A Single Self-Supervised Model for Many Speech Modalities Enables Zero-Shot Modality Transfer

Anonymous Author(s)
Affiliation
Address
email

Abstract

While audio-visual speech models can yield superior performance and robustness compared to audio-only models, their development and adoption are hindered by the lack of labeled and unlabeled audio-visual data and the cost to deploy one model per modality. In this paper, we present u-HuBERT, a self-supervised pre-training framework that can leverage both multimodal and unimodal speech with a unified masked cluster prediction objective. By utilizing modality dropout during pre-training, we demonstrate that a single fine-tuned model can achieve performance on par or better than the state-of-the-art modality-specific models. Moreover, our model fine-tuned only on audio can perform well with audio-visual and visual speech input, achieving zero-shot modality generalization for speech recognition and speaker verification. In particular, our single model yields 1.2%/1.4%/27.2% speech recognition word error rate on LRS3 with audio-visual/audio/visual input.

1 Introduction

Speech processing has been moving from developing domain-specific models toward building a general model for all domains. For example, early studies presented models dedicated for robust speech recognition [Kingsbury et al., 1998] and distant speech recognition [Feng et al., 2014, Kim et al., 2017] by taking into account acoustic properties for each domain. In contrast, recent work has built general domain models by pooling data from different domains and even different languages, including labeled ones for supervised learning [Chan et al., 2021, Likhomanenko et al., 2020, Pratap et al., 2020] and unlabeled ones for self-supervised learning [Kawakami et al., 2020, Hsu et al., 2021b, Conneau et al., 2020]. These studies show that in addition to removing the need of deploying and maintaining multiple models, a single general domain model can in fact achieve performance on par or superior to domain-specific models, especially for domains with less data, and improve generalization performance to unseen domains.

In contrast, there are few attempts and even fewer successful ones on developing a versatile model that can process many speech modalities [Makino et al., 2019, Sari et al., 2021]. In general, researchers develop models with different input speech modalities, such as visual speech [Chung and Zisserman, 2016], audio-visual speech [Xu et al., 2020], multi-channel speech [Haeb-Umbach et al., 2021], 3D talking faces [Richard et al., 2021], Electromyographic (EMG) muscle signals [Diener and Schultz, 2018], ultrasonic oral cavity scan [Hueber et al., 2010], or ultrasonic microphone [Livescu et al., 2009] in isolation, despite the fact that many of them are used to solve the same task, like speech and speaker recognition. Moreover, while additional modalities, such as visual speech, can often boost the performance and robustness on top of audio speech, models taking such input are less studied and deployed in real world application. This is mainly due to the lack of both labeled and unlabeled data for these alternative modalities, hindering the development of high-performing self-supervised or supervised modal-specific models that are trained exclusively on one type of input. We argue that to

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

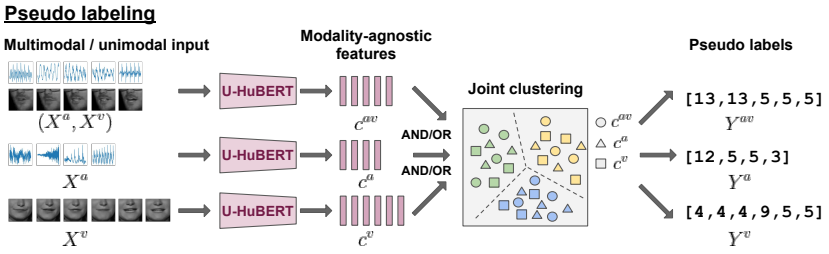

Figure 1: u-HuBERT learns modality-agnostic features that are clustered to produce pseudo labels with a shared codebook for different speech modalities. See Section 3.2 for more information.

tackle the data scarcity issue, it essential to utilize not only the speech from the target modality, but also speech in other modalities to build a general speech model for many modalities. By doing so, more data can be used to learn components shared across different modalities, which is expected to bring bigger gains to low-resource modalities.

In this paper, we present unified hidden unit BERT (u-HuBERT), which leverages unlabeled speech data of many different modalities for pre-training, including both unimodal and multimodal speech. Architecture-wise, u-HuBERT consists of a set of modality-specific feature extractors and a shared transformer backend. The model is trained with a unified masked cluster prediction objective for both unimodal and multimodal speech data. Specifically, we adopt modality dropout [Neverova et al., 2014] for multimodal data to simulate different combination of input modalities and encourage the model to learn modality-agnostic frame-level representations. The pre-trained u-HuBERT can be fine-tuned with either multimodal or unimodal data, and we demonstrate that in both cases the resulting model can be used for many speech modalities. We verify the effectiveness on two popular speech processing tasks: speech recognition and speaker verification. To the best of our knowledge, this is the first single model for many speech modalities that yields performance on par with the state-of-the-art modality-specific models and is capable of zero-shot modality generalization.

2 Related Work

Comparison to AV-HuBERT Our work is an extension of AV-HuBERT, which studies pre-training exclusively with audio-visual data. While the model architecture and the form of pre-training objective are the same, u-HuBERT generalizes AV-HuBERT to utilize both multimodal and unimodal speech for pre-training by mapping various input to a modality-agnostic feature space and creating a shared target space for masked prediction. We empirically demonstrate the benefits of including additional unimodal data for pre-training. The fine-tuning and evaluation protocols also differ significantly. AV-HuBERT effectively learns modality-specific models (e.g., visual speech recognizer) given labeled data in the target modality for fine-tuning. In contrast, this work focuses on building a single model for all modalities seen during pre-training with labeled data in any subset of modalities.

Joint unimodal and multimodal pre-training Recent cross-modal contrastive approaches such as CLIP [Radford et al., 2021] and ALIGN [Jia et al., 2021] have shown impressive results yet they can only be trained with multimodal data and require a large-scale dataset, which can be hard to obtain for many speech modalities (e.g., paired EMG-audio speech). FLAVA [Singh et al., 2021] and SLAM [Bapna et al., 2021] are two frameworks that utilize both unimodal and multimodal data for image-text and speech-text pre-training, respectively. There are three key differences between u-HuBERT and these two models. First, for multimodal speech, different modalities such as audio and visual speech are temporally aligned. Instead of concatenating the feature sequences from unimodal encoders, we fuse the features frame-by-frame before passing it to the shared encoder. Second, while all three models utilize a masked token prediction objective, u-HuBERT considers a shared target token space for all unimodal and multimodal data, obtained by clustering modality-agnostic features extracted from the u-HuBERT model itself. In contrast, SLAM uses w2v-BERT [Chung et al., 2021] tokens for speech and SentencePiece [Schuster and Nakajima, 2012] for text. FLAVA on other hand uses a separately trained dVAE tokenizer for image and the same as SLAM for text, and it predicts tokens for unimodal and multimodal data at the output of different modules, differing from u-

HuBERT. In addition, both FLAVA and SLAM pre-train with a number of auxiliary objectives, while u-HuBERT only uses one, which simplifies hyperparameter selection. Third, different modalities of speech are often used to solve the same task, such as speech recognition or speaker verification. Our evaluation focuses on building a single model that can solve the same task with different modalities despite not having labeled data for each modality, which is not considered by SLAM or FLAVA.

Single model for many input modalities Makino et al. [2019] and Sari et al. [2021] explored building a single model for supervised audio-visual/audio/visual speech recognition and speaker verification, respectively. Both studies show significant degradation comparing unified models with modality-specific ones. Furthermore, these models are single-task supervised and hence not capable of zero-shot modality transfer when trained on unimodal data. Omnivore [Girdhar et al., 2022] is a single supervised model for many vision modalities, including image (RGB), 3D image (RGBD), and video (RGB sequence), which also considers a model architecture with modality-specific encoders (RGB and depth) followed by a shared Swin Transformer backend [Liu et al., 2021]. Transformer [Liu et al., 2021] to learn shared representations across modalities. Both Omnivore and u-HuBERT demonstrates cross-modal generalization, but Omnivore achieves it through supervised learning with different tasks, and u-HuBERT achieves it with unified self-supervised learning. Finally, if we considered different languages as different modalities, our work is also related to mBERT [Devlin et al., 2019], mBART [Liu et al., 2020], TLM [Yao et al., 2021], and XLM [Lample and Conneau, 2019], fine-tuning which enables generalizing natural language understanding [Pires et al., 2019] and machine translation [Johnson et al., 2017] to languages without labeled data. Pre-training on multimodal speech is analogous to pre-training on bitext, while the former requires less manual effort to obtain by collecting more measurements when speech is produced. In addition, time-synchronicity between speech modalities provides additional self-supervision, enabling u-HuBERT to more easily learn frame-level modality-agnostic representations.

3 Method

3.1 Background: audio-visual HuBERT

We build our framework upon AV-HuBERT [Shi et al., 2022a], a self-supervised representation learning model designed to pre-train a single type of multimodal speech data: audio-visual speech. AV-HuBERT first encodes temporally aligned audio speech $X^a \in \mathbb{R}^{T_a \times C_a}$ and visual speech $X^v \in \mathbb{R}^{T_v \times C_v \times H \times W}$ into feature sequences $h^a \in \mathbb{R}^{T \times D_h}$ and $h^v \in \mathbb{R}^{T \times D_h}$ of the same lengths, which are then fused frame-by-frame $\texttt{FUSE}(h^a, h^v) \in \mathbb{R}^{T \times D_c}$ and processed by a Transformer backend [Vaswani et al., 2017] to produce contextualized audio-visual representations $c^{av} \in \mathbb{R}^{T \times D_c}$.

This model is trained by iterating two steps. First, a K-means clustering model is trained on some audio-visual features to produce frame-level targets $Y^{av} \in [1, \cdots, K]^T$, where K is the number of clusters. Second, the AV-HuBERT model is trained with a masked cluster prediction task, where spans from the audio-visual input are randomly masked and the model learns to predict the cluster assignment of the masked frames given the context. The model is forced to learn the underlying linguistic structure of audio-visual speech to infer the content in the unseen regions based on the context. The intermediate feature of a trained model, which is better than previous iteration, is used to produce targets for the subsequent iteration. The pre-trained AV-HuBERT model performs well on both multimodal downstream tasks like audio-visual speech recognition [Shi et al., 2022b] and unimodal tasks like lip-reading [Shi et al., 2022a] when fine-tuned for a specific modality.

3.2 Self-supervised pre-training with many modalities

We extend the framework to utilize not only multimodal data, but also unimodal data that can be more easily obtained in large quantities. The main idea is to apply modality dropout to multimodal speech along with a shared Transformer [Vaswani et al., 2017] to learn modality-agnostic features that can be quantized by a shared codebook to produce pseudo labels for masked prediction.

Given unimodal and multimodal speech data from many modalities, we assume that there is an anchor modality that is paired with all other modalities. For simplicity, assume the anchor modality is audio, and we have unlabeled audio, visual, and audio-visual speech. In the first iteration, because a modality-agnostic feature extractor is not available yet, we cluster the anchor modality features to produce targets for unimodal and multimodal speech that contain the anchor modality. Concretely speaking,

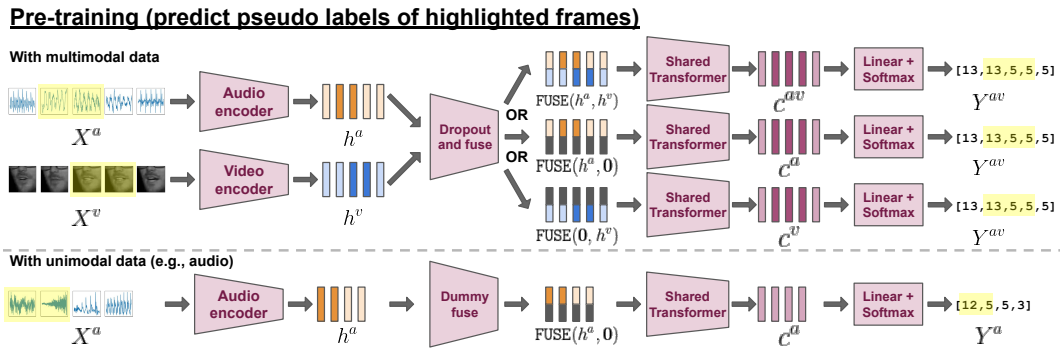

Figure 2: u-HuBERT pre-trains on multimodal and unimodal speech with a unified objective: predicting cluster assignments of the frames that are masked at the input (highlighted by yellow shades). For multimodal data, a subset of modalities are randomly selected to predict the same targets.

this produces pseudo paired data for audio speech X^a and audio-visual speech (X^a, X^v) with audio-based cluster assignment Y^a as the target. Like AV-HuBERT, each modality has its own feature extractor: f^a for audio and f^v for video, and feature sequences are fused frame-by-frame before being passed onto the Transformer g to produce contextualized features $c^{av} = g(\texttt{FUSE}(f^a(X^a), f^v(X^v)))$. When a modality is absent, its feature sequence is simply replaced by zero vectors, for example, $c^a = g(\texttt{FUSE}(f^a(X^a), \mathbf{0}))$. Following HuBERT [Hsu et al., 2021a], this model is trained with a masked prediction objective.

To encourage the model to learn modality-agnostic features, modality dropout is applied to multimodal data to randomly drop a subset of the modalities, effectively creating multiple copies of the data with the same target but different input modalities, as depicted in the top half of Figure 2. While modality dropout was also used in AV-HuBERT, it was implemented for a different reason – to reduce the discrepancy between pre-training and fine-tuning for unimodal downstream tasks, which resulted in only minor improvement (reducing the WER from 57.0% to 55.3% as shown in Table D.1 in Shi et al. [2022a]). In contrast, we will demonstrate that modality dropout is essential for mixed modality pre-training and zero-shot generalization.

After the first iteration, the model can be used to extract modality-agnostic features for all unimodal and multimodal data, which are clustered to produce the pseudo label for next iteration of pre-training as illustrated in Figure 1. For multimodal data like audio-visual speech, one can produce pseudo labels using features extracted from the complete multimodal input (e.g., Y^{av} from audio-visual input), or those extracted from input with a subset of modalities (e.g., Y^v using only the video stream). We opt for the former and study the impact of these choices in Secton 4.2.

3.3 Multimodal and unimodal fine-tuning with partial modalities

Once the model is pre-trained, we remove the cluster prediction head (the linear layer on top of transformer used to predict cluster assignment distribution) and add a downstream task-specific prediction head. The pre-trained model can be fine-tuned on labeled multimodal speech, unimodal speech, or speech with mixed modalities. More importantly, the modalities included in the fine-tuning data do not necessarily cover the modalities seen during pre-training, yet the fine-tuned model can still handle the downstream task on all pre-trained modalities, which we refer to as *zero-shot modality generalization*. This is because the pre-trained model has successfully learned modality-agnostic representations, where same speech manifested in different modalities are mapped to similar representations. Thus mapping from the representation of one modality (e.g., audio) to a target output (e.g., text) can be directly applied to other pre-trained modalities (e.g., video).

Similar to mBART [Liu et al., 2020] for zero-shot machine translation, catastrophic forgetting [Lee et al., 2019, Üstün et al., 2021] may occur during fine-tuning and hence hinders its generalization across modalities. Particularly, this is because the distributions of Transformer input and output at earlier layers are less modality-invariant. To alleviate the issue, we apply modality dropout when multimodal data is used, and explore a few techniques to prevent the representations from drifting too much, including freezing the entire pre-trained model for a number of steps and freezing a number

of initial layers (up to where more modality-agnostic representations are produced) throughout fine-tuning.

4 Experiments

4.1 Pre-training setup

We evaluate u-HuBERT on audio, visual, and audio-visual speech, where larger quantities of unlabeled and labeled data are available. Two audio-visual and one audio-only datasets are used for pre-training: (1) **LRS3** [Afouras et al., 2018] with 433 hours of English audio-visual speech, (2) **VoxCeleb2-En (VC2-En)** with 1,326 hours of English YouTube audio-visual speech filtered from VoxCeleb2 [Chung et al., 2018] by Shi et al. [2022a], and (3) **TED-LIUM release 3 (TD)** [Hernandez et al., 2018] with 452 hours of English audio collected from the same domain as LRS3. To improve noise robustness, we apply online noise augmentation where each utterance is corrupted with additive noise at 0dB sampled from MUSAN [Snyder et al., 2015] with a probability of 0.25. More details can be found in Appendix A. We adopt the AV-HuBERT-LARGE model Shi et al. [2022a][1]: the video encoder is a modified ResNet-18 model [He et al., 2016] that takes lip region-of-interest (ROI) frames sampled at 25Hz at input and produce a sequence of 512 dimensional embeddings at the same frame rate; audio is represented as filter bank features sampled at 100 Hz, stacked every four frames to match the video frame rate, and projected with a linear layer to 1024-D features. `FUSE` concatenates features frame-by-frame and linearly projects them to the embedding dimension of the shared Transformer. The shared Transformer has 24 layers, with 16 heads, embedding dimension 1024, and FFN dimension 4096, using the pre-norm residual connection setup [Nguyen and Salazar, 2019].

To directly compare with the publicly available fifth iteration noise-augmented AV-HuBERT-LARGE model trained on LRS3 and VC2-En, the same feature extractor (fourth iteration BASE model) and the same codebook (learned from multimodal features C^{av}) are used to generate pseudo labels for multimodal LRS3 and VC2-En and unimodal TD. We then pre-train u-HuBERT on the combined data for 1M updates on 64 32GB V100 GPUs using the Adam optimizer [Kingma and Ba, 2015] and a learning rate of 0.002. Gradient norm is clipped at 1.0. to stablize training. A batch size of maximal 40 seconds per GPU is used. When pre-trained on multimodal data, audio and video are dropped with a probability of 0.25 and 0.25, respectively.

4.2 Learning modality-agnostic features and codebooks with modality dropout

We argue that modality dropout leads to learning modality-agnostic representations through forcing the model to predict the same target with different subsets of input modalities given multimodal input, which in turn enables construction of a shared codebook for all modalities. To verify this, we compare two models that are identical except for whether modality dropout is activated: the public noise-augmented AV-HuBERT LARGE model trained on LRS3 and VC2-En with modality dropout and its counterpart. For each model, we infer audio, visual, and audio-visual features for each LRS3 validation utterances, randomly sample 500 frames, pool the features, and run t-SNE [Van der Maaten and Hinton, 2008] dimension reduction for visualization (Figure 3). It is visually evident that without modality dropout, features extracted from different modalities have different distributions (bottom row), contrary to those from the model with modality dropout (top row).

We next adopt the phone normalized mutual information (PNMI) metric presented in Hsu et al. [2021a] to quantify cluster quality. For each model, we create four codebooks by clustering audio-visual features C^{av}, audio features C^{a}, visual features C^{v}, or all three combined $\cup_m C^m$, quantize each type of features with each codebook, and measure the normalized mutual information $\in [0, 1]$ between frame-level phone labels and cluster assignment (Table 1). Because features inferred from different modalities exhibit similar distributions with modality dropout, all four codebooks lead to similar PNMI for a given feature. In contrast, without modality dropout, cross-quantization (e.g., Y^{av} with C^{v} codebook) often leads to significantly worse PNMI, and PNMI is worse for all *(codebook, feature)* combinations compared with that from the modality dropout model. Note for a given codebook, the PNMI still differs among modalities which is due to the discrepancy of information carried per modality.

[1] `https://github.com/facebookresearch/av_hubert/blob/main/avhubert/conf/pretrain/noise_large_vox_iter5.yaml`

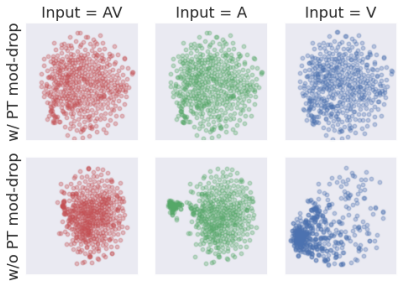

Figure 3: t-SNE projection of AV/A/V speech representations learned by models differing in modality dropout.

K-means training data	PNMI (↑) w/ PT mod-drop Y^{av}	Y^{a}	Y^{v}	w/o PT mod-drop Y^{av}	Y^{a}	Y^{v}
$\cup_m C^m$	43.60	43.72	38.01	41.89	36.59	21.78
C^{av}	42.89	42.96	36.89	42.83	38.46	26.17
C^{a}	43.18	43.28	37.11	42.87	37.28	20.99
C^{v}	42.23	42.57	37.68	6.85	13.68	23.54

Table 1: Cluster quality (PNMI) of each modality $Y^{av}/Y^{a}/Y^{v}$ using codebooks learned from features of different modalities, obtained from models pre-trained with and without modality dropout.

4.3 Speech recognition

Fine-tuning setup LRS3 *trainval* and *pretrain*, combined for 433 hours, are used for training, with the same 1,200 utterances as Shi et al. [2022a] split for validation. Results on the *test* split are reported. Noise augmentation with MUSAN used in pre-training is also adopted during fine-tuning. A 9-layer randomly initialized Transformer decoder with 8 attention heads and 1024/4096-D embedding/FFN is appended to the pre-trained u-HuBERT, and the entire model is fine-tuned with a cross-entropy loss predicting the next text token given the input and the previous text tokens. Text is represented as unigram-based subword units [Kudo, 2018] with a vocabulary size of 1,000. Modality dropout is applied when fine-tuning on multimodal input. Additionally, we sweep learning rate, number of updates, steps to freeze, and layers to freeze throughout fine-tuning. We evaluate each model, regardless of the modality used for fine-tuning, on all kinds of input modalities and report the word error rates (WERs) without fusing external language models. To gauge the robustness to noise, we follow the same protocol in Shi et al. [2022b] to noisy audio and audio-visual test sets by adding babble noise at 0 decibel (dB) to the audio stream for evaluation. Detailed hyperpamater setups can be found in Appendix A, and more ablation studies can be found in Appendix B.1.

Modality dropout is essential for building a single model Table 2 compares fine-tuning from scratch, as well as from pre-trained models with and without modality dropout, using labeled audio, visual, or audio-visual speech, and highlights the results of zero-shot modality transfer where labeled data in test modality are not used. First, models without pre-training are significantly worse. Second, when fine-tuning on audio-visual data without modality dropout, performance drops significantly on unimodal data if pre-training modality dropout is not applied (e.g., 1.7% to 21.4% on clean audio). This can be alleviated with fine-tuning modality dropout, improving the average WER from 24.5% to 12.3%, but is still behind applying it in both pre-training and fine-tuning (11.4% average).

The benefit of applying pre-training modality dropout is magnified in unimodal fine-tuning: the model fine-tuned only on labeled audio data yields the same performance on clean and noisy audio-visual input (1.3% and 4.6%) as the model fine-tuned with labeled audio-visual data, while being slight behind on video input compared to the model fine-tuned with labeled video data (31.6% vs. 28.7%). When it is compared to the model pre-trained without modality dropout and fine-tuned also with labeled audio, huge gains are observed on zero-shot scenarios, for example 96.8% versus 31.6% V-WER and 18.0% versus 4.6% noisy AV-WER. Similar trend can be observed when fine-tuning with labeled visual speech.

Zero-shot modality transfer is not achieved through memorization Because the dataset used for fine-tuning is a subset of the pre-training data, there could be suspicions that the model memorizes the audio-visual pairs from pre-training and associates them with the labeled unimodal data to enable zero-shot transfer. To avoid cheating, we fine-tune our model on LibriSpeech [Panayotov et al., 2015], an audio-only read audiobook dataset that is out-of-domain relative to the pre-training data composed of oratory talks (Table 3). Despite the degradation compared to fine-tuning on in-domain LRS3 data, the model still performs decently on zero-shot transfer scenarios compared to models pre-trained without modality dropout in Table 2.

Table 2: Speech recognition results on LRS3 test. PT indicates if a model is pre-trained (on LRS3 and VC-2). PT mod-drop-p and FT-drop-p denote if modality dropout is applied in pre-training and fine-tuning, respectively. FT mod denotes the labeled data used for fine-tuning: audio-visual (AV), audio (A), or visual (V). Zero-shot scenarios are highlighted with blue shades.

PT	PT mod-drop	FT mod	FT mod-drop	AV-WER Clean	AV-WER Noisy	A-WER Clean	A-WER Noisy	V-WER	Avg-WER
fine-tuned on 433h									
✗	n/a	AV	✗	3.8	17.2	28.2	87.6	83.4	44.0
✓	✗	AV	✗	1.3	4.8	21.4	52.6	42.3	24.5
✓	✓	AV	✗	1.2	5.2	1.7	25.5	32.4	13.2
✗	n/a	AV	✓	3.6	15.9	4.6	44.8	63.7	26.5
✓	✗	AV	✓	1.3	4.1	1.8	23.1	31.0	12.3
✓	✓	AV	✓	1.3	4.6	1.5	20.5	29.1	11.4
✗	n/a	A	n/a	✗	✗	4.0	37.3	✗	✗
✓	✗	A	n/a	1.5	18.0	1.6	20.9	96.8	27.8
✓	✓	A	n/a	1.3	4.6	1.4	19.3	31.6	11.6
✗	n/a	V	n/a	✗	✗	✗	✗	60.3	✗
✓	✗	V	n/a	11.3	21.8	80.3	97.7	28.0	47.8
✓	✓	V	n/a	2.1	5.1	2.3	20.9	28.7	11.8

Table 3: Fine-tuning on out-of-domain labeled audio speech that does not overlap with the pre-training data, and evaluating on LRS3 *test*. Zero-shot scenarios are highlighted with blue shades.

PT data	PT mod-drop	FT data	FT mod	AV (Clean)	AV (Noisy)	A (Clean)	A (Noisy)	V
LRS3+VC2-En	✗	LibriSpeech	A	6.2	11.0	20.9	65.9	64.7
LRS3+VC2-En	✓	LibriSpeech	A	4.5	8.9	4.8	27.6	39.4

Comparison with state-of-the-art models Table 4 compares u-HuBERT with models from the literature, all of which are modality-specific. In contrast, u-HuBERT is the only single model that can be fine-tuned on *any* modality and test on *all* modalities while achieving performance on par or better than the best modality-specific models. Furthermore, compared to AV-HuBERT, u-HuBERT can be pre-trained on additional unimodal data to yield even better performance.

Table 4: Comparison with state-of-the-art audio, visual, and audio-visual speech recognition results reported on LRS3. Zero-shot scenarios are highlighted with blue shades. † refers to *filtered* public datasets where the list of used utterances are not public. ‡ refers to non-public datasets.

Method	Unlab data AV (hr)	Unlab data A (hr)	Lab data Mod	Lab data Dataset (hr)	WER (%) AV	WER (%) A	WER (%) V
supervised learning							
Ma et al. [2021]	-	-	AV	LRW, LRS3 (590)	2.3	✗	✗
Ma et al. [2021]	-	-	A	LRW, LRS3 (590)	✗	2.3	✗
Ma et al. [2021]	-	-	V	LRW, LRS3 (590)	✗	✗	43.3
Xu et al. [2020]	-	-	AV	LRW, LRS3 (590)	6.8	✗	✗
Ma et al. [2022]	-	-	V	LRW, LRS2, LRS3, AVSpeech† (1,459)	✗	✗	31.5
Afouras et al. [2021]	-	-	V	LRS2, LRS3, MV-LRS‡, TEDx$^{\ddagger}_{ext}$ (2,676)	✗	✗	30.7
Makino et al. [2019]	-	-	AV	YT31k‡ (31,000)	4.5	✗	✗
Makino et al. [2019]	-	-	A	YT31k‡ (31,000)	✗	4.8	✗
Makino et al. [2019]	-	-	V	YT31k‡ (31,000)	✗	✗	33.6
Serdyuk et al. [2021]	-	-	V	YT90k‡ (90,000)	✗	✗	25.9
Serdyuk et al. [2021]	-	-	V	LRS3, YT90k‡ (90,433)	2.3	✗	✗
semi-supervised and self-supervised learning							
[Shi et al., 2022b]	LRS3, VC2-En (1,759)	-	AV	LRS3 (433)	1.4	✗	✗
[Shi et al., 2022b]	LRS3, VC2-En (1,759)	-	A	LRS3 (433)	✗	1.6	✗
[Shi et al., 2022a]	LRS3, VC2-En (1,759)	-	V	LRS3 (433)	✗	✗	28.6
Afouras et al. [2020]	VC2-clean† (334)	-	V	LRW, LRS3 (590)	✗	✗	59.8
u-HuBERT	LRS3, VC2-En (1,759)	-	AV	LRS3 (433)	1.3	1.5	29.1
u-HuBERT	LRS3, VC2-En (1,759)	-	A	LRS3 (433)	1.3	1.4	31.6
u-HuBERT	LRS3, VC2-En (1,759)	-	V	LRS3 (433)	2.1	5.1	28.7
u-HuBERT	LRS3, VC2-En (1,759)	TD (452)	AV	LRS3 (433)	1.2	1.4	27.2

4.4 Speaker verification

We repeat the experiment on speaker verification to study if u-HuBERT enables construction of a single model and achieves zero-shot generalization on a different task.

Fine-tuning setup We follow the protocol of SUPERB [Yang et al., 2021], a standard benchmark for self-supervised speech representation evaluation. Specifically, we use VoxCeleb1 (VC1) [Nagrani et al., 2017] which contains 352 hours of speech from 1,251 speakers for fine-tuning. Here the utterance-level speaker identity is the target label. For visual modality, we use the face image frames provided by Nagrani et al. [2018] and follow the same preprocessing steps as Shi et al. [2022c] to get the lip regions. Noise augmentation following Shi et al. [2022b,c] is used to simulate noisy environments. The noise is sampled from {Babble, Speech, Music, Noise} categories in MUSAN corpus [Snyder et al., 2015] at SNR in {-10, -5, 0, 5, 10} dB and we report the average performance over 20 configurations.

Speaker verification aims to evaluate how well the *utterance-level* representation captures speaker similarity based on some distance metric in the space. An X-vector [Snyder et al., 2018] network is added following the SUPERB protocol, which first weighted-sums frame-level u-HuBERT representations over all Transformer layers, and then pools them temporally into a sequence-level representation. The X-vector is randomly initialized and only added during fine-tuning. The model is trained with the widely used AMS-Softmax loss [Wang et al., 2018] with scale and margin being set to 0.4 and 30 respectively. For evaluation, we compute the cosine similarity between two utterances and measure the equal error rate (EER). During fine-tuning, the u-HuBERT model is frozen while only the downstream X-Vector network is fine-tuned. The model is trained with Adam [Kingma and Ba, 2015]. The learning rate is searched within the range of [1e-7, 1e-1] and is kept fixed throughout training. Note the above fine-tuning and evaluation paradigm is based on SUPERB, which is for a direct comparison with prior works on the same benchmark. See Appendix A and B.2.

Table 5: Speaker verification results on VC1 test. Headers are the same as Table 2. Only pre-trained models are considered, because representation extractors are not updated following SUPERB.

PT	PT mod-drop-p	FT mod	FT mod-drop-p	AV-EER Clean	AV-EER Noisy	A-EER Clean	A-EER Noisy	V-EER	Avg-EER
fine-tuned on VC1									
✓	✗	AV	✗	2.46	5.49	7.40	23.44	10.73	9.90
✓	✓	AV	✗	2.75	6.00	4.67	14.75	9.60	7.55
✓	✗	AV	✓	**2.31**	**4.09**	6.94	23.12	6.69	8.63
✓	✓	AV	✓	2.54	5.20	**4.23**	**14.72**	6.53	6.64
✓	✗	A	n/a	4.63	9.02	6.70	22.38	20.03	12.56
✓	✓	A	n/a	3.85	8.92	4.25	14.46	17.90	9.88
✓	✗	V	n/a	10.34	9.57	24.57	35.18	6.30	17.19
✓	✓	V	n/a	8.60	9.99	15.35	23.05	**5.01**	12.40

Zero-shot performance Good audio-visual speaker representations aim to capture a number of related features, such as voice, appearance, and speaking style (wording, intonation, rhythm). However, in contrast to speech recognition where content is highly correlated with each modality, for speaker verification each modality by itself can only infer a subset of these speaker features (e.g., appearance can barely be inferred from audio). This imposes an inherent limitation to zero-shot transfer.

We present the results in Table 5. We found u-HuBERT fine-tuned with audio-only data can be used for audio-visual speaker verification and yield better performance. In particular, it achieves 3.85%/8.92% EER with audio-visual input compared to 4.25%/14.46% with audio input on clean/noisy test set. With labeled audio data, the model learns to map voice and speaking style features to speaker representations. Hence, when tested with additional visual input, the model gains robustness on inferring voice and speaking style features, such that the performance on noisy test set is significantly improved (14.46% $\rightarrow$ 8.92% EER). However, this model does not learn to map appearance features (which could still be preserved in early Transformer layers that the X-vector head also takes as input [Chang et al., 2022]) to speaker representations. As a result, the model performs much worse with video input (17.90%) compared to the model fine-tuned on visual speech (5.01%), which learns to map appearance and potentially speaking style features to speaker representations. Similarly, the

audio fine-tuned model is also slightly behind audio-visual fine-tuned model when tested on audio-visual input (3.85% vs. 2.54%), as the latter could learn to map all features to speaker representations. On the other hand, when fine-tuning u-HuBERT with only visual speech, the model struggles to generalize for audio-visual and audio input because mapping from voice to speaker is not learned.

Effect of modality dropout Modality dropout plays an important role in building a single model with multimodal labeled data and enabling zero-shot transfer with only unimodal labeled data. In unimodal fine-tuning (audio-only or video-only), the performance consistently drops across all settings but one (noisy AV-EER of model fine-tuned on V) when the model is pre-trained without modality dropout. In audio-visual fine-tuning, modality dropout greatly improves the performance of a unified audio-visual model under single-modal test setting (A-EER: 7.40% $\rightarrow$ 4.23%, V-EER: 10.73% $\rightarrow$ 6.53%). Generally, the single model trained with modality dropout in both pre-training and fine-tuning achieves a comparable performance compared to modality-specific systems (2.54% vs. 2.31% in clean AV-EER, 4.23% vs. 4.25% in clean A-EER, 6.52% vs. 5.00% in V-EER). The value of dropout can be tuned in order to favor test performance of one specific modality, which we show in the Appendix B.1.1.

5 Discussion

We assume that there is an anchor modality where all other modalities appear in some multimodal data paired with it. This enables the model to predict shared targets and learn modality-agnostic representations for all modalities, such that non-anchor unimodal data can be used in the subsequent iterations. While this may seem limiting, most multimodal speech are paired with audio [Richard et al., 2021, Hueber et al., 2010, Richmond et al., 2011, Livescu et al., 2009] since audio is the primary measurement for speech and other modalities are often supplementary. This makes the assumption of having audio as the anchor modality practical. Furthermore, the idea can be easily extended to settings where any modality can be connected to the anchor via some intermediate modality. In short, a modality that is N-hop away from the primary anchor can be used in the $(N + 1)$-th iteration.

The only case not supported u-HuBERT is pre-training on modalities that cannot be connected through any multimodal data, for example, on a collection of audio, visual, audio-visual and EMG speech. A shared codebook cannot be derived for EMG and the rest of the modalities, and it is unclear if the model can still benefit from joint training. The setup has also been found much more challenging on other pairs of modalities: for example, SLAM Bapna et al. [2021] shows that without paired text and speech data, joint pre-training on unlabeled speech and text degrades the performance compared to a model pre-trained only on speech. We leave such exploration to future work. In addition, we also include ethical discussion in Appendix C.

6 Conclusion

Recent studies in applied machine learning trend toward building things that are *general*, for example, general perception module [Jaegle et al., 2021], general self-supervised objective [Baevski et al., 2022], single model for many tasks [Brown et al., 2020, Kaplan et al., 2020, Aghajanyan et al., 2022] and/or many input [Girdhar et al., 2022, Hu and Singh, 2021], instead of optimizing for performance on very specific setups. Our work contributes to this line of research by presenting a unified self-supervised objective to pre-train a model on speech of many modalities. Specifically, it can be fine-tuned for many different speech tasks on labeled data in any combination of pre-trained modalities, resulting in a single model that can process all combinations of pre-trained modalities to perform the fine-tuned task.

We envision our work can bring substantial benefit to multimodal speech processing. Given there are significantly more unlabeled audio compared to unlabeled multimodal speech, being able to utilize unlabeled audio can greatly improve the quality of multimodal speech representations. Furthermore, there are few or even no labeled multimodal speech for many tasks, such as speech translation. Hence, it is essential to build a model capable of zero-shot input modality transfer to perform these tasks with modalities without labeled data.

References

T. Afouras, J. S. Chung, and A. Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition, 2018. arXiv:1809.00496.

T. Afouras, J. S. Chung, and A. Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *ICASSP*, 2020.

T. Afouras, A. Zisserman, et al. Sub-word level lip reading with visual attention. *arXiv preprint arXiv:2110.07603*, 2021.

A. Aghajanyan, B. Huang, C. Ross, V. Karpukhin, H. Xu, N. Goyal, D. Okhonko, M. Joshi, G. Ghosh, M. Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

A. Bapna, Y.-a. Chung, N. Wu, A. Gulati, Y. Jia, J. H. Clark, M. Johnson, J. Riesa, A. Conneau, and Y. Zhang. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*, 2021.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021.

H.-J. Chang, S.-w. Yang, and H.-y. Lee. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091. IEEE, 2022.

S. Chen et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*, 2021.

J. S. Chung and A. Zisserman. Lip reading in the wild. In *ACCV*, 2016.

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250, 2021.

A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

L. Diener and T. Schultz. Investigating objective intelligibility in real-time emg-to-speech conversion. In *INTERSPEECH*, 2018.

X. Feng, Y. Zhang, and J. Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1759–1763. IEEE, 2014.

R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022.

R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, T. N. P. University, H. Germany, A. J. Aachen, J. H. University, Baltimore., Usa, N. C. S. Laboratories, Kyoto, and Japan. Far-field automatic speech recognition. *Proceedings of the IEEE*, 109:124–148, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

F. Hernandez, V. Nguyen, S. Ghannay, N. A. Tomashenko, and Y. Estève. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. *ArXiv*, abs/1805.04699, 2018.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*, 2021a.

W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021b.

R. Hu and A. Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52(4):288–300, 2010.

A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021.

C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord. Learning robust and multilingual speech representations. *arXiv preprint arXiv:2001.11128*, 2020.

S. Kim, I. R. Lane, S. Kim, and I. Lane. End-to-end speech recognition with auditory attention for multi-microphone distance speech recognition. In *Interspeech*, pages 3867–3871, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

B. E. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1-3):117–132, 1998.

T. Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *ACL*, 2018.

G. Lample and A. Conneau. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

C. Lee, K. Cho, and W. Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*, 2019.

T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve. Rethinking evaluation in asr: Are our models robust enough? *arXiv preprint arXiv:2010.11745*, 2020.

Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

K. Livescu, B. Zhu, and J. R. Glass. On the phonetic information in ultrasonic microphone signals. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4621–4624, 2009.

P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP*, 2021.

P. Ma, S. Petridis, and M. Pantic. Visual speech recognition for multiple languages in the wild. *arXiv preprint arXiv:2202.13084*, 2022.

T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *Interspeech*, 2019.

A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018.

N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 12 2014. doi: 10.1109/TPAMI.2015.2461544.

T. Q. Nguyen and J. Salazar. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*, 2019.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *arXiv preprint arXiv:2007.03001*, 2020.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, and Y. Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1182, October 2021.

K. Richmond, P. Hoole, and S. King. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

L. Sari, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf. A multi-view approach to audio-visual speaker verification. 2021.

M. Schuster and K. Nakajima. Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012.

D. Serdyuk, O. Braga, and O. Siohan. Audio-visual speech recognition is worth $32 \times 32 \times 8$ voxels. *arXiv preprint arXiv:2109.09536*, 2021.

B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022a.

B. Shi, W.-N. Hsu, and A. Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022b.

B. Shi, A. Mohamed, and W.-N. Hsu. Learning lip-based audio-visual speaker embeddings with av-hubert, 2022c. arXiv:2205.07180.

A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. Flava: A foundational language and vision alignment model. *ArXiv*, abs/2112.04482, 2021.

D. Snyder, G. Chen, and D. Povey. Musan: A music, speech, and noise corpus. *ArXiv*, abs/1510.08484, 2015.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333. IEEE, 2018.

A. Üstün, A. Bérard, L. Besacier, and M. Gallé. Multilingual unsupervised neural machine translation with denoising adapters. *arXiv preprint arXiv:2110.10472*, 2021.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25:926–930, 2018.

B. Xu, C. Lu, Y. Guo, and J. Wang. Discriminative multi-modality speech recognition. In *CVPR*, 2020.

S. Yang et al. SUPERB: Speech processing Universal PERformance Benchmark. In *Interspeech*, 2021.

X. Yao, Y. Zheng, X. Yang, and Z. Yang. Nlp from scratch without large-scale pretraining: A simple and efficient framework. *ArXiv*, abs/2111.04130, 2021.

P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng. Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks. *IEEE Transactions on Image Processing*, 29:1061–1073, 2019.