Improving Model Stability and Training Efficiency in Fast, High Quality Expressive Voice Conversion System

Zhiyuan Zhao, Jingjun Liang, Zehong Zheng, Linhuang Yan
Zhiyong Yang, Wan Ding, Dongyan Huang
zhiyuan.zhao@ubtrobot.com
UBTECH Robotics Corp
Shenzhen, China

ABSTRACT

Voice conversion (VC) systems have made significant progress owing to advanced deep learning methods. Current research is not only concerned with high-quality and fast audio synthesis, but also richer expressiveness. The most popular VC system was constructed from the concatenation of an automatic speech recognition module with a text-to-speech module (ASR-TTS). Yet this system suffers from errors in recognition and pronunciation and it also requires a large amount of data for a pre-trained ASR mode l. We propose an approach to improve the model stability and training efficiency of a VC system. Firstly, a data redundancy reduction method is used to balance the distribution of vocabulary to avoid uncommon words being ignored during the training process; by adding connectionist temporal classification (CTC) loss, the word error rate (WER) of our system reduces to 3.02%, which is 5.63 percentage points lower than that of the ASR-TTS system (8.65%), and the inference speed (e.g., real-time rate 19.32) of our VC system is much higher than that of the baseline system (real-time rate 2.24). Finally, emotional embedding is added to the pre-trained VC system to generate expressive speech conversion. The results show that after fine-tuning on the multi-emotional dataset, the system can achieve high quality and expressive speech synthesis.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**.

KEYWORDS

voice conversion, redundancy reduction, disentangle correlation

ACM Reference Format:
Zhiyuan Zhao, Jingjun Liang, Zehong Zheng, Linhuang Yan and Zhiyong Yang, Wan Ding, Dongyan Huang. 2021. Improving Model Stability and Training Efficiency in Fast, High Quality Expressive Voice Conversion System. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3461615.3491106

1 INTRODUCTION

Voice conversion is the process of converting a source voice to a target voice while keeping the linguistic content unchanged. It aims at transforming the timbre of a source speaker to that of a target speaker through a conversion function. Traditional approaches, such as the Gaussian mixture model (GMM) [2], hidden Markov model (HMM) [5], the non-negative matrix factorisation [20] and the vocoding algorithms [7, 13, 14], produce robotic-like sounding results. Recently, deep neural network (DNN) [4], recurrent neural networks such as Long-Short Term Memory (LSTM) [16] and Transformer [11] have been used to address this problem. Neural network models are able to generate speech almost indistinguishable from human recordings. However, there is some deficiency in neural-network-based multi-speaker VC systems.

Firstly, the neural network method relies on big data and powerful computation [3], but the available data and computing resources are limited. Meanwhile, the multi-speaker model is more difficult to converge than the single-speaker model, and the mel-to-mel model is more difficult to train than the text-to-mel model. Therefore, it is necessary to collect a sufficient amount of data to improve the generalisation ability of the model, and to remove the redundancy in order to save training time. In a previous study the impact of different types and sizes of datasets on the model performance has been explored [10]. It was found that the model stability is not directly related to the scale of the dataset, but may be related to the characteristics of the data itself.

In this paper, we made the following assumptions. We propose one of the factors that determine the model quality and stability is the scale of vocabulary which the training dataset contains. Generally, within a certain range, the larger the training dataset, the more vocabulary it contains. But there is a limit to the number of words that can be used in daily conversation. Over that range, a larger dataset may contain redundant data, leaving a small proportion of rare words under-trained. In the experiment, We sub-sampled the common words in the dataset and raised the proportion of rare words to balance the word distribution, as well as to reduce the size of the dataset. We show that the stability of the model is related to the richness of the vocabulary of the training dataset. We hope this discovery could inspire later researchers in choosing recording texts to build more effective datasets.

Secondly, pronunciation errors are a common problem for VC systems. One view is that the model does not fully disentangle speech contents and the speaker identity in input utterances [21]. It requires disentangling representations of the speech content and speaker identity in input utterances, otherwise, the input speaker characteristics may affect the synthetic similarity or lead to mispronunciation. The current, widely used approach is to extract text or phonetic posteriorgram (PPG) using an Automatic Speech Recognition (ASR) system [17, 22]. However, training of a high-precision ASR system is very challenging.

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3491106

Considering the limitations of existing solutions, we propose a one-to-many training strategy to disentangle correlations between speaker identity and speech contents in input utterances without ASR. We employ a TTS model to construct parallel data pairs, applying synthesised speech signals as sources, and parallel personalised corpus as targets. It is straightforward to obtain a single-speaker dataset of dozens of hours to build a medium-quality TTS model. Through the TTS model, we can generate an infinite amount of parallel data to train a multi-speaker voice conversion system.

Thirdly, current research on speech synthesis is no longer limited to high-quality requirements, but further demands expressive speech to meet diverse needs in the fields of dubbing, audiobooks, emotional interaction, etc. Therefore, we add an emotional embedding into our VC system; based on the pre-trained model, only a few customised emotion utterances are needed to synthesise high quality audios.

The organisation of this paper is as follows: Section 2 presents the model architecture and explains the data redundancy reduction algorithm and one-to-many training strategy. Section 3 shows our experimental results to support our proposed methods. Finally, we conclude this paper in Section 4.

2 PROPOSED METHOD

In subsection2.1, we propose a seq2seq multi-speaker expressive voice conversion architecture shown in Figure 1. In subsection2.2, We describe a data redundancy reduction algorithm to extract a subset of the entire vocabulary and reduce redundancy to improve training efficiency. In subsection 2.3, we present a one-to-many training strategy, a hypothesis that can disentangle the correlation between speaker features and linguistic content by fixing the input speaker, whereby realise multi-speaker voice conversion.

2.1 Acoustic model

The VC encoder is adopted to extract speech contents from input utterances and eliminate speaker characteristics. The extracted mel-spectrogram is passed into two convolutional layers for downsampling. Then the features pass a stack of FFT blocks, which is similar to Fastspeech. The speaker encoder aims at extracting the characteristics of the target speaker, which was initially proposed in a speaker verification method [19]. We randomly choose five utterances from the target speaker and input their mel-spectrogram slices into the speaker encoder for training, and in the inference step, it accepts the utterance of a target speaker to extract the speaker feature. The emotion encoder simply receives an ID representing different emotions, and then concatenates those with the speaker representations of the speaker encoder.

The combination of features from the emotion and the speaker encoder is extended to the same length as the mel hidden representation in the time domain. Then these two vectors are concatenated as input of Length Regulator. We use a pre-trained parrotron-style acoustic model to generate aligned duration information as ground truth. Alignment features are extracted from each parallel data pair between synthesised utterances and real human utterances.

The decoder part consists of a stack of FFT blocks, a linear layer, and a postnet. We use Mean Square Error (MSE) loss to evaluate the gap between predicted and target mel spectrum, and the distance between ground truth and predicted duration. We added an

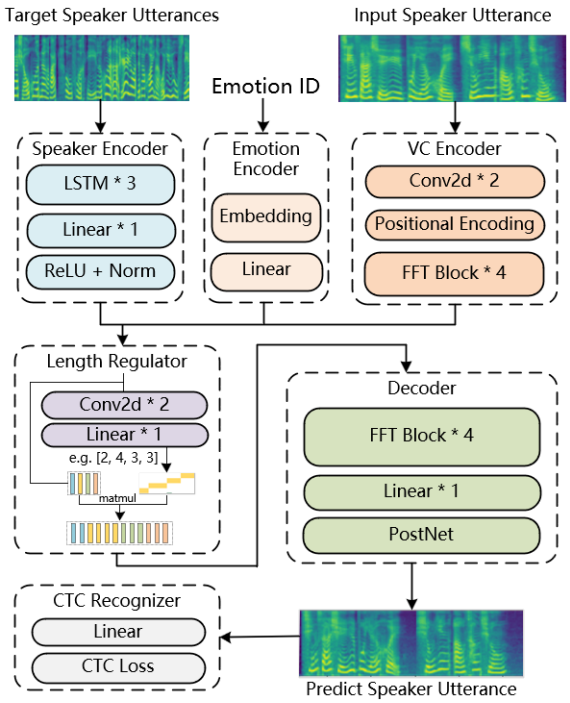

Figure 1: Overview of the model architecture. Target speaker utterances are mel spectrum slices random chosen from multiple target speaker sentences.

auxiliary CTC recogniser [8, 12] module to predict the phoneme of the output speech, conditioned on the latent representation of the VC encoder, calculating CTC loss to guide the system towards improved speech intonation.

2.2 Data redundancy reduction

We first normalise the text and use Jieba [6] for word segmentation. Then we count all words and the number of occurrences of each word, store the result in a dictionary, with the key being a word and the value a list of sentences containing this word. These words are sorted in ascending order by the number of occurrences. Starting with the word that occurs once, the first sentence containing that word is stored and the other words contained in that sentence are removed from the dictionary. We then repeat this step until the dictionary is empty. The stored sentences make up the reductive dataset, the remaining sentences cover all the words in the original dataset, reducing the repetition of the most frequent words.

2.3 One-to-many training strategy

As Figure 1 shows, assuming the source sequence as $x = (x_1, x_2, \ldots, x_n)$ where n represents the time frame of input, and the predicted sequence generated by decoder as $y = (y_1, y_2, \ldots, y_m)$ by the VC model, where m represents the time frame of the target. The predicted sequence should be as close to the target sequence $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ as possible. The latent variables in conversion, including linguistic contents $c = (c_1, c_2, \ldots, c_n)$ in each input frame, source speaker characteristics $s \in \{s_1, s_2, \ldots, s_i\}$, target speaker characteristics $\hat{s} \in \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_t\}$, where $i \in \{1, 2, \ldots, j\}$,

$t \in \{1, 2, \ldots, k\}$, i and t represent source and target, j and k are the numbers of source and target speakers.

The VC Encoder is aimed to strip s_i from x, and remain c only. The probability distribution of the input sequence can be expressed as

$$P(x|c, s_i) = \frac{P(c|x, s_i)P(x|s_i)}{P(c|s_i)} \tag{1}$$

To disentangle correlations between s_i and c, we construct a one-to-many VC system. Collecting a large scale parallel dataset of human recordings is difficult, we therefore introduce a TTS model to generate the source speech utterances with the same contents of human recordings. This approach manually fixes the source speaker and makes this variable become a constant s_0. Thus Eq. (1) becomes

$$P(x|c, s_0) = \frac{P(c|x, s_0)P(x|s_0)}{P(c|s_0)} = \frac{P(c|x)P(x)}{P(c)} \tag{2}$$

The probability distribution of the target sequences can be expressed as

$$P(y|c, \hat{s}_t) = \frac{P(c|y, \hat{s}_t)P(y|\hat{s}_t)}{P(c|\hat{s}_t)} \tag{3}$$

As $\hat{s}_t$ and c are extracted from different utterances, they can be considered independent. Combined with Eq. (2), we have

$$P(y|c, \hat{s}_t) = \frac{P(c|y)P(y|\hat{s}_t)}{P(c)} = \frac{P(c|y)P(y|\hat{s}_t)P(x|c)}{P(c|x)P(x)} \tag{4}$$

Therefore, the predicted sequence y is only related to source sequence x, input contents c, and target speaker characteristics $\hat{s}_t$. This implies that the source speaker characteristics s are separated from the speech content.

3 EXPERIMENTS

In this section, we introduce our experimental setup, followed by three experiments to demonstrate the effectiveness of our proposed methods.

3.1 Experimental setup

3.1.1 Datasets and features. The training datasets include an internal multi speaker dataset, ST-CMDS[15], and the combination of these datasets after data reduction. We introduce an emotion dataset, which contains three different types of emotions: sadness, happiness, and anger. The neutral statement, containing 20 hours mandarin recording of a female voice, is treated as the fourth emotion. Details are shown in table 1, where 90% of the data is used for training and 10% is used for testing. The number in column **Hours** and **Speakers** may be smaller than the meta information of the original dataset, because some audio was deleted due to poor quality, and the silence at the beginning and end of the audio is trimmed to facilitate training.

Speech signals from the dataset are firstly downsampled to 16 kHz and trimmed silence in the head and tail. Then we calculate the waveform using the Short-Time Fourier transform (STFT), and the Hann window function, with a window size of 1024 and hop length of 256. We map the power spectrum onto the 80 dim mel scale using filter-bank and limit the frequency between 40 Hz and 8 kHz, followed by a log dynamic range compression.

Table 1: Details of training dataset. Words are counted using the method in subsection 2.2.

Dataset	Speakers	Hours	Utterances	Words
Internal-Female	1	20	16000	37340
Internal-Expressive	3	6	6000	9884
Internal-Multi-Speaker	42	20	21000	21402
ST-CMDS	850	68	91200	60906
Combined (Internal+ST)	896	114	134200	91825
Reductive (Internal+ST)	896	40	47749	91825

3.1.2 Model details. Each convolutional layer in the VC Encoder contains 32 channels, with a kernel size of 5x5, and a stride of 2x2, which is used to downsample the sequence in time-frequency and mel channel (from 80 dim to 20 dim after convolutional layers). The FFT and linear layer width in the speaker encoder is 256, and the layer width in the decoder is 512. The postnet contains five 1-D convolution layers with 512 channels and a kernel size of five, followed by a 1D-batch normalisation after each layer and a dropout as final stage. The speaker encoder[1] is pre-trained on all the datasets listed in table 1, using generalised end-to-end (GE2E) loss [19] to ensure that the representation corresponds to each speakers. Its parameters are not updated in the later training of the VC encoder and decoder.

In order to eliminate source speaker characteristics and extract latent expressions of speech contents in the VC Encoder, we construct an TTS system based on MTTS[2] and WORLD [13] to synthesise large amounts of parallel data. The TTS structure consists of Jieba [6] for Chinese word segmentation, Conditional Random Fields(CRF) [18] in prosody predictor, a 3 x 256 bidirectional LSTM (Bi-LSTM) stack as duration predictor and a 3 x 512 Bi-LSTM network as the acoustic model. It is pre-trained on a 20-hour single mandarin female dataset.

3.2 Data redundancy reduction

Generally, a larger training dataset will enhance the performance of the model and expend more computing resources, yet the relationship between dataset size and model performance is not always positively correlated. In this experiment, we show that the stability of the model is not directly related to the size of the dataset, but directly proportional to the amount of vocabulary contained in the training dataset.

Table 2: The stability of models trained on different dataset (TT means training time).

Dataset	Hours (H)	TT (H)	WER(%)	CER(%)
Internal-Multi-Speaker	20	12	3.913	2.200
Combined(Int+ST)	114	57	3.440	1.945
Reductive(Int+ST)	40	21	3.020	1.675

We use datasets at three different scales for training, as shown in table 2. The VC encoder and decoder are trained from scratch to

[1]https://github.com/CorentinJ/Real-Time-Voice-Cloning
[2]https://github.com/Jackiexiao/MTTS

260 epochs, with a batch size of 64 and a learning rate of 1e-4. 500 texts outside the dataset were chosen randomly to generate audios for testing. For each emotion, we generate 125 audios and calculate Word Error Rate (WER) and Character Error Rate (CER) [1]. As tables 1 and 2 show, after the same epochs training on the dataset, the WER declines from 3.44% to 3.02%. The hours of training time is approximately one-third of the original, while the reductive dataset maintains the same amount of vocabulary as the combined dataset. We suggest that this is because, after reduction, the distribution of the vocabulary is more balanced, so that some infrequently used words have been fully trained.

3.3 The ablation study of CTC

In this experiment, we verify the effectiveness of the CTC Loss in our proposed model and compare the stability of our method with the ASR combined TTS method. The CTC module refers to the open-source code [3]. The training dataset is the combination of the internal dataset and ST-CMDS after data reduction, which achieved the lowest WER and CER in the previous experiment. The testing dataset is the same as above. We construct a baseline VC system adopting the Kaldi CVTE Mandarin Model V2 [4] combined with MTTS [2].

Table 3: Evaluate the WER and CER of a ASR-TTS system and verify the effectiveness of CTC module.

Model	WER(%)	CER(%)
kaldi-CVTE + MTTS	8.646	5.765
proposed Without CTC	6.539	3.961
proposed with CTC	3.020	1.675

As table 3 shows, the WER and CER in the model without CTC module is much higher than those in our proposed method. As the ASR system does not perform well in sentences that are meaningless and composed by randomly chosen words according to their POS tags. By adding CTC Recogniser as an auxiliary module, combining CTC loss and mel loss to calculate back propagation, the pronunciation accuracy of the synthesised audio can be enhanced.

3.4 Expressive voice conversion system

The acoustic model is trained 2000k iterations on all datasets except emotion dataset in table 1 on a single Nvidia V100 GPU, with a batch size of 64 and a learning rate of 1e-4. Both the proposed system and the baseline system are fine-tuned with 200k iterations on the emotion dataset. We use World [13] as the vocoder of the baseline and HiFi-GAN [9] as the vocoder of proposed method. After training on all datasets, we generate a Ground True Alignment (GTA) mel spectrogram of each emotion to fine-tune the vocoder.

The test sentences and audios were randomly selected. We conduct a Mean Option Score (MOS) test to evaluate the clarity, naturalness and similarity of the synthesised speech. For each emotion, the test set is composed of two utterances synthesised by the baseline model, two utterances generated by our proposed model, and one ground truth audio. Over 50 listeners are invited to rate the MOS and discriminate different emotions. All listeners are native speakers. We calculate the average score of MOS with 95% confidence intervals. The average training time (seconds per iteration) and the inference speed n indicates that waveforms are generated n times faster than real-time.

Table 4: MOS and speed results of ground truth, baseline and our proposed method. (TT means Training Time and IS means Inference Speed)

Method	TT	IS	Clarity	Naturalness	Similarity
Ground Truth	/	/	4.60 ± 0.10	4.51 ± 0.14	4.67 ± 0.12
CVTE+MTTS	$0.89s/it$	2.24	3.84 ± 0.08	3.39 ± 0.07	3.16 ± 0.09
Proposed	$0.44s/it$	19.32	4.30 ± 0.05	3.94 ± 0.06	4.20 ± 0.08

Table 5: MOS details of each emotion.

Emotion	Naturalness	Clarity	Similarity
Happiness	4.26 ± 0.10	3.97 ± 0.12	4.49 ± 0.10
Anger	4.38 ± 0.10	3.90 ± 0.13	4.02 ± 0.14
Neutral	4.19 ± 0.10	3.88 ± 0.12	4.27 ± 0.17
Sadness	4.37 ± 0.09	4.02 ± 0.12	3.98 ± 0.21

Table 4 shows that the MOS score of our proposed method achieves 4.19 in average, which is higher than the baseline system, the training time is half of the baseline, and the inference speed (real-time rate) is 8.6 times faster than the baseline. Some samples can be found here [5]. Table 5 lists the MOS details of each emotion. Each emotion has been scored more than 300 times, the average correct rate is 82.54%. According to the confusion matrix in figure 2, happiness is easier to be mistaken as neutral, compared with other miss classifications.

4 CONCLUSION

In this paper, we firstly propose a data redundancy reduction method to reduce redundant data in the training dataset, which balances the word distribution for improving model stability and improve training efficiency. After data reduction, the training time is approximately one-third of the original VC system, while the WER is reduced from 3.44% to 3.02%. Secondly, we construct a one-to-many VC system to separate the source speaker characteristics from the speech content. By adding CTC loss, it significantly reducing pronunciation errors. Finally, we add an emotional embedding to the pre-trained VC system to synthesise multi-emotional speech. Our results show that the converted speech achieves high quality in MOS (4.19 in average), while the training time is much half of the baseline and the inference speed is 8.6 times faster than the baseline. According to the confusion matrix, we find that the average recognition rate among four different emotion is 82.54% and happiness is more likely to be mistaken for neutral than other miss categorisation.

[3] https://github.com/bfs18/tacotron2
[4] https://kaldi-asr.org/models/m2
[5] https://approximetal.github.io/multi-speaker-VC/

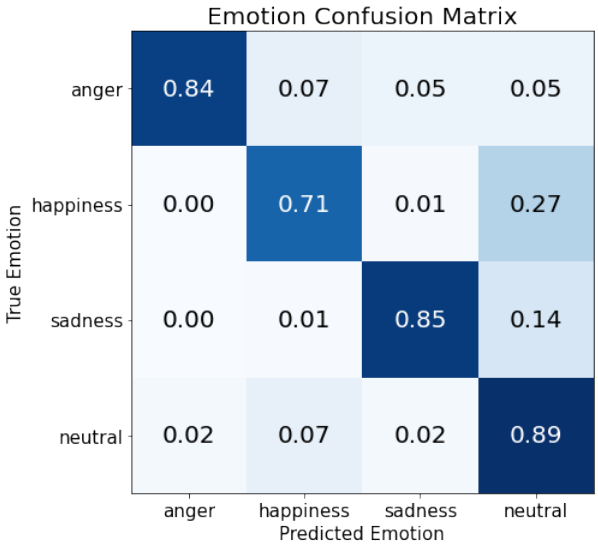

Figure 2: Emotion Confusion Matrix

REFERENCES

[1] [n. d.]. Computing Error Rates. ([n. d.]). https://sites.google.com/site/textdigitisation/qualitymeasures/computingerrorrates

[2] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2012. GMM-based emotional voice conversion using spectrum and prosody features. *American Journal of Signal Processing* 2, 5 (2012), 134–138.

[3] Fadi Biadsy, Ron J Weiss, Pedro J Moreno, Dimitri Kanvesky, and Ye Jia. 2019. Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation. (2019).

[4] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai. 2014. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 12 (2014), 1859–1872.

[5] Minghui Dong, Chenyu Yang, Yanfeng Lu, Jochen Walter Ehnes, Dongyan Huang, Huaiping Ming, Rong Tong, Siu Wa Lee, and Haizhou Li. 2015. Mapping frames with DNN-HMM recognizer for non-parallel voice conversion. In *2015 APSIPA*. IEEE, 488–494.

[6] et al. Junyi Sun. 2013. "Jieba": Chinese text segmentation. (2013). https://github.com/fxsjy/jieba

[7] Hideki Kawahara. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology* 27, 6 (2006), 349–353.

[8] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 ICASSP*. IEEE, 4835–4839.

[9] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *arXiv preprint arXiv:2010.05646* (2020).

[10] Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, and Klimkov Viacheslav. 2018. Effect of data reduction on sequence-to-sequence neural TTS.

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

[12] Peng Liu, Xixin Wu, Shiyin Kang, Guangzhi Li, Dan Su, and Dong Yu. 2019. Maximizing mutual information for tacotron. *arXiv preprint arXiv:1909.01145* (2019).

[13] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems* 99, 7 (2016), 1877–1884.

[14] Nathanaël Perraudin, Peter Balazs, and Peter L Søndergaard. 2013. A fast Griffin-Lim algorithm. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 1–4.

[15] ST-CMDS-20170001_1. 2017. Free ST Chinese Mandarin Corpus. (2017). https://openslr.org/38/

[16] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. 2015. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. (2015), 4869–4873.

[17] vcc2020__at__vc challenge.org. [n. d.]. Voice Conversion Challenge 2020. ([n. d.]). http://www.vc-challenge.org/

[18] Hanna M Wallach. 2004. Conditional random fields: An introduction. *Technical Reports (CIS)* (2004), 22.

[19] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 ICASSP*. IEEE, 4879–4883.

[20] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li. 2014. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 10 (2014), 1506–1521.

[21] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448* (2019).

[22] Xiao Zhou, Zhen-Hua Ling, and Simon King. 2020. The Blizzard Challenge 2020. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. 1–18. https://doi.org/10.21437/VCC_BC.2020-1