CASIA Voice Conversion System for the Voice Conversion Challenge 2020

Zheng Lian[1,2]*, Jianhua Tao*[1,2,3]*, Zhengqi Wen*[1] *and Rongxiu Zhong*[1,2]

[1]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{zheng.lian}@nlpr.ia.ac.cn

Abstract

This paper presents our CASIA (Chinese Academy of Sciences, Institute of Automation) voice conversion system for the Voice Conversation Challenge 2020 (VCC 2020). The CASIA voice conversion system can be separated into two modules: the conversion model and the vocoder. We first extract linguistic features from the source speech. Then, the conversion model takes these linguistic features as the inputs, aiming to predict the acoustic features of the target speaker. Finally, the vocoder utilizes these predicted features to generate the speech waveform of the target speaker. In our system, we utilize the CBHG conversion model and the LPCNet vocoder for speech generation. To better control the prosody of the converted speech, we utilize acoustic features of the source speech as additional inputs, including the pitch, voiced/unvoiced flag and band aperiodicity. Since the training data is limited in VCC 2020, we build our system by combining the initialization using a multi-speaker data and the adaptation using limited data of the target speaker. The results of VCC 2020 rank our CASIA system in the second place with an overall mean opinion score of 3.99 for speaker quality and 84% accuracy for speaker similarity.

Index Terms: Voice conversion, Phonetic posteriorgrams (PPGs), LPCNet vocoder, Speaker adaptation

1. Introduction

Voice conversion (VC) aims to modify the source speaker's voice to sound like that of the target speaker while keeping the content consistent. VC frameworks have been utilized in widespread applications, including development of personalized speaking aids for speech-impaired subjects [1], novel vocal effects of singing voices [2], and a voice changer to generate various types of expressive speech [3].

Voice Conversation Challenge (VCC) plays an important role in the development of VC approaches. Organizers provide an open platform for participants to test their systems. The first VCC is organized in 2016. It requires participants to build VC models using parallel training samples, which contain pairs of the same transcription utterances spoken by different speakers. Following this, VCC 2018 increases difficulties. It requires participants to build their systems using non-parallel training data. This year is the third VCC [4]. It contains two tasks: VC within the same language (task 1) and cross-lingual VC (task 2). We participate in all tasks. Our participating system, the CASIA (Chinese Academy of Sciences, Institute of Automation) VC system, is elaborated in this paper.

The conventional VC approach usually needs parallel training data [5–7]. However, VCC 2020 [4] only provides few parallel training data in the task 1. What's worse, the parallel training data is unavailable in the task 2. To deal with this problem, researches propose some methods for non-parallel VC. Variational autoencoder (VAE) [8] has been successfully proposed for non-parallel VC [9, 10]. However, VAE suffers from the risk of over-smoothing. To deal with this problem, generative adversarial network (GAN) [11] and its variants (such as CycleGAN [12, 13] and StarGAN [14, 15]) use a discriminator to amplify this artifact in the loss function. However, the discriminator's discernment may not correspond well to human auditory perception, thus degrading the sound quality of converted speech. Recently, there is another track of research [16, 17] that applies phonetic posteriorgrams (PPGs) for non-parallel VC. PPGs are of frame-level linguistic information representations extracted from the speaker-independent automatic speech recognition system. The PPGs based VC framework can be separated into two modules: the conversion model and the vocoder. The conversion model takes PPGs as the inputs, aiming to predict the acoustic features of the target speaker. Then, the vocoder utilizes these predicted features to generate the speech waveform of the target speaker.

In VCC 2020, we utilize the PPGs based VC framework. Our system uses a CBHG conversion model [18] and a LPCNet vocoder [19] for speech generation. The CBHG [18] conversion model has a bank of 1-D convolutional filters, highway networks [20] and a bidirectional Gated Recurrent Unit (GRU) [21]. Previous works [18] have verified that this structure can effectively capture context information in feature sequences. The LPCNet vocoder [19] combines linear prediction with recurrent neural networks. Previous works [19] have verified that this vocoder can better control the spectral shape. To better control the prosody of converted speech, we utilize acoustic features of the source speech as additional inputs, including the pitch, voiced/unvoiced flag and band aperiodicity. To release the impact of the low-resource training samples, we apply the speaker adaptation strategy in the training process. We first develop the average voice model using the multi-speaker data. Then the average model is adapted to the target speaker via limited training samples. According to the official evaluation results [4, 22], our system reaches the 3.99 for speech quality and 84% accuracy for speaker similarity, ranking the second place.

2. Proposed Method

2.1. Framework Overview

Let us define the acoustic features $Y \in \mathbb{R}^{T\times D_a}$, the pitch $f0 \in \mathbb{R}^{T\times 1}$, the band aperiodicity (bap) $f_{bap} \in \mathbb{R}^{T\times 1}$, the voiced/unvoiced flag (vuv) $f_{vuv} \in \mathbb{R}^{T\times 1}$, the PPGs $L \in \mathbb{R}^{T\times D_p}$. Here, T is the number of frames.

In the training stage (Figure 1(a)), we first extract Y, $f0$, f_{bap}, f_{vuv} and L from the target speech. Then we concatenate $f0$, f_{bap}, f_{vuv} and L together, represented as $F = [L; f0; f_{vuv}; f_{bap}] \in \mathbb{R}^{T\times(D_p+3)}$. Finally, the CBHG conver-

10.21437/VCC_BC.2020-19

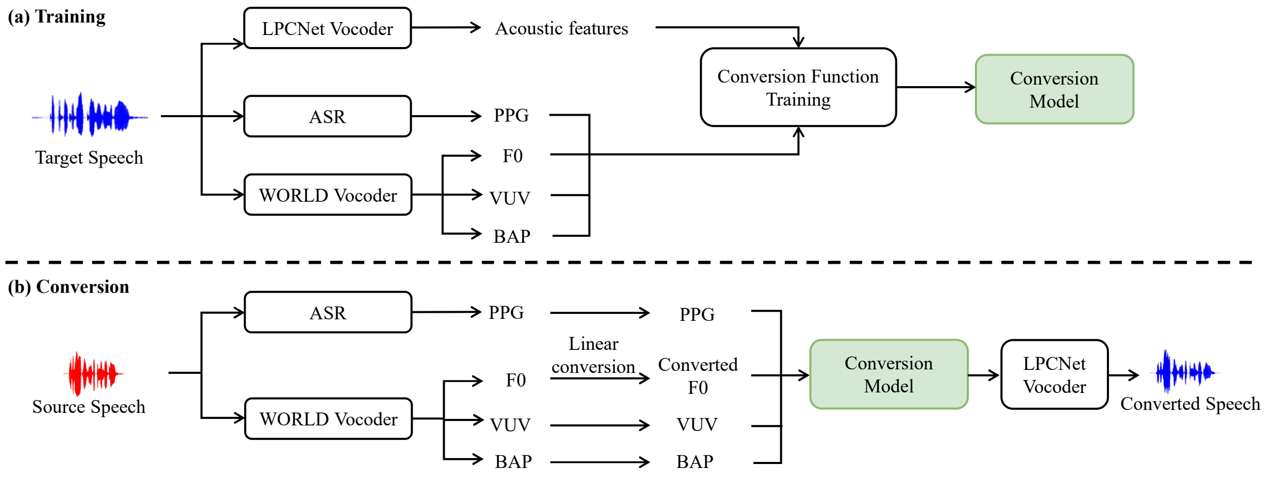


Figure 1: *The framework of our CASIA VC system.*

sion model is learned to convert F into the acoustic features Y.

In the conversion stage (Figure 1(b)), we first extract $f0$, f_{bap}, f_{vuv} and L from the source speech. A linear conversion is applied to convert $f0$ from the source speaker to the target speaker:

$$\log f0_y = (\log f0_x - \mu_x) * \left(\frac{\sigma_y}{\sigma_x}\right) + \mu_y \tag{1}$$

where μ_x (or μ_y) and σ_x (or σ_y) are the mean value and variance value of the source speaker's (or the target speaker's) $\log f0$, respectively. $\log f0_x$ and $\log f0_y$ are the source and converted $f0$ in logarithmic domain, respectively. Then, we concatenate L, converted $\log f0$, f_{vuv} and f_{bap} together. These representations are fed into the conversion model, aiming to predict the converted acoustic features. Finally, we feed these converted acoustic features to the speaker-dependent LPCNet vocoder for speech generation.

2.2. Linear Prediction Coding Net (LPCNet)

Vocoders influence the quality of converted speech. In this paper, we choose the LPCNet vocoder [19] for speech generation. LPCNet is a variant of WaveRNN [23], which generates speech samples from Bark-Frequency Cepstral Coefficients (BFCCs) [24], pitch period and pitch correlation parameters.

In this work, we use the code published by the Mozilla team [19] with some modifications. Since VCC 2020 focuses on the 24 kHz conversion strategies, we modify the original 16 kHz LPCNet to 24 kHz LPCNet. To better control high frequency features, we increase feature dimension of BFCCs features to 30 dimensions. To extract more accurate pitch trajectory, we use the reaper toolkit in the WORLD vocoder [25] for the pitch estimation. Totally, we extract 32-D features, including 30-D BFCCs, 1-D pitch period and 1-D pitch correlation.

2.3. Training with limited samples

Speaker-adaptive approaches [26–28] have been proposed to obtain models for the target speaker using limited training samples. In these methods, they first develop average voice models using multi-speaker data. Then, the average models are adapted to the target speaker with limited data.

Since the training data is limited in VCC 2020, we apply speaker-adaptive approaches for the CBHG conversion model and LPCNet vocoder. As for the CBHG conversion model, we first train a multi-speaker average model. The PPGs augmented with one-hot speaker embedding vectors are used as the inputs. Then, we adapt the pre-trained average model to the target speaker. As for the LPCNet vocoder, we first train an initialization model with a multi-speaker dataset without additional speaker embeddings. Then, we adapt the initialization model to the target speaker with limited data.

3. Experiments and Discussion

Firstly, we present our experimental databases. Then, we illustrate implementation details of our proposed method. Finally, we compare our method with other participants via the official subjective measures [4, 22]. Speech samples are available online at https://zeroqiaoba.github.io/VC-Demo..

3.1. Corpus Description

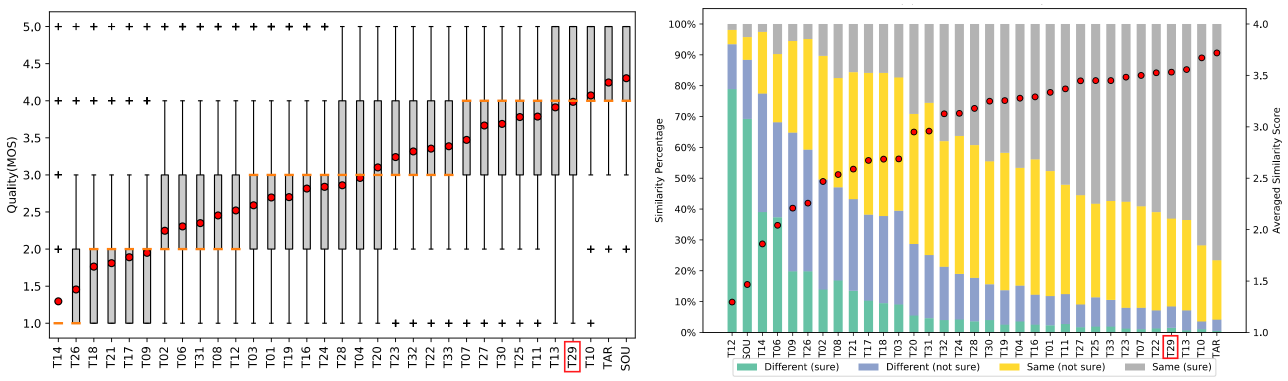
We evaluate our propose method on two tasks of VCC 2020. The first task requires to build an VC system within the same language. The second task needs to build a cross-lingual VC system. Concretely, the source speakers are in English. While the target speakers are in Finnish, German and Mandarin.

Due to the small amount of training data in VCC 2020, we increase training samples by means of speed perturbation. It is a technique of changing speech speed while keeping the tone unchanged. We randomly choose speed factor from 0.6, 0.8, 1.0 and 1.2.

For the multi-speaker average conversion modal and LPCNet vocoder, an internal Mandarin dataset of CASIA and the public available VCTK [29] English dataset are employed. Our Mandarin dataset contains 80 speakers and VCTK English dataset contains 108 speakers.

3.2. Implementation Details

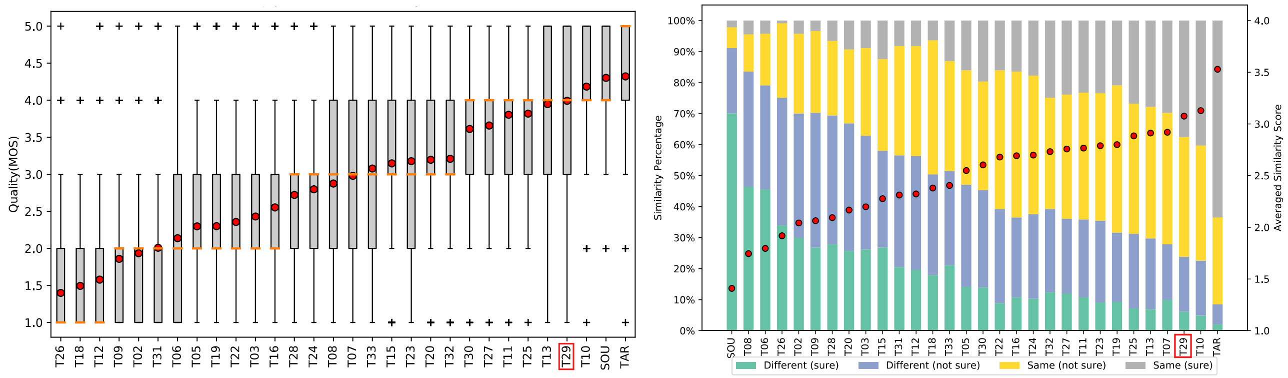
To extract PPGs from the input speech, we build a TDNN-LSTM based acoustic model using the Kaldi toolkit [30]. The input features are 40-D filter-bank features with the 25ms win-



(a) MOS results for task 1.

(b) Similarity results for task 1.

Figure 2: *Subjective test results for task 1 when averaging all speaker pairs. The CASIA VC system is T29.*



(a) MOS results for task 2.

(b) Similarity results for task 2.

Figure 3: *Subjective test results for task 2 when averaging all speaker pairs. The CASIA VC system is T29.*

dow size and 10ms window shift. Outputs of the last LSTM layer are utilized as the frame-level lexical features, PPGs. Totally, 512-D PPGs are extracted for each input waveform.

Acoustic features are extracted with 10ms window shift. The WORLD vocoder [25] is utilized to extract the pitch, vuv and bap. The LPCNet vocoder [19] is utilized to extract 32-D acoustic features.

The CBHG [18] conversion model has $K = 16$ sets of 1-D convolutional filters. The k-th set convolutional filter contains 128 filters with width k ($k \in [1, K]$). These convolutional layers are followed with one highway network (4 layers with 64 hidden units) and one bi-directional GRU layer (64 units for each GRU component).

In the training process, we choose the Adam optimizer with a learning rate of 0.001 for parameter optimization. We train our models for at least 100k steps with a batch size of 32. Gradient clipping is also used for regularization.

3.3. Evaluation Results of VCC 2020

In VCC 2020, the quality of the speech samples and their similarity to the target speaker are evaluated using the official subjective evaluation [4, 22]. The organizers recruit 206 Japanese listeners (110 female, 96 male) and 68 English listeners (32 female, 33 male and 3 for others) to evaluate the converted speech.

Subjective test results on two tasks are shown in Figure 2∼3, respectively. Besides the submitted systems, results of three baselines (*T11*, *T16*, *T22*), as well as the original source speaker (*SOU*) and target speaker (*TAR*) are also listed. Our CASIA VC system is denoted as *T29*. Figure 2(a) shows that our system achieves an MOS of 3.99 for speech quality in task 1, compared with 3.79 for the baseline *T11* and 4.07 for the top system *T10*. Figure 2(b) shows that our system achieves a similarity score of 84% in task 1, compared with 79% for the baseline *T11* and 89% for the top system *T10*. Figure 3(a) shows that our system achieves an MOS of 3.99 for speech quality in task 2, compared with 3.80 for the baseline *T11* and 4.18 for the top system *T10*. Figure 3(b) shows that our system achieves a similarity score of 69% in task 2, compared with 59% for the baseline *T11* and 71% for the top system *T10*. Overall, the results of VCC 2020 rank our system in the second place.

4. Conclusions

We present our CASIA VC system developed for the VCC 2020 in this paper. Our system adopts the CBHG structure to convert source speech's PPGs into target speaker's acoustic features. Then the LPCNet vocoder is utilized for speech generation. To better control the prosody of converted speech, the auxiliary acoustic features (including the pitch, voiced/unvoiced flag and band aperiodicity) of the source speech are utilized as additional inputs. To deal with the impact of limited training samples, speaker-adaptive strategies are also applied. The results of VCC 2020 rank our system in the second place with an MOS of 3.99 for speaker quality and 84% accuracy for speaker similarity.

5. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).

6. References

[1] A. B. Kain, J.-P. Hosom, X. Niu, J. P. Van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007.

[2] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 3277–3281.

[3] O. Turk and M. Schroder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 965–973, 2010.

[4] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020 — intra-lingual semi-parallel and cross-lingual voice conversion –," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020, pp. XXX–XXX.

[5] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[6] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 96, no. 10, pp. 1946–1953, 2013.

[7] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2016, pp. 1–6.

[10] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Voice conversion based on cross-domain features using variational auto encoders," in *International Symposium on Chinese Spoken Language Processing*, 2018, pp. 51–55.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[12] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *European Signal Processing Conference*, 2018, pp. 2100–2104.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6820–6824.

[14] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *IEEE Spoken Language Technology Workshop*, 2018, pp. 266–273.

[15] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," 2019, pp. 679–683.

[16] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[17] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6790–6794.

[18] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proceedings of the Interspeech*, pp. 4006–4010, 2017.

[19] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5891–5895.

[20] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[22] R. K. Das, T. Kinnunen, W.-C. Huang, Z. Ling, J. Yamagishi, Y. Zhao, X. Tian, and T. Toda, "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions –," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020, pp. XXX–XXX.

[23] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," 2018, pp. 2415–2424.

[24] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[25] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[26] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[27] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive hmm-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[28] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[29] V. Christophe, Y. Junichi, and M. Kirsten, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *The Centre for Speech Technology Research (CSTR)*, 2016.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.