ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413930

TOWARDS DATA SELECTION ON TTS DATA FOR CHILDREN'S SPEECH RECOGNITION

Wei Wang, Zhikai Zhou, Yizhou Lu, Hongji Wang, Chenpeng Du, Yanmin Qian

MoE Key Lab of Artificial Intelligence, AI Institute
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{wangwei.sjtu, zhikai.zhou, luyizhou4, jijijiang77, duchenpeng, yanminqian}@sjtu.edu.cn

ABSTRACT

Although great progress has been made on automatic speech recognition (ASR) systems, children's speech recognition still remains a challenging task. General ASR systems for children's speech suffer from the lack of corpora and mismatch between children's and adults' speech. Efforts have been made to reduce such mismatch by applying normalization methods to generate modified adults' speech for ASR training. However, modified adults' data can reflect the characteristics of children's speech to a very limited extent. In this work, we adopt text-to-speech data augmentation to improve the performance of children's speech recognition system. We find that the children's TTS model generates speech with inconsistent quality due to children's substandard pronunciations of phonemes, and the ASR system suffers when trained with these additional synthesized data. To solve this problem, we propose data selection strategies on the TTS augmented data, and the effectiveness of the synthesized data can be substantially boosted for children's ASR modeling. We show that the speaker embedding similarity based data selection strategy can obtain the best position: relative 14.0% and 14.7% CER reduction for child conversation and child reading test set respectively compared to the baseline model trained on real data.

Index Terms— children's speech recognition, data augmentation, text-to-speech, data selection

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems has been improved significantly since the introduction of deep neural networks. Provided with large amounts of training data and advanced model structures, ASR models are now able to achieve human parity performance [1]. However, as far as we know, although many efforts have been made, children's speech recognition still remains a challenging task.

One challenge of children's speech recognition is the lack of data since children's corpora are difficult to collect. Moreover, children's physical and articulatory characteristics and expressions have inherently high variability [2]. To overcome these difficulties, vocal tract length normalization (VTLN) is proposed to reduce inter-speaker acoustic variability [3]. Pitch and formant modification [4, 5] were applied to reduce the acoustic mismatch between children's and adults' voice. However, the above approaches have not fundamentally solved the lack of data on children's speech.

In recent years, text-to-speech (TTS) based data augmentation for ASR has been widely applied and achieved good performance [6, 7]. However, the usage of synthesized speech generated by the TTS system trained on children's speech data is problematic since children's speech involves substandard or unclear pronunciation. As a result, the quality of synthesized speech is inconsistent under such circumstances. In this work, we present data selection for FastSpeech2 [8] synthesized children's speech. ASR model trained on TTS data is compared with VTLN normalization and pitch modification for children's speech recognition. We propose data selection approaches based on:

1. Character error rate (CER) of an ASR model trained on real data.
2. Resynthesis with CER filtered reference speech.
3. Normalized frame-wise acoustic posterior of a GMM-HMM model trained on real data.
4. Genuine score of a synthetic speech discrimination system trained on real and synthetic data.
5. Speaker embedding's cosine similarity of synthesized speech with its reference speech.

We perform our experiments on the SLT2021 CSRC data set and obtain the best result with speaker embedding similarity based selection: relative 14.7% and 14.0% CER reduction for child conversation and child reading test set respectively compared to the baseline model trained with real data.

The remainder of this paper is organized as follows. In Section 2, we introduce the ASR and TTS system adopted in our experiments. Then in Section 3, the proposed data selection methods are illustrated. The detailed experimental results and analysis are described in Section 4, and finally we conclude the paper in Section 5.

2. SYSTEM DESCRIPTION

2.1. Transformer-based E2E for ASR

Transformer is a sequence-to-sequence network constructed with an encoder and a decoder network. The encoder network is a stack of several transformer modules. Each transformer module consists of a multi-head self-attention and several fully connected feed-forward layers [9]. The encoder takes acoustic features as input and maps it into high-level representation. For ASR tasks, usually, a front-end CNN network is adopted to apply time-scale down-sampling [10].

The decoder network process the representation from the encoder with attention mechanism and outputs the predicted tokens in an auto-regressive fashion. For each decoding step, the decoder emits the posteriors of the next token given previous output tokens.

The transformer model is trained with the joint CTC-attention framework to improve robustness and achieve fast convergence [11,

Yanmin Qian is the corresponding author.

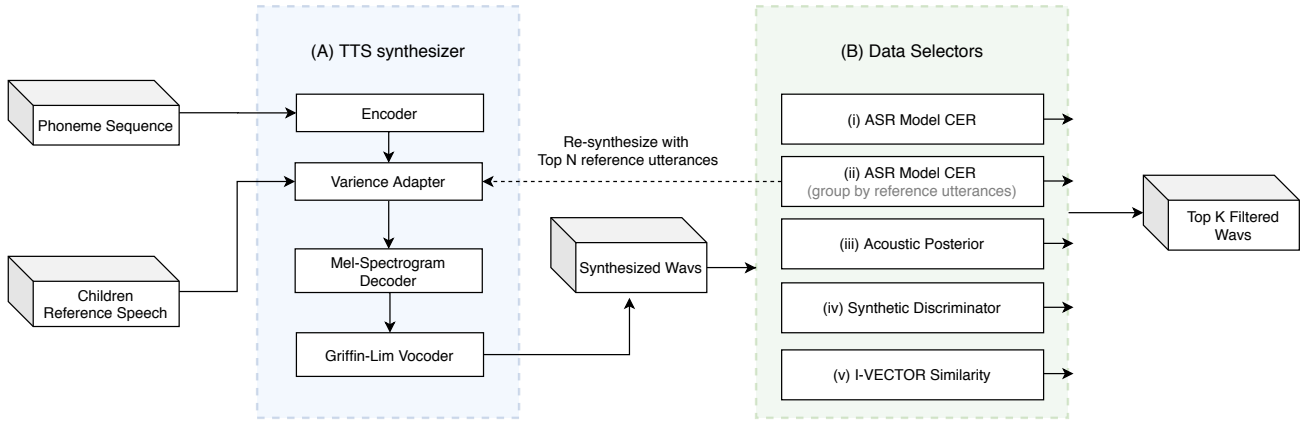

Fig. 1: **The proposed data selection procedure for augmented TTS data, utterances are filtered according to their scores**: (i): Synthesized utterances obtain CER scores from an ASR model trained on real data. (ii): CER scores of synthesized utterances are grouped by their reference utterances and each reference utterance obtains an averaged score of its synthesized utterances. Top N references are further fed into TTS synthesizer to generate K synthesized utterances. (iii): A GMM-HMM acoustic model is adopted to calculate the frame-wise posteriors for each utterance. The posteriors are normalized by the number of voiced frames to give a score to each utterance. (iv): A synthetic speech discrimination model is applied to give each synthesized utterance a genuine score. (v): A synthesized utterance is given a score by the cosine similarity of its i-vector with the i-vector of its reference as shown in Figure 2. For all the above methods, synthesized utterances with top K scores are used for training. (Here K=300hrs and N=1hr)

12]. Denote $\mathcal{L}_{ctc}$ and $\mathcal{L}_{s2s}$ as the CTC and S2S objective loss, the loss function of joint CTC-attention network is defined as:

$$\mathcal{L}_{jca} = \lambda \mathcal{L}_{ctc} + (1-\lambda)\mathcal{L}_{s2s} \quad (1)$$

A tunable coefficient $\lambda \in [0, 1]$ is applied to control the contribution of each loss. Joint CTC/attention decoding [13] is adopted to predict the output sequence, where S2S scores together with CTC prefix scores are combined to make the decision.

We combine Chinese characters and English BPE subwords for the modeling units [14] as final units. SpecAugment [15] is applied for all data throughout our experiments.

2.2. FastSpeech 2 for TTS

We follow the Transformer based TTS model in [8]. The feed-forward Transformer (FFT) block, a stack of self-attention and 1D-convolution, is adopted in Fastspeech 2. Some variance information is introduced to ease the one-to-many mapping problem. Besides speech mel-spectrogram, the model is trained and predicts the audio's duration, pitch, and energy.

In this work, some modifications are conducted for data augmentation. In order to generate speech from children, either in training or inference, we take the ground-truth of pitch and energy extracted from user-specified templates as input into the hidden sequence to predict the target speech. As is shown on the left in Figure 1, the FFT based encoder transforms the phoneme sequence into the hidden sequence. A variance adaptor then adds variance information (such as pitch) into the sequence. After that, the decoder predicts the mel-spectrograms. The output mel-spectrograms are reconstructed by Griffin-lim [16] for rapid turn around.

3. TTS DATA SELECTION

Children have substandard pronunciation for some phonemes, and a phoneme spoken by different children speakers might sound very different. Under such circumstances, it is hard to train a TTS model that generates speech with consistent quality, and the ASR model might suffer when trained with these unfiltered TTS data. Therefore, we propose data selection strategies to select high-quality speech that is beneficial for ASR model training.

3.1. Character Error Rate Selection

A straightforward idea is to select data based on character error rate (CER) measured by a baseline ASR model trained on real speech. Utterances with lower CER imply they are not severely distorted and are valid utterances for the ASR model. The CER criterion can help filter out synthesized speech with very low quality which are harmful to ASR training.

However, since these utterances are already well recognized by the baseline ASR model, the improvement of training on CER filtered utterances might be limited.

3.2. Normalized Frame-wise Acoustic Posterior

Traditional GMM-HMM acoustic models can model speech characteristics well. The posteriors of GMM-HMM model alignments directly represent how likely the synthesized speech matches its transcript. We calculate a score based on alignments of the GMM-HMM model: the frame-wise posteriors are normalized by the length of voiced speech of an utterance. Posteriors of silence ("sil") and non-speech noise ("spn") frames are ignored in the calculation.

$$score_{gmm} = \frac{\sum_{i=1}^{N} \log P(O_i|W_i)}{N-k}, W_i \notin \{sil, spn\} \quad (2)$$

where N is the total number of frames of an utterance, k is the number of silence, and non-speech noise, $P(O_i|W_i)$ is the probability of observation (acoustic feature) conditioned on phoneme sequence.

3.3. Reference Speech Selection and Re-synthesize

A FastSpeech2 model generates speech from a child speech template and a given phoneme sequence. Through CER selection, we find that synthesized speech with lower CER tends to be synthesized from the same references. That is, the trained TTS system is more

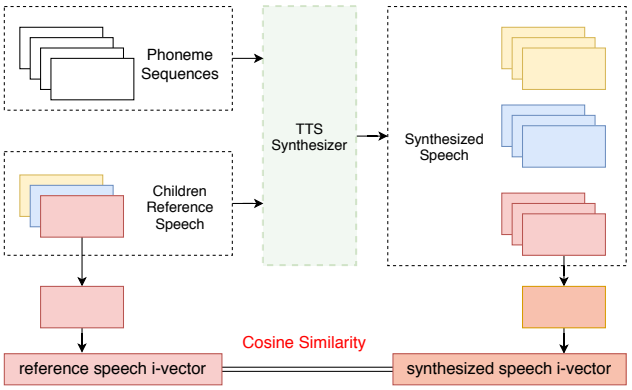

Fig. 2: **Speaker embedding similarity-based data selection pipeline**: Utterances in the right box with the same color are synthesized by the reference with corresponding color in the left box. The cosine similarity between the i-vector of a synthesized utterance and its reference is calculated as its score. The structure of the TTS synthesizer is illustrated in Figure 1.

adept at synthesizing valid speech from specific references. Thus we further use these selected references to synthesize speech with more transcripts and use them for training without further filtering.

By synthesizing with filtered references, low-quality or invalid synthesized speech that might be harmful to ASR training are expected to be mostly avoided. However, since all utterances are generated from the same group of selected references, the variety of synthesized speech might be limited.

3.4. Synthetic Speech Discrimination Genuine Score

The advancement of speech synthesis technologies means well-trained synthetic speech can be almost perceptually indistinguishable from real speech. A TTS speech discriminator (usually with a binary output) is trained to detect whether an utterance is recorded from a human (genuine) or is synthesized by computers (synthetic). A synthesized utterance having a higher genuine score from the discriminator means it more successfully deceives the discriminator. Utterances with higher genuine score have higher similarity with natural utterances from the perspective of the neural discriminator and are filtered from synthesized data for ASR training.

3.5. Speaker Embedding Similarity

The synthesized utterance by a FastSpeech2 model is expected to have characteristics of their reference, and such similarity can be measured by speaker embedding.

Here, we adopt i-vector[17] as the speaker embedding and measure the score of a synthesized speech by its i-vector's cosine similarity with its reference. A higher similarity can imply the synthesized utterance is of higher quality .

$$score_{ivec} = \frac{< i, i_{ref} >}{\|i\| \cdot \|i_{ref}\|} \quad (3)$$

where i means i-vector embedding of the utterance.

3.5.1. Illustration of i-vector based selection

As is shown in Fig. 3, all audio samples are transformed into embedding based on statistics pooling along time. The figure shows the effectiveness of our i-vector based data selection method: most of TTS utterances that are very different from children speech are discarded by our selector (green dots on the left). Although some synthesized utterances that are similar to children speech are also discarded, the selected high-quality utterances are still very helpful for training ASR model.

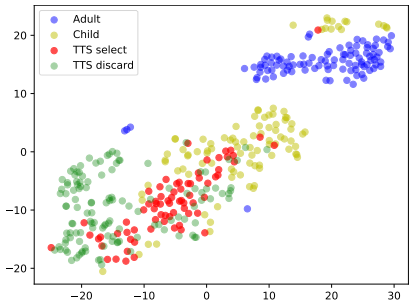

Fig. 3: **Distribution of audio samples with t-SNE.** 100 utterances are sampled from both adult speech and children speech, respectively. Then 200 utterances are sampled from synthesized data. Adult utterances are plotted in blue, yellow ones are real child speech. Red dots are selected data, and green ones are discarded ones.

4. EXPERIMENTS

4.1. Dataset

Our experiments are performed on the children's speech dataset from SLT2021 CSRC (children speech recognition challenge), which contains adult reading set, child reading set, and child conversation set. The language of all three sets is mandarin. And all speech data are in 16kHz, 16bit, and single-channel format.

Both our dev set and test set contain 1500 utterances from child reading and child conversation sets respectively. The training set is all data excluding the dev and test set, and is referred to as 'REAL' in our experiments.

Table 1: Details of CSRC Dataset

Part	Duration	# Utt.	# Spk.	Ages
Adult Reading	341.4h	243,056	1,999	18-60
Child Reading	28.6h	29,837	927	7-11
Child Conversation	29.5h	30,525	54	4-11

4.2. Experiment Setup

4.2.1. ASR Setup

The input of the model is an 80-dimensional log Mel-filterbank with 25ms window length computed every 10ms and pitch feature of 3 dimensions. The spec-augment [15] is conducted on speech features. We adopt 20 layers of encoder and 6 layers of the decoder with 2048 hidden units. Each layer is a Transformer block with 8 heads of 64 dimension self-attention layer. Dropout is set to 0.1 for each block and position-wise feedforward. For multitask learning(MTL), the weight for CTC and attention is set to 0.3 and 0.7. The modeling units are 3669 Chinese characters units and 100 English BPE units.

4.2.2. TTS Setup

The text-to-speech system is Transformer implemented on ESPNet [18]. The encoder contains 6 feed-forward Transformer blocks.

Each block has 2 heads with 384-dimensional attention hidden sizes and phoneme embedding. The decoder has 6 feed-forward Transformer blocks, which has the same hyper-parameters as the encoder. For TTS target, 320D Mel-filterbank with 16000 sampling frequency, 1024 FFT points, 800 points for window length, and 200 points shift is extracted. Three-dimensional pitch feature is computed with a window size of 50ms with 12.5ms shift, and 16000 sampling frequency by Kaldi [19].

4.2.3. Synthesized speech discriminator setup

We adopt the Light CNN architecture as the discriminator, which was the best system in the ASVspoof 2017 Challenge [20]. It also performed well in the ASVspoof 2019 Challenge in both replay and synthetic speech discrimination sub-tasks [21, 22]. The detailed model structure is the same as that of our previous work [23].

The front-end feature is the 257-dimensional log power spectrogram, which is extracted by computing 512-point Short-Time Fourier Transform (STFT) every 10 ms with a window size of 25 ms. We adopt the cross-entropy loss criterion as well as the SGD optimizer with a learning rate of 0.001 and a momentum of 0.9.

4.3. Evaluation Results

4.3.1. Comparison with existing approaches

VTLN and prosody modification normalization approaches are compared with ASR model trained on additional TTS unfiltered data (REAL 400hrs + TTS 300hrs). We follow the prosody modification method in [4]. The tool SoX implemented based on WSOLA [24] is adopted to modify the audio signal's tempo while keeping the original pitch and spectral unchanged. The factor λ is set to 1.1 to tune up the prosody of adults' utterances. For VTLN, the linear-VTLN model in Kaldi [19] is trained, starting from an existing system based on LDA+MLLT GMM-HMM. Then the VTLN warping factors are computed for each speaker. After that, the Mel-filterbank feature is re-generated with the VTLN warping factors for normalization.

The results show that additional TTS unfiltered data leads to slight degradation on child conversation set and inferior improvement on child reading set compared with other two approaches.

Table 2: Results (CER%) compared with existing approaches

Traing Data	Conversation	Reading
REAL (baseline)	27.16	8.05
VTLN	26.04	7.46
Prosody modi.	25.33	7.56
REAL + TTS unfiltered	27.34	7.86

4.3.2. Comparison among proposed data selectors

Table 3: Results (CER%) of data selection methods

Training Data (REAL +)	Conversation (dev / test)	Reading (dev / test)
TTS random	26.14 / 27.34	7.63 / 7.86
CER	23.41 / 24.48	7.03 / 7.32
GMM posterior	23.56 / 24.36	6.79 / 6.97
Reference resynthesis	24.92 / 25.71	6.54 / 6.99
Synthetic discrimination	23.21 / 24.42	7.03 / 7.33
I-VECTOR similarity	**22.62 / 23.35**	**6.26 / 6.87**
+ Prosody modi. & VTLN	**22.57 / 23.24**	**6.10 / 6.67**

For all experiments in Table 3, 1500 hours of synthesized data is first generated by the TTS model. Then each selection method is conducted to filter data. The comparison among data selection methods is performed on 20% (about 300hrs) filtered utterances, which performs the best on our dev set in Table 4. For unfiltered condition, data is also randomly selected 20% for a fair comparison. All proposed data selection methods achieve a lower CER than real data and unfiltered data.

Re-synthesizing with filtered references performs the worst on child conversation set among our proposed methods. This can be ascribed to two reasons: (i) The re-synthesized speech has not been filtered with any selectors, and contains invalid or severely distorted utterances. (ii) All re-synthesized speech corresponds to the same group of references (1hr), limiting the variety of synthesized data.

Synthetic speech discrimination based selection performs the worst on child reading set among our proposed methods. The filtered utterances might contains long silence frames that cannot provide enough information to be detected as synthetic by our discrimination model.

The i-vector similarity based selection performs the best on both test sets. The similarity between a synthesized utterance and its reference can effectively measure its quality. High similarity implies that the utterance reflects the speaker's characteristics of its reference well and can be considered as valid training data for ASR model.

4.3.3. Effect of data selection threshold

Table 4 shows how the amount of TTS selected data affects ASR performance. Training with scarce synthesized speech (5%) brings limited improvement, while an overly loose threshold (40%) might introduce distorted data that is harmful to ASR model training.

Table 4: **Effect of amount of data selected (CER%)**. Experiments in this table is performed on REAL data + i-vector similarity selected data (from 1500hrs TTS speech) with different thresholds.

Training Data (REAL +)		Conversation (dev / test)	Reading (dev / test)
5%	(75hrs)	23.51 / 24.30	6.96 / 7.27
10%	(150hrs)	23.20 / 24.26	6.72 / 7.07
20%	(300hrs)	**22.62 / 23.35**	**6.26 / 6.87**
40%	(600hrs)	23.39 / 24.28	6.87 / 7.24

5. CONCLUSIONS

This paper presented data selection of text-to-speech data augmentation for children speech recognition. Experiments show that appropriate data selection methods for augmented TTS data can significantly improve the performance of the ASR system. Data selection with speaker embedding (i-vector) similarity based selection method obtains the best position, with 14.0% and 14.7% relative improvement over the baseline on child conversation and child reading test set, respectively. Furthermore, by applying TTS data augmentation together with prosody modification and VTLN, we observe 14.4% and 17.1% relative improvement over the baseline.

6. ACKNOWLEDGEMENTS

This work was supported by the China NSFC projects (No. 62071288 and No. U1736202). Experiments have been carried out on the PI super- computer at Shanghai Jiao Tong University.

7. REFERENCES

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," 2017.

[2] Prashanth Gurunath Shivakumar and Panayiotis Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," 2018.

[3] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 135–140.

[4] Chenda Li and Yanmin Qian, "Prosody Usage Optimization for Children Speech Recognition with Zero Resource Children Speech," in *Proc. Interspeech 2019*, 2019, pp. 3446–3450.

[5] H. Kumar Kathania, S. Reddy Kadiri, P. Alku, and M. Kurimo, "Study of formant modification for children asr," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7429–7433.

[6] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," 2020.

[7] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[8] 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," 2020.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017.

[10] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. Interspeech 2019*, 2019, pp. 1408–1412.

[11] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[12] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[13] Takaaki Hori, Shinji Watanabe, and John Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017, pp. 518–529, Association for Computational Linguistics.

[14] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[15] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019.

[16] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[17] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[18] Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[20] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin, "Audio replay attack detection with deep learning frameworks.," in *Interspeech*, 2017, pp. 82–86.

[21] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," *arXiv preprint arXiv:1904.05576*, 2019.

[22] Yexin Yang, Hongji Wang, Heinrich Dinkel, Zhengyang Chen, Shuai Wang, Yanmin Qian, and Kai Yu, "The sjtu robust anti-spoofing system for the asvspoof 2019 challenge," *Proc. Interspeech 2019*, pp. 1038–1042, 2019.

[23] Hongji Wang, Heinrich Dinkel, Shuai Wang, Yanmin Qian, and Kai Yu, "Cross-domain replay spoofing attack detection using domain adversarial training," *Proc. Interspeech 2019*, pp. 2938–2942, 2019.

[24] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, vol. 2, pp. 554–557 vol.2.