TRANSPEECH: SPEECH-TO-SPEECH TRANSLATION WITH BILATERAL PERTURBATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct speech-to-speech translation (S2ST) with discrete units leverages recent progress in speech representation learning, where a sequence of discrete representations derived in a self-supervised manner, are predicted from the model and passed to a vocoder for speech synthesis, still facing the following challenges: 1) Acoustic multimodality: the discrete units derived from speech with same content could be indeterministic due to the acoustic property (e.g., rhythm, pitch, and energy), which causes deterioration of translation accuracy; 2) high latency: current S2ST systems utilize autoregressive models which predict each unit conditioned on the sequence previously generated, failing to take full advantage of parallelism. In this work, we propose TranSpeech, a speech-to-speech translation model with bilateral perturbation. To alleviate the acoustic multimodal problem, we propose bilateral perturbation (BiP), which consists of the style normalization and information enhancement stages, to learn only the linguistic information from speech samples and generate more deterministic representations. With reduced multimodality, we step forward and become the first to establish a non-autoregressive S2ST technique, which repeatedly masks and predicts unit choices and produces high-accuracy results in just a few cycles. Experimental results on three language pairs demonstrate that BiP yields an improvement of 2.9 BLEU on average compared with a baseline textless S2ST model. Moreover, our parallel decoding shows a significant reduction of inference latency, enabling speedup up to 21.4x than autoregressive technique. [1]

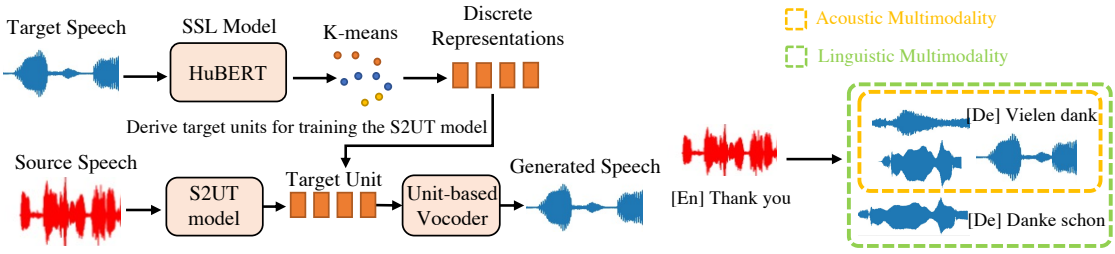
1 INTRODUCTION

Speech-to-speech translation (S2ST) aims at converting speech from one language into speech in another, significantly breaking down communication barriers between people not sharing a common language. Among the conventional method (Lavie et al., 1997; Nakamura et al., 2006; Wahlster, 2013), the cascaded system of automatic speech recognition (ASR), machine translation (MT), or speech-to-text translation (S2T) followed by text-to-speech synthesis (TTS) have demonstrated reasonable results yet suffering from expensive computational costs. Compared to these cascaded systems, recently proposed direct S2ST literature (Jia et al., 2019; Zhang et al., 2020; Jia et al., 2021; Lee et al., 2021a;b) demonstrate the benefits of lower latencies as fewer decoding stages are needed.

Among them, Lee et al. (2021a;b) leverage recent progress on self-supervised discrete units learned from unlabeled speech for building textless S2ST systems, further supporting translation between unwritten languages. As illustrated in Figure 1(a), the unit-based textless S2ST system consists of a speech-to-unit translation (S2UT) model followed by a unit-based vocoder that converts discrete units to speech, leading to a significant improvement over previous literature.

However, the current development of the unit-based textless S2ST system is hampered by the acoustic multimodal challenges (as illustrated in the orange dotted box in Figure 1(b)): different from the language tokens (e.g., bpe) used in the text translation, the self-supervised representation derived from speech with the same content could be different due to a variety of acoustic conditions (e.g., speaker identity, rhythm, pitch, and energy), including both linguistic content and acoustic information. As such, the indeterministic training target for speech-to-unit translation fails to yield good results.

[1] Audio samples are available at `https://TranSpeech.github.io/`.



(a) Direct speech-to-speech translation (S2ST) system (b) Multimodality challenges

Figure 1: In subfigure (b), we have the following observations: 1) In the orange dotted box, speech with the same content "Vielen dank" could be different due to a variety of acoustic conditions. Therefore, this acoustic multimodality poses a challenge for generating deterministic representations for accurate translation. 2) In the green dotted box, there are multiple correct target translations ("Danke schon" and "Vielen dank") for the same source word/phrase/sentence ("Thank you"). As such, this linguistic multimodality (Gu et al., 2017; Wang et al., 2019) prevents models from properly capturing the distribution of target translations.

In this work, we propose TranSpeech, a fast speech-to-speech translation model with bilateral perturbation. To tackle the acoustic multimodal challenge, we propose a **Bilateral Perturbation (BiP)** technique that finetunes a self-supervised speech representation learning model with CTC loss to generate deterministic representation agnostic to acoustic variation. Based on preliminary speech analysis by decomposing a signal into linguistic and acoustic information, the bilateral perturbation consists of the 1) **style normalization** stage which eliminates the acoustic-style information in speech and creates the style-agnostic "pseudo text" for finetuning; and 2) **information enhancement** stage which applies information bottleneck to create speech samples variant in acoustic conditions (i.e., rhythm, pitch, and energy) while preserving linguistic information. The proposed bilateral perturbation guarantees the speech encoder to learn only the linguistic information from acoustic-variant speech samples, significantly reducing the acoustic multimodality in unit-based S2ST.

The proposed bilateral perturbation eases acoustic multimodality and makes it possible for non-autoregressive (NAR) generation. As such, we further step forward and become the first to establish a NAR S2ST technique, which repeatedly masks and predicts unit choices and produces high-accuracy results in just a few cycles. Experimental results on three language pairs demonstrate that BiP yields an improvement of 2.9 BLEU on average compared with baseline textless S2ST models. The parallel decoding algorithm requires as few as 2 iterations to generate outperformed samples, enabling a speedup by up to 21.4x compared to the autoregressive baseline. TranSpeech further enjoys a speed-performance trade-off with advanced decoding choices, including multiple iterations, length beam, and noisy parallel decoding, trading by up to 3 BLEU points in translation results. The main contributions of this work include:

- Through preliminary speech analysis, we propose bilateral perturbation which assists in generating deterministic representations agnostic to acoustic variation. This novel technique alleviates the acoustic multimodal challenge and leads to significant improvement in S2ST.
- We step forward and become the first to establish a non-autoregressive S2ST technique with a mask-predict algorithm to speed up the inference procedure. To further reduce the linguistic multimodality in NAR translation, we apply the knowledge distillation technique and construct a less noisy and more deterministic corpus.
- Experimental results on three language pairs demonstrate that BiP yields the promotion of 2.9 BLEU on average compared with baseline textless S2ST models. In terms of inference speed, our parallel decoding enables speedup up to 21.4x compared to the autoregressive baseline.

2 Background: Direct Speech-to-Speech Translation

Direct speech-to-speech translation has made huge progress to date. Translatotron (Jia et al., 2019) is the first direct S2ST model and shows reasonable translation accuracy and speech naturalness. Translatotron 2 (Jia et al., 2021) utilizes the auxiliary target phoneme decoder to promote translation quality but still needs phoneme data during training. UWSpeech (Zhang et al., 2020) builds the

vector-quantized variational auto-encoder (VQ-VAE) model and discards transcript in the target language, while paired speech and phoneme corpora of written language are required.

Most recently, a direct S2ST system (Lee et al., 2021a) has been proposed to take advantage of self-supervised learning (SSL) and demonstrates its outperformed translation results without using the text data. However, the majority of SSL models are trained by reconstructing (Chorowski et al., 2019) or predicting unseen speech signals (Chung et al., 2019), which would inevitably include factors unrelated to the linguistic content (i.e., acoustic condition). As such, the indeterministic training target for speech-to-unit translation fails to yield good results.

The following textless S2ST system (Lee et al., 2021b) further demonstrates to obtain the speaker-invariant representation of speech by finetuning the SSL model to disentangle the speaker-dependent information. However, this system only constrains speaker identity, and the remaining aspects (i.e., content, rhythm, pitch, and energy) are still lumped together. In this work, we follow the common textless setup as illustrated in Figure 1(a), and propose a bilateral perturbation technique that generates more deterministic units agnostic to a broader variety of acoustic conditions, including the rhythm, pitch, and energy. With addressed acoustic multimodality, we design a much more challenging NAR S2ST technique, especially for applications where low latency is required, which is relatively overlooked. More related works have been attached in Appendix A in the supplementary materials.

3 Speech Analysis and Bilateral Perturbation

3.1 Acoustic Multimodality

The majority of SSL models are trained by reconstructing (Chorowski et al., 2019) or predicting unseen speech signals (Chung et al., 2019), which would inevitably include factors unrelated to the linguistic content (i.e., acoustic condition). As reported in previous textless S2ST system (Lee et al., 2021b), speech representations predicted by the self-supervised pre-trained model include both linguistic and acoustic information. As such, derived representations of speech samples with the same content can be different due to the acoustic variation (e.g., speaker identity, rhythm, pitch, and energy), and the indeterministic training target for speech-to-unit translation (as illustrated in Figure 1(a)) fails to yield good results. To address this multimodal issue, we conduct a preliminary speech analysis and introduce the bilateral perturbation technique.

3.2 Speech Analysis

In this part, we decompose speech variations into linguistic content and acoustic condition (e.g., speaker identity, rhythm, pitch, and energy) and provide a brief primer on each of these components.

Linguistic Content represents the meaning of speech signals. To translate a speech sample to another language, learning the linguistic information from the speech signal is crucial.

Speaker Identity is perceived as the voice characteristics of a speaker. It is reflected by the formant frequencies, which are the resonant frequency components in the vocal tract.

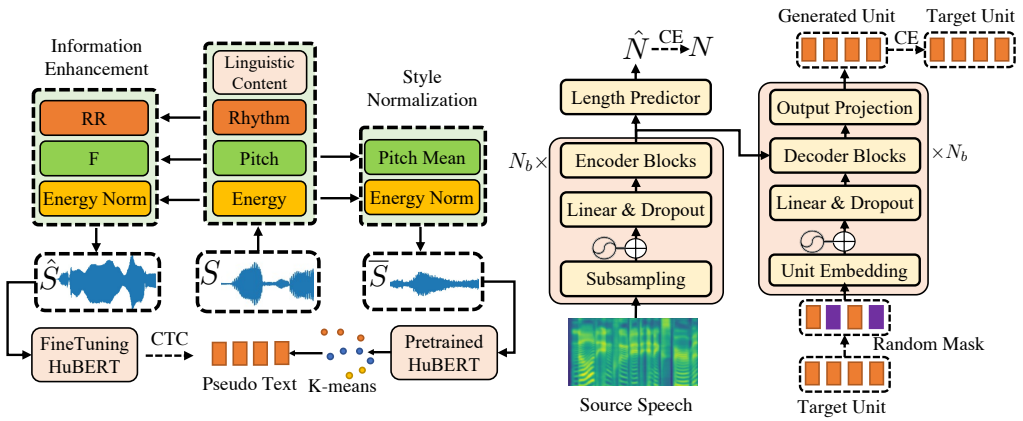
Rhythm characterizes how fast the speaker utters each syllable, and the duration variation plays a significant role in acoustic variation.

Pitch is an essential component of intonation. The pitch contour is generally considered as the result of a constant attempt to hit the pitch targets of each syllable, subject to physical constraints.

Energy indicates the frame-level magnitude of mel-spectrograms and directly affects the volume of speech, where stress and tone represent different energy values.

3.3 Bilateral Perturbation

To alleviate the multimodal problem and increase the translation accuracy in a textless S2ST system, we propose a bilateral perturbation technique that disentangles the acoustic variation and generates deterministic speech representations according to the linguistic content. Specifically, we leverage the success of connectionist temporal classification (CTC) finetuning (Baevski et al., 2019) with a pre-trained speech encoder, using the perturbed input speech and normalized target. Since the



(a) Speech Analysis and Bilateral Perturbation

(b) TranSpeech

Figure 2: In subfigure(a), we use $S, \overline{S}, \hat{S}$ to denote the original speech and perturbed speeches used in style normalization and information enhancement, respectively. RR: random resampling. F: a chain function for random pitch shifting of source speech. In subfigure(b), the "sinusoidal-like symbol" denotes the positional encoding, we have N_b encoder and decoder blocks. During training, we randomly select the masked position and compute the cross-entropy loss (denoted as "CE").

mechanism to obtain speaker-invariant representation has been well-studied (Lee et al., 2021b; Hsu et al., 2020), we focus on the more challenging acoustic conditions in a single-speaker scenario, including rhythm, pitch, and energy variations.

3.3.1 Overview

The overview of the information flow is shown in Figure 2(a), and we consider tackling the multi-modality in bilateral sides for CTC finetuning, including 1) **style normalization** stage to eliminate the acoustic information in the CTC target and create the acoustic-agnostic "pseudo text"; and 2) **information enhancement** stage which applies bottleneck on acoustic features to create speech samples variant in acoustic conditions (e.g., rhythm, pitch, and energy) while preserving linguistic content information. In the final, we train an ASR model using the perturbed speech as input and the "pseudo text" as the target.

As a result, according to speeches with acoustic variation, the ASR model with CTC decoding is encouraged to learn the "average" information referring to linguistic content and generate deterministic representations, significantly reducing multimodality and promoting speech-to-unit translation. In the following subsections, we present the bilateral perturbation technique in detail:

3.3.2 Style Normalization

To create the acoustic-agnostic "pseudo text" for CTC finetuning, the acoustic-style information should be eliminated and disentangled: 1) We first compute the averaged pitch fundamental frequency $\overline{p}$ and energy $\overline{e}$ values in original dataset S; and 2) for each sample in S, we conduct pitch shifting to $\overline{p}$ and normalize its energy to $\overline{e}$, resulting in a new dataset $\overline{S}$ with the averaged acoustic condition, where the style-specific information has been eliminated; finally, 3) the self-supervised learning (SSL) model encodes $\overline{S}$ and creates the normalized targets for CTC finetuning.

3.3.3 Information Enhancement

According to the speech samples with different acoustic conditions, the ASR model is supposed to learn the deterministic representation referring to linguistic content. As such, we apply the following functions as information bottleneck on acoustic features (e.g., rhythm, pitch, and energy) to create highly acoustic-variant speech samples $\hat{S}$, while the linguistic content remains unchanged, including 1) formant shifting fs, 2) pitch randomization pr, 3) random frequency shaping using a parametric equalizer peq, and 4) random resampling RR.

- For rhythm information, random resampling RR divides the input into segments of random lengths, and we randomly stretch or squeeze each segment along the time dimension.

- For pitch information, we apply the chain function $F = fs(pr(peq(x)))$ to randomly shift the pitch value of original speech.
- For energy information, we take an average from a log-mel spectrogram along the frequency axis.

The perturbed waveforms $\hat{S}$ are highly variant on acoustic features (i.e., rhythm, pitch, and energy) while preserving linguistic information. It guarantees the speech encoder to learn the "acoustic-averaged" information referring to linguistic content and generate deterministic representations. The hyperparameters of the perturbation functions have been included in Appendix C.

3.4 ANALYSIS

To visualize the acoustic multimodality and demonstrate the effectiveness of proposed bilateral perturbation, we apply the information bottleneck on acoustic features (i.e., rhythm, pitch, and energy) to create perturbed speech samples $\hat{S}_r, \hat{S}_p, \hat{S}_e$, respectively. As illustrated in Figure 5 in Appendix D, we plot the spectrogram and pitch contours of the original and acoustic-perturbed samples and present the derived representations. The **unit error rate (UER)** is further adopted as an evaluation matrix to measure the undeterminacy and multimodality according to acoustic variation, and we have the following observations: 1) In the pre-trained SSL model, the acoustic dynamics result in UERs by up to 22.7% (in rhythm), indicating the distinct alteration of derived representations. The pre-trained SSL model learns both linguistic and acoustic information given speech, and thus the units derived from speech with the same content can be indeterministic; however, 2) with the proposed bilateral perturbation (BiP), a distinct drop of UER (in energy) by up to 82.8% could be witnessed, demonstrating the efficiency of BiP in producing deterministic representations referring to linguistic content.

Table 1: We calculate UER between units derived from original and perturbed speeches respectively using the **pre-trained** and **fine-tuned** SSL model, which is calculated averaged over the dataset. It measures the ability of the SSL model to generate acoustic-agnostic representations referring to linguistic content.

Acoustic	Pretrained	BiP-Tuned
Reference	0.0	0.0
Rhythm $\hat{S}_r$	22.7	10.2
Pitch $\hat{S}_p$	16.3	4.3
Energy $\hat{S}_e$	10.5	1.8

4 TRANSPEECH

The S2ST pipeline has been illustrated in Figure 2(a), we 1) use the SSL HuBERT (Hsu et al., 2021) tuned by BiP to derive discrete units of target speech; 2) build the sequence-to-sequence model TranSpeech for speech-to-unit translation (S2UT) and 3) apply a separately trained unit-based vocoder to convert the translated units into waveform.

In this section, we first overview the encoder-decoder architecture for TranSpeech, following which we introduce the knowledge distillation procedure to alleviate the linguistic multimodal challenges. Finally, we present the mask-predict algorithm in both training and decoding procedures and include more advanced decoding choices.

4.1 ARCHITECTURE

The overall architecture has been illustrated in Figure 2(b), and we put more details on the encoder and decoder block in Appendix B.

Conformer Encoder. Different from previous textless S2ST literature (Lee et al., 2021b), we use conformer blocks (Gulati et al., 2020) in place of transformer blocks (Vaswani et al., 2017). The conformer model (Guo et al., 2021; Chen et al., 2021) has demonstrated its efficiency in combining convolution neural networks and transformers to model both local and global dependencies of audio in a parameter-efficient way, achieving state-of-the-art results on various downstream tasks. Furthermore, we employ the multi-head self-attention with a relative sinusoidal positional encoding scheme from Transformer-XL (Dai et al., 2019), which promotes the robustness of the self-attention module and generalizes better to different utterance lengths.

Non-autoregressive Unit Decoder. Currently, S2ST systems utilize the autoregressive S2UT models and suffer from high inference latency. Given the N' frames source speech $X = \{x_1, \ldots, x_{N'}\}$,

autoregressive model θ factors the distribution over possible outputs $Y = \{y_1, \ldots, y_N\}$ by $p(Y \mid X; \theta) = \prod_{i=1}^{N+1} p(y_i \mid y_{0:i-1}, x_{1:N'}; \theta)$, where the special tokens $y_0(\langle bos \rangle)$ and $y_{N+1}(\langle eos \rangle)$ are used to represent the beginning and end of all target units.

Unlike the relatively well-studied non-autoregressive (NAR) MT (Gu et al., 2017; Wang et al., 2019; Gu et al., 2019; Ghazvininejad et al., 2019), building NAR S2UT models that generate units in parallel could be much more challenging due to the joint linguistic and acoustic multimodality. Yet the proposed bilateral perturbation eases this acoustic multimodality and makes it possible for NAR modeling. As such, we further step forward and become the first to establish a NAR S2ST model θ.

It assumes that the target sequence length N can be modeled with a separate conditional distribution p_L, and the distribution becomes $p(Y \mid X; \theta) = p_L(T \mid x_{1:N'}; \theta) \cdot \prod_{i=1}^{N} p\left(y_i \mid x_{1:N'}; \theta\right)$. The target units are conditionally independent of each other, and the individual probabilities p is predicted for each token in Y. Since the length of target units N should be given in advance, TranSpeech predicts it by pooling the encoder outputs into a length predictor.

4.2 Linguistic Multimodality

As illustrated in Figure 1(b), there might be multiple valid translations for the same source utterance, and thus this linguistic multimodality degrades the ability of NAR models to properly capture the target distribution. To alleviate this linguistic multimodality in NAR translation, we apply knowledge distillation to construct a sampled translation corpus from an autoregressive teacher, which is less noisy and more deterministic than the original one. The knowledge of the AR model is distilled to the NAR model, assisting to capture the target distribution for better accuracy.

4.3 Mask-Predict

The NAR unit decoder applies the mask-predict algorithm (Ghazvininejad et al., 2019) to repeatedly reconsider unit choices and produce high-accuracy translation results in just a few cycles.

Training. During training, the target units are given conditioned on source speech sample X and the unmasked target units Y_{obs}. As illustrated in Figure 2(b), given the length N of the target sequence, we first sample the number of masked units from a uniform distribution $n \sim \mathrm{Unif}(\{1, \cdots, N\})$, and then randomly choose the masked position. For the learning objective, we compute the cross-entropy (CE) loss with label smoothing between the generated and target units in masked places, and the CE loss for target length prediction is further added.

Decoding. In inference, the algorithm runs for pre-determined T times of iterative refinement, and we perform a *mask* operation at each iteration, followed by *predict*.

In the first iteration $t = 0$, we predict the length N of target sequence and mask all units $Y = \{y_1, \ldots, y_N\}$. In the following iterations, we mask n units with the lowest probability scores p:

$$Y^t_{mask} = \arg\min_i(p_i, n) \qquad Y^t_{obs} = Y \backslash Y^t_{mask}, \tag{1}$$

where n is a function of the iteration t, and we use linear decay $n = N \cdot \frac{T-t}{T}$ in this work.

After masking, TranSpeech predicts the masked units Y^t_{mask} conditioned on the source speech X and unmasked units Y_{obs}. We select the prediction with the highest probability p for each $y_i \in Y^t_{mask}$ and update its probability score accordingly:

$$y^t_i = \arg\max_w P(y_i = w \mid X, Y^t_{obs}; \theta) \qquad p^t_i = \max_w P(y_i = w \mid X, Y^t_{obs}; \theta) \tag{2}$$

4.4 Advanced Decoding Choices

Target Length Beam. It has been reported (Ghazvininejad et al., 2019) that translating multiple candidate sequences of different lengths can improve performance. As such, we select the top K length candidates with the highest probabilities and decode the same example with varying lengths in parallel. In the following, we pick up the sequence with the highest average log probability as our result. It avoids distinctly increasing the decoding time since the computation can be batched.

Noisy Parallel Decoding. The absence of the AR decoding procedure makes it more difficult to capture the target distribution in S2ST. To obtain the more accurate optimum of the target distribution

and compute the best translation for each fertility sequence, we use the autoregressive teacher to identify the best overall translation.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Following the common practice in the direct S2ST pipeline, we apply the publicly-available pre-trained multilingual HuBERT (mHuBERT) model and unit-based HiFi-GAN vocoder (Polyak et al., 2021; Kong et al., 2020) in this work and leave them unchanged. In this section, we present the experimental setup of TranSpeech with proposed bilateral perturbation.

Dataset. For a fair comparison, we use the benchmark CVSS-C dataset (Jia et al., 2022), which is derived from the CoVoST 2 (Wang et al., 2020b) speech-to-text translation corpus by synthesizing the translation text into speech using a single-speaker TTS system. To evaluate the performance of the proposed model, we conduct experiments on three language pairs, including French-English (Fr-En), English-Spanish (En-Es), and English-French (En-Fr).

Model Configurations and Training. For bilateral perturbation, we finetune the publicly-available mHuBERT model for each language separately with CTC loss until 25k updates using the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$). Following the practice in textless S2ST (Lee et al., 2021b), we use the k-means algorithm to cluster the representation given by the well-tuned mHuBERT into a vocabulary of 1000 units. TranSpeech computes 80-dimensional mel-filterbank features at every 10-ms for the source speech as input, and we set N_b to 6 in encoding and decoding blocks. In training the TranSpeech, we remove the auxiliary tasks for simplification and follow the unwritten language scenario. TranSpeech is trained until convergence for 200k steps using 1 Tesla V100 GPU. A comprehensive table of hyperparameters is available in Appendix B.

Evaluation and Baseline models. For translation accuracy, we pre-train an ASR model to generate the corresponding text of the translated speech and then calculate the BLEU score (Papineni et al., 2002) between the generated and the reference text. In decoding speed, latency is computed as the time to decode the single n-frame speech sample averaged over the test set using 1 V100 GPU.

We compare TranSpeech with other systems using the publicly-available implementation in *fairseq* framework (Ott et al., 2019), including 1) Direct ASR, where we transcribe all the S2ST data with open-sourced ASR models as a reference and compute BELU; 2) Direct TTS, where we first synthesize speech samples with target units in unit-based vocoder, and then transcribe the speech to text and compute BELU; 3) S2T+TTS cascaded system, where we first train the S2T basic transformer model (Wang et al., 2020a), and then apply TTS model (Ren et al., 2020; Kong et al., 2020) text-to-speech generation; 4) basic transformer (Lee et al., 2021a) without using text, and 5) basic norm transformer (Lee et al., 2021b) with speaker normalization.

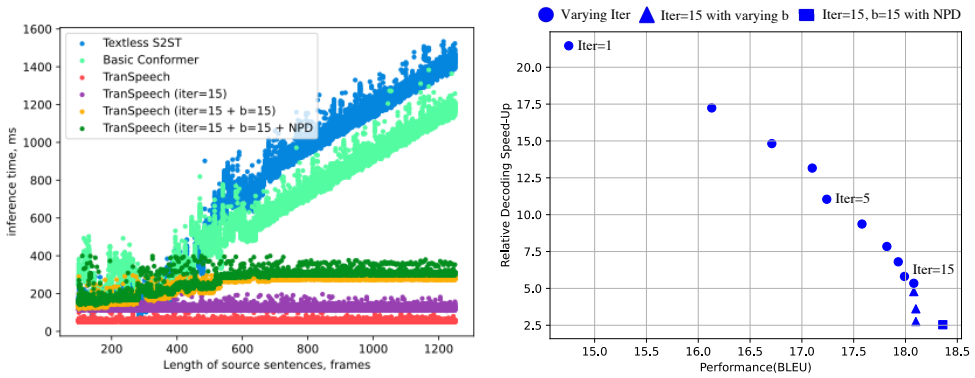
5.2 TRANSLATION ACCURACY AND SPEECH NATURALNESS

Table 2 summarizes the translation accuracy and inference latency among all systems, and we have the following observations: 1) **Bilateral perturbation (3 vs. 4)** improves S2ST performance by a large margin of 2.9 BLEU points. The proposed techniques address acoustic multimodality by disentangling the acoustic information and learning linguistic representation given speech samples, which produce more deterministic targets in speech-to-unit translation. 2) **Conformer architecture (2 vs. 3)** shows a 2.2 BLEU gain of translation accuracy. It combines convolution neural networks and transformers as joint architecture, exhibiting outperformed ability in learning local and global dependencies of an audio. 3) **Knowledge distillation (6 vs. 7)** is demonstrated to alleviate the linguistic multimodality between source and target language, where training on the distillation corpus provides a distinct promotion of around 1 BLEU points.

When considering the **speed-performance trade-off in the NAR unit decoder**, we find that more iterative cycles (7 vs. 8), or advanced decoding methods (e.g., length beam (8 vs. 9) and noisy parallel decoding (9 vs. 10)) further lead to an improvement of translation accuracy, trading up to 1.5 BLEU points during decoding. In comparison with baseline systems, TranSpeech yields the highest BLEU scores than the best publicly-available direct S2ST baselines (2 vs. 6) by a considerable margin; in

Table 2: **Translation quality (BLEU scores (↑)) and inference speed (frame/second (↑)) comparison with baseline systems.** We set beam size to 5 in autoregressive decoding, and apply 5 iterative cycles in NAR naive decoding. †: In this work, we remove the auxiliary task (e.g., source and target CTC, auto-encoding) in training the S2ST system for simplification. Though the S2ST system can be further improved with the auxiliary task, this is beyond our focus. BiP: Bilateral Perturbation; NPD: noisy parallel decoding; b: length beam in NAR decoding.

ID	Model	BiP	Fr-En	En-Fr	En-Es	Speed	Speedup
Autoregressive models							
1	Basic Transformer (Lee et al., 2021a)†	✗	15.44	15.28	10.07	870	1.00×
2	Basic Norm Transformer (Lee et al., 2021b)†	✗	15.81	15.93	12.98		
3	Basic Conformer	✗	18.02	17.07	13.75	895	1.02×
4	Basic Conformer	✔	**22.39**	**19.65**	**14.94**		
Non-autoregressive models with naive decoding							
5	TranSpeech - Distill	✗	14.86	14.12	10.27	**9610**	**11.04**×
6	Transpeech - Distill	✔	16.23	15.9	10.94		
7	TranSpeech	✔	**17.24**	**16.3**	**11.79**		
Non-autoregressive models with advanced decoding							
8	TranSpeech (iter=15)	✔	18.03	16.97	12.62	4651	5.34×
9	TranSpeech (iter=15 + b=15)	✔	18.10	17.05	12.70	2394	2.75×
10	TranSpeech (iter=15 + b=15 + NPD)	✔	**18.39**	**17.50**	**12.77**	2208	2.53×
Cascaded systems							
11	S2T + TTS	/	27.17	34.85	32.86	/	/
12	Direct ASR	/	71.61	50.92	68.75	/	/
13	Direct TTS	/	82.41	76.87	83.69	/	/



(a) Translation latency (b) Performance-speed trade-off.

Figure 3: The translation latency is computed as the time to decode the n-frame speech sample, averaged over the whole test set using 1 NVIDIA Tesla V100. b: length beam. NPD: noisy parallel decoding.

fact, only 2 mask-predict iterations (see Figure 3(b)) are necessary for achieving a new state-of-the-art on textless S2ST.

5.3 Decoding Speed

We visualize the relationship between the translation latency and the length of input speech in Figure 3(a). As can be seen, the autoregressive baselines have a latency linear in the decoding length. At the same time, NAR TranSpeech is nearly constant for typical lengths, even with multiple cycles of mask-predict iterative refinement. We further illustrate the versatile speed-performance trade-off for NAR decoding in Figure 3(b). TranSpeech enables a speedup up to 21.4x compared to the autoregressive baseline. On the other, it could alternatively retain the highest quality with BELU 18.39 while gaining a 253% speedup.

Table 3: **Two examples comparing translations produced by TranSpeech and baseline models.** We use the bond fonts to indicate the the issue of **noisy and incomplete translation.**

Source:	l'origine de la rue est liée à la construction de la place rihour.
Target:	the origin of the street is linked to the construction of rihour square.
Basic Conformer:	the origin of the street is linked to the construction of **the.**
TranSpeech:	th origin of the **seti** is linked to the construction of the **rear**.
TranSpeech+BiP:	the origin of the street is linked to the construction of the **ark.**
TranSpeech+BiP+Advanced:	the origin of the street is linked to the construction of the work.
Source:	il participe aux activités du patronage laïque et des pionniers de saint-ouen.
Target:	he participates in the secular patronage and pioneer activities of saint ouen.
Basic Conformer:	he participated in the activities of the late patronage a **d** see.
TranSpeech:	he **takes in the patronage activities in of saint.**
TranSpeech+BiP:	he participated in the activities of the lake patronage and **say pointing**
TranSpeech+BiP+Advanced:	he participated in the activities of the wake patronage and saint pioneers

5.4 Case Study

We present several translation examples sampled from the Fr-En language pair in Table 3, and have the following findings: 1) Models trained with original units suffer severely from the issue of *noisy and incomplete translation* due to the indeterministic training targets, while with the bilateral perturbation brought in, this multimodal issue is largely alleviated; 2) the advanced decoding methods lead to a distinct improvement in translation accuracy. As can be seen, the results produced by the TranSpeech with advanced decoding (more iterations and NPD), while of a similar quality to those produced by the autoregressive basic conformer, are noticeably more literal.

5.5 Ablation Study

We conduct ablation studies to demonstrate the effectiveness of several detailed designs in this work, including the bilateral perturbation and the conformer architecture in TranSpeech. The results have been presented in Table 4, and we have the following observations: 1) Style normalization and information enhancement in bilateral perturbation both demonstrate a performance gain, and they work in a joint effort to learn deterministic representations, leading to improvements in translation accuracy. 2) Replacing the relative positional encoding in the self-attention layer by the vanilla one (Vaswani et al., 2017) witnesses a distinct degradation in translation accuracy, demonstrating the outperformed capability of modeling both local and global audio dependencies brought by architecture designs.

Table 4: **Ablation study results.** SN: style normalization; IE: information enhancement; PE: positional encoding.

ID	Model	PE	Fr-En	En-Fr	En-Es
1	Basic Conformer	Relative	18.02	17.07	13.75
2	Basic Conformer + IE	Relative	21.98	19.60	14.91
3	Basic Conformer + SN	Relative	21.54	18.53	13.97
4	Basic Conformer	Absolute	17.23	16.19	13.06

6 Conclusion

In this work, we propose TranSpeech, a speech-to-speech translation model with bilateral perturbation. To tackle the acoustic multimodal issue in S2ST, the bilateral perturbation, which included style normalization and information enhancement, had been proposed to learn only the linguistic information from acoustic-variant speech samples. It assisted in generating deterministic representation agnostic to acoustic conditions, significantly reducing the acoustic multimodality and making it possible for non-autoregressive (NAR) generation. As such, we further stepped forward and became the first to establish a NAR S2ST technique. TranSpeech took full advantage of parallelism and leveraged the mask-predict algorithm to generate results in a constant number of decoding iterations. To address linguistic multimodality, we applied knowledge distillation by constructing a less noisy and more deterministic sampled translation corpus. Experimental results demonstrated that BiP yields an improvement of 2.9 BLEU on average compared with a baseline textless S2ST model. Moreover, TranSpeech showed a significant improvement in inference latency, which required as few as 2 iterations to generate outperformed samples, enabling a sampling speed of up to 21.4x faster than the autoregressive baseline. We envisage that our work serve as a basis for future textless S2ST studies.

REFERENCES

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Nanxin Chen, Shinji Watanabe, Jesús Villalba, Piotr Żelasko, and Najim Dehak. Non-autoregressive transformer for speech recognition. *IEEE Signal Processing Letters*, 28:121–125, 2020.

Sanyuan Chen, Yu Wu, Zhuo Chen, Jian Wu, Jinyu Li, Takuya Yoshioka, Chengyi Wang, Shujie Liu, and Ming Zhou. Continuous speech separation with conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5749–5753. IEEE, 2021.

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

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

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

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

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

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5874–5878. IEEE, 2021.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. *arXiv preprint arXiv:2012.15454*, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: Robust direct speech-to-speech translation. *arXiv preprint arXiv:2107.08661*, 2021.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. Cvss corpus and massively multilingual speech-to-speech translation. *arXiv preprint arXiv:2201.03713*, 2022.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. Janus-iii: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 99–102. IEEE, 1997.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*, 2021a.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021b.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376, 2006.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adam Polyak and Lior Wolf. Attention-based wavenet autoencoder for universal voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6800–6804. IEEE, 2019.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*, 2021.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pp. 7836–7846. PMLR, 2020.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wolfgang Wahlster. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020a.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020b.

Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Non-autoregressive machine translation with auxiliary regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5377–5384, 2019.

Bang Yang, Fenglin Liu, and Yuexian Zou. Non-autoregressive video captioning with iterative refinement. 2019.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. Uwspeech: Speech to speech translation for unwritten languages. *arXiv preprint arXiv:2006.07926*, 59:132, 2020.

Appendices

TranSpeech: Speech-to-Speech Translation With Bilateral Perturbation

A Related Work

A.1 Self-Supervised Representation Learning

There has been an increasing interest in self-supervised learning in the machine learning and speech processing community. Wav2Vec 2.0 (Baevski et al., 2020) trains a convolutional neural network to distinguish true future samples from random distractor samples using a contrastive predictive coding (CPC) loss function. HuBERT (Hsu et al., 2021) is trained with a masked prediction with masked continuous audio signals. The majority of self-supervised representation learning models are trained by reconstructing (Chorowski et al., 2019) or predicting unseen speech signals (Chung et al., 2019), which would inevitably include factors unrelated to the linguistic content (i.e., acoustic condition).

A.2 Non-autoregressive Sequence Generation

An autoregressive model takes in a source sequence and then generates target sentences one by one with the causal structure during the inference process. It prevents parallelism during inference, and thus the computational power of GPU cannot be fully exploited. To reduce the inference latency, (Gu et al., 2017) introduces a non-autoregressive (NAR) transformer-based approach with explicit word fertility, and identifies the multimodality problem of linguistic information between the source and target language. (Ghazvininejad et al., 2019) introduced the masked language modeling objective from BERT (Devlin et al., 2018) to non-autoregressively predict and refine translations. Besides the study of neural machine translation, many works bring NAR model into other sequence-to-sequence tasks, such as video caption (Yang et al., 2019), speech recognition (Chen et al., 2020) and speech synthesis (Ren et al., 2019). In contrast, we focus on non-autoregressive generation in direct S2ST, which is relatively overlooked.

B Model Architectures

In this section, we list the model hyper-parameters of TranSpeech in Table 5.

Hyperparameter		TranSpeech
Conformer Encoder	Conv1d Layers	2
	Conv1d Kernel	(5, 5)
	Encoder Block	6
	Encoder Hidden	512
	Encoder Attention Heads	8
	Encoder Dropout	0.1
Length Predictor	Projection Dim	512
Unit Decoder	Unit Dictionary	1000
	Decoder Block	6
	Decoder Hidden	512
	Decoder Attention Headers	8
	Decoder Dropout	0.1
Total Number of Parameters		67 M

Table 5: Hyperparameters of TranSpeech.

C INFORMATION ENHANCEMENT

We apply the following functions (Qian et al., 2020; Choi et al., 2021) on acoustic features (e.g., rhythm, pitch, and energy) to create acoustic-perturbed speech samples $\hat{S}$, while the linguistic content remains unchanged, including 1) formant shifting fs, 2) pitch randomization pr, 3) random frequency shaping using a parametric equalizer peq, and 4) random resampling RR. As shown in Figure 4, we further illustrate the mel-spectrogram of the single-perturbed utterance in bilateral perturbation.

- For fs, a formant shifting ratio is sampled uniformly from $\mathrm{Unif}(1, 1.4)$. After sampling the ratio, we again randomly decided whether to take the reciprocal of the sampled ratio or not.
- In pr, a pitch shift ratio and pitch range ratio are sampled uniformly from $\mathrm{Unif}(1, 2)$ and $\mathrm{Unif}(1, 1.5)$, respectively. Again, we randomly decide whether to take the reciprocal of the sampled ratios or not. For more details for formant shifting and pitch randomization, please refer to Parselmouth `https://github.com/YannickJadoul/Parselmouth`.
- peq represents a serial composition of low-shelving, peaking, and high-shelving filters. We use one low-shelving HLS, one high-shelving HHS, and eight peaking filters HPeak.
- RR denotes a random resampling to modify the rhythm. The input signal is divided into segments, whose length is randomly uniformly drawn from 19 frames to 32 frames (Polyak & Wolf, 2019). Each segment is resampled using linear interpolation with a resampling factor randomly drawn from 0.5 to 1.5.

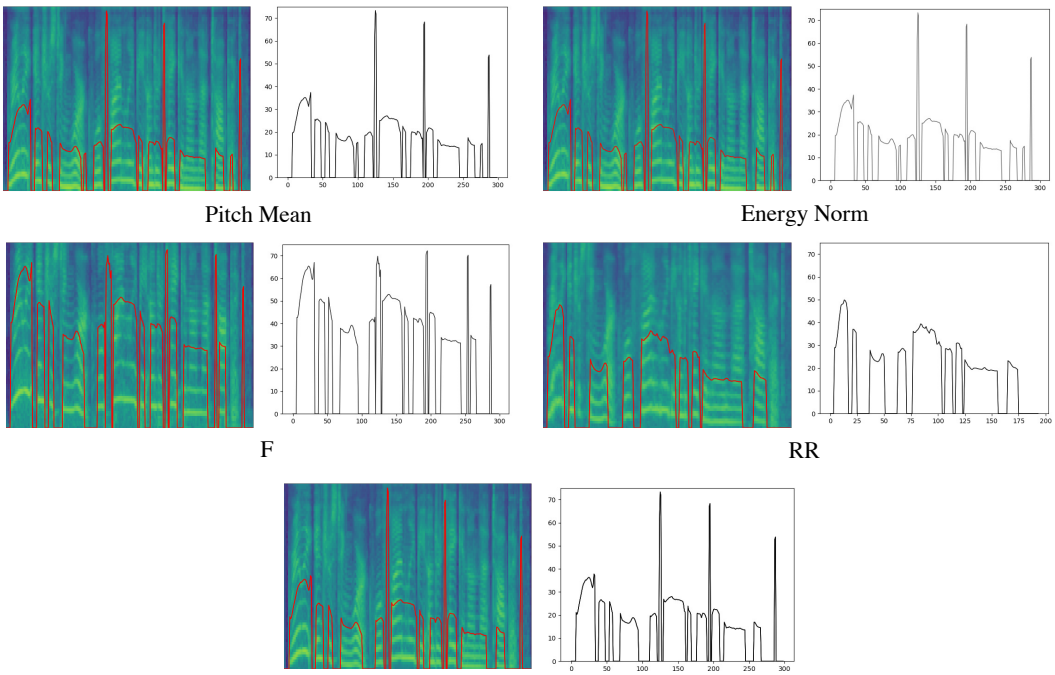


Figure 4: Spectrogram and pitch contours of the utterance with the single-perturbed acoustic condition, remaining the linguistic content ("really interesting work will finally be undertaken on that topic") unchanged. RR: random resampling. F: a chain function $F = fs(pr(peq(x)))$ for random pitch shifting.

D VISUALIZATION OF ACOUSTIC-PERTURBED SPEECH SAMPLES

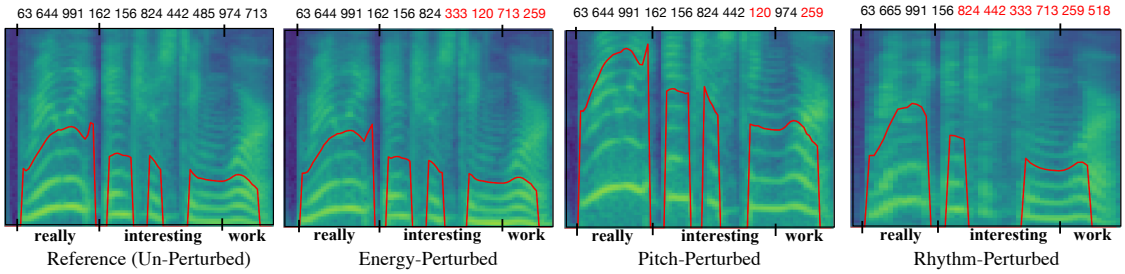


Figure 5: Spectrogram and pitch contours of speech sample with the perturbed acoustic condition, remaining the linguistic content ("really interesting work.") unchanged. The altered units are printed in red upside the spectrogram.