. Tiny-Attention Adapter: Contexts Are More Important Than the Number of Parameters. *arXiv preprint arXiv:2211.01979* (2022).

[640] Shengkui Zhao and Bin Ma. 2023. MossFormer: Pushing the Performance Limit of Monaural Speech Separation using Gated Single-Head Transformer with Convolution-Augmented Joint Self-Attentions. *arXiv preprint arXiv:2302.11824*

(2023).

[641] Shengkui Zhao, Trung Hieu Nguyen, and Bin Ma. 2021. Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6648–6652.

[642] Shengkui Zhao, Hao Wang, Trung Hieu Nguyen, and Bin Ma. 2021. Towards natural and controllable cross-lingual voice conversion based on neural tts model and phonetic posteriorgram. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5969–5973.

[643] Wei Zhao and Zheng Yang. 2023. An Emotion Speech Synthesis Method Based on VITS. *Applied Sciences* 13, 4 (2023), 2225.

[644] Xing Zhao, Shuang Yang, Shiguang Shan, and Xilin Chen. 2020. Mutual information maximization for effective lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 420–427.

[645] Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu. 2021. Interactive speech and noise modeling for speech enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14549–14557.

[646] Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning*. PMLR, 12736–12746.

[647] Yibin Zheng, Xinhui Li, Fenglong Xie, and Li Lu. 2020. Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6734–6738.

[648] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9299–9306.

[649] Pan Zhou, Ruchao Fan, Wei Chen, and Jia Jia. 2019. Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding. *arXiv preprint arXiv:1911.00203* (2019).

[650] Donghui Zhu and Ning Chen. 2022. Multi-Source Domain Adaptation and Fusion for Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2103–2116.

[651] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. 2018. Arbitrary talking face generation via attentional audio-visual coherence learning. *arXiv preprint arXiv:1812.06589* (2018).

[652] Hongning Zhu, Kong Aik Lee, and Haizhou Li. 2021. Serialized multi-layer multi-head attention for neural speaker embedding. *arXiv preprint arXiv:2107.06493* (2021).

[653] Xiaolian Zhu, Shan Yang, Geng Yang, and Lei Xie. 2019. Controlling Emotion Strength with Relative Attribute for End-to-End Speech Synthesis. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 192–199. https://doi.org/10.1109/ASRU46091.2019.9003829