INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Incorporating Cross-speaker Style Transfer for Multi-language Text-to-Speech

Zengqiang Shang[1,2], *Zhihua Huang*[2,3], *Haozhe Zhang*[1,2], *Pengyuan Zhang*[1,2,*], *Yonghong Yan*[1,2]

[1]Key Laboratory of Speech Acoustics Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, China

shangzengqiang@hccl.ioa.ac.cn

Abstract

Recently multilingual TTS systems using only monolingual datasets have obtained significant improvement. However, the quality of cross-language speech synthesis is not comparable to the speaker's own language and often comes with a heavy foreign accent. This paper proposed a multi-speaker multi-style multi-language speech synthesis system (M3), which improves the speech quality by introducing a fine-grained style encoder and overcomes the non-authentic accent problem through cross-speaker style transfer. To avoid leaking timbre information into style encoder, we utilized a speaker conditional variational encoder and conducted adversarial speaker training using the gradient reversal layer. Then, we built a Mixture Density Network (MDN) for mapping text to extracted style vectors for each speaker. At the inference stage, cross-language style transfer could be achieved by assigning any speaker's style type in the target language. Our system uses existing speaker style and genuinely avoids foreign accents. In the MOS-speech-naturalness, the proposed method generally achieves 4.0 and significantly outperform the baseline system.

Index Terms: multilingual text-to-speech, domain-adversarial training, conditional variational encoder, fine-grained style encoder

1. Introduction

Recent speech synthesis systems have obtained good results and produce natural speech due to deep learning advances. However, the speech quality, target similarity, expressiveness, and robustness are still not satisfied for synthetic speech in multilingual TTS. The performance further deteriorates when the languages or styles differ significantly. To solve the problem, we first need to understand the composition of spoken speech, which contains content, speaker, and style factors. The content part is determined by linguistic information and contains one or more languages. Speaker information refers to static speaker-related information representing the person's timbre and is invariant to language and content. Other information is generalized as speech style, including semantics, emotions, and prosody, representing dynamic style information. Generally, the style information is weakly correlated to the speaker and the content.

To build a multilingual TTS system over monolingual datasets, the first obstacle we will face is that timbre and content are mixed together. As speaker identity is bound to the language, it is necessary to disentangle the timbre from language. Here, Zhang et al. [1] and Nekvinda et al. [2] achieve the language speaker separation using adversarial speaker training. Chen et al.[3] use ResCNN based speaker encoder that enable timbre extraction. The second problem is text conflicts for those languages which are similar in writing form. This could be effectively resolved by using phoneme input or introducing separated encoder for each language [4] [2]. The most challenging one is that cross-language synthesis often comes with a heavy foreign accent, which has rarely been studied. In the work of Zhang et al. [1], foreign accents can be relieved to a certain extent by carefully adjusting the degree of adversarial speaker training. But it is time-consuming to tune the super parameters and it can not guarantee to solve this problem completely. As far as my knowledge, there exist no effective approaches for non-authentic accents problem in multilingual TTS.

We attempt to incorporate cross-speaker style transfer, which transfers the source speaker's style to the target speaker, to overcome the non-authentic accent problem and improve the speech quality in multilingual TTS. There have been several approaches proposed for style transfer, and they can be classified into: 1) Coarse-Grained Style Transfer (CST) techniques [5] [6] and 2) Fine-Grained Style Transfer (FST) techniques [7] [8] [9] [10] [11]. While CST techniques focus on capturing sentence-level style features like emotion, which can be transferred across sentences of different text [12], FST techniques focus on capturing style features like rhythm, emphasis, melody, and loudness, which can not necessarily be transferred between sentences of different text [13]. Both CST and FST techniques get latent representations from either a Mel-spectrogram or hand-crafted features. In CST methods, the latent representation is in the form of a single time-independent vector, while in FST methods, time-dependent latent representations are obtained. The time-dependency of latent representations in FST can be either at the phoneme level or the Mel-spectrogram frame-level.

In this paper, we proposed a FastSpeech based multi-speaker, multi-style, multi-language TTS (M3), which utilizes a phone-level style transfer model capable of transferring style from any source speaker to any target speakers within training data. We can summarize our contributions as follows:

1) We overcome the foreign accent problem in cross-lingual synthesis by using existing authentic style at inference.
2) Our system supports multi-language speech synthesis for each available timbre and allows the user to customize each language's speaking style.
3) Our system could stabilize the training and significantly improve the speech quality by incorporating the fine-grained encoder.

The paper is organized as follows. Section 2 describes our system, which explains how it can decouple content, speaker, and style. Then the style predictor will be briefly introduced. Section 3 demonstrates and discusses the evaluation results. The conclusion is presented in Section 4.

* corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-1265

2. Method

To focus on the spectrogram generation part, we use universal vocoder HIFIGAN [14] to generate waveforms from spectrograms. Our M3 system consists of a FastSpeech based acoustic model and a predictor to fine-grained style code. We first describe the acoustic model's structure and explain how it works using only monolingual datasets. Then we describe the configuration of auxiliary style predictor.

2.1. Acoustic model structure

As shown in Fig.1, the acoustic model consists of three informational components, including fine-grained style encoder, multilingual text encoder, and speaker embedding, which separately contain dynamic style information speaker-independent linguistics information, and static timbre information. Moreover, a decoder component is added to fuse those information into Mel spectrograms.

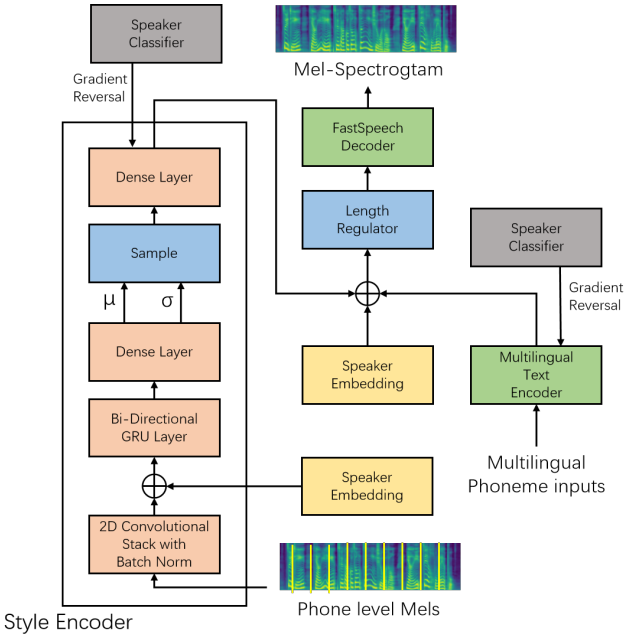

Figure 1: *The architecture of the acoustic model*

2.1.1. Multilingual text encoder

As FastSpeech2 [15], we obtain the time-point through force alignment and use phoneme as input. Experiments [1] have proved that using phoneme inputs can reduce the pronunciation errors and results in more fluent speech in multilingual TTS. Each language has its phoneme set in our model, and all languages share the same self-attention based text encoder. Besides, to enable cross-language voice cloning, we also conducted an adversarial speaker classifier with a gradient reversal layer on the encoding representation to remove residual speaker information.

2.1.2. Fine-grained style encoder

The fine-grained style encoder obtains temporal style representations from the aligned Mel-spectrograms at the phone-level, which maps each Mel-chunk into a fixed-length representation. We adopt a conditional variational encoding structure to extract speaker-independent style information, enabling cross-speaker style transfer without contaminating target timbre.

First, phoneme duration is obtained by performing text-speech force alignment[16]. Then Mel spectrograms are split into segments by the phoneme boundaries. As in Fig.1, we first apply three 2D convolution layers with batch norm for each segment. Different form [11], we retain the batch norm here because introducing instance norm will limit the component's learning ability. To reduce the presence of stationary features like speaker identity, we add speaker embedding as conditional input, which is concatenated with the learned feature map and sent to a bi-directional Gated Recurrent Unit (GRU) that maps variable-length representation into fixed-length vector z_s . Then time-independent latent style representation z is obtained by passing z_s to a variational encoder, which is introduced to mitigate the sparseness of encoded representations. Furthermore, we constrain the dimension of latent space to 3 to function as an information bottleneck. We assume a prior distribution, $p(z) = N(z; 0, I)$, and train the model to maximize the evidence lower bound (ELBO) defined in Equation 1. We use an anneal factor γ to avoid posterior collapse. To further squeeze out the timbre information, we also add an adversarial speaker classifier with a gradient reversal layer over the mapped latent representation.

$$\mathbb{L} = \mathbb{E}_{q_\theta(Z|[x_0,x_1,..,x'_{T}-1],S_{style})}[log(p_\theta(X|Y,Z,S)] - \gamma \sum_{i=0}^{T'-1} D_{KL}(q_\theta(z^i|x_i,S_{style})||p(z^i)) \quad (1)$$

2.1.3. Decoder

Follow in the implementation of FastSpeech, our decoder is consists of a stack of feedforward transformer blocks. The decoder predict the output Mel-spectrograms $X = [x^0, x^1, ..., x^{T-1}]$, given the phoneme encodings, Y, the latent representations, $Z = [z^0, z^1, ..., z^{T'-1}]$ and speaker embedding S. For all languages, we share the same decoder. We introduce the speaker embedding in a table look-up manner for both decoder and fine-grained style encoder. Since speaker embedding is learned through training, adversarial speaker classifier is necessary to help learning a good representations. Here T and T' represent the length of Mel frames and phoneme sequence, respectively. We optimize the acoustic model by minimizing the mean square error of Mel spectrograms.

2.2. Style predictor

Once finished the acoustic model training, we can extract the style code z for the training set. We assume that style coding has a strong correlation with the speaker's attributes and content. So we establish a mapping from phoneme sequences to style codes for each speaker in datasets. The speaker's attributes mentioned here may include long-term characteristics such as speaking habits and accent types. We use three 1D convolution layers with layer norm followed by an MDN[17] layer, which can increase the diversity of the generation. We optimize the style predictor by maximizing the likelihood estimation. For simplicity, we apply table-lookup speaker embedding to distinguish different speaker styles.

2.3. Style transfer across languages

After carefully designing each module, our model can decouple the style, content, and timbre in the multilingual TTS framework. It allows the arbitrary combination of those attributes to

enable flexible synthesis. So cross-speaker style transfer can be easily accomplished by combining timbre and style from different speakers. Possible combinations can be categorized into two classes, including within-language transfer and cross-language transfer as in Fig.2. Within-language transfer, one person can speak in another style, provides more choices for TTS in commercial use. For instance, we can design our own TTS voice by picking timbre and style from different people to meet our needs when we are not satisfied with the recorder's style. Cross-language style transfer enables cross-language speech synthesis. So multilingual speech synthesis problem is naturally resolved using style transfer in our system. Another advantage of our multilingual TTS is that it genuinely avoids non-authentic foreign language pronunciation because it uses existing style patterns.

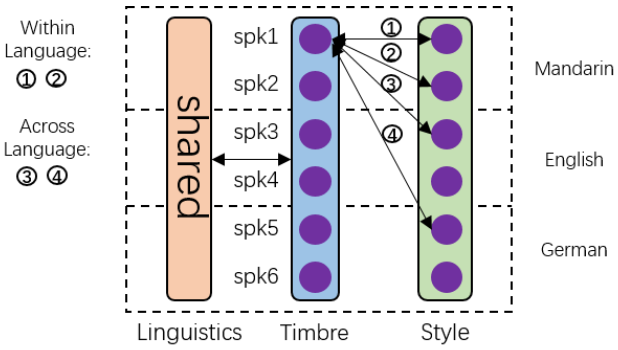

Figure 2: *Possible combination among linguistics, timbre and style.*

3. Experiments

3.1. Data

We train models using both openly available and proprietary datasets in three languages. Table 1 shows our data structure used for training.

Table 1: *Data source structure*

Languages	Mandarin	English	German
Source	Internal/ databaker[1]	ljspeech[18]/ vctk[19]	Common Voice[20]
Duration(h)	14.4/12	24/1.2	2.4
Num Speakers	2/1	1/1	1

3.2. Model and training setup

The synthesizer network uses FastSpeech architecture, with additional inputs consisting of learned speaker embedding (64-dim) and mapped style embedding (64-dim), concatenated and passed to the decoder at each step. The generated speech is represented as a sequence of 80-dim log-Mel spectrogram frames, computed from 50ms windows shifted by 12.5ms.

The speaker classifiers are fully-connected networks with two 256 unit hidden layers followed by a softmax layer. The acoustic model and speaker classifiers are trained with weight 1.0 and $1/log(speakernum)$, respectively. We scale the gradient from speaker classifier at gradient reversal layer with different factors in different models.

[1]https://www.data-baker.com/open_source.html

We use ADAM optimizer with $\beta_1 = 0.9$,$\beta_2 = 0.98$, and apply the same learning rate schedule as [21] with initial learning rate of 10^{-3} for acoustic model and style predictor.

3.3. Overall evaluation

We chose the baseline $(Baseline)$ model, which follows the multilingual TTS works of Zhang et al. [1], but were adapted to fit into FastSpeech architecture for a fairer comparison. Furthermore, We also created another baseline $(M3\ w/o\ FSE)$ by abandoning the fine-grained style encoder from our M3 model. We first evaluated the naturalness and speaker similarity among the three models. For each language, we chose one speaker for similarity and naturalness tests. We synthesized 100 samples per model and are rated by ten raters. For the M3 system, we synthesized foreign speech with randomly selected styles. Then we compared style similarity for both within-language style transfer and cross-language style transfer using subjective evaluation. Generally, it is challenging to transfer style across gender. So we selected speakers considering different factors, including gender, dialect. We select three speakers from mandarin, two from English and one from German. Among them, spk4 is spoken with an Indian accent. spk3 and spk5 are male speakers. Audio samples are available on GitHub[2].

3.3.1. Naturalness

We rely on crowd-sourced Mean Opinion Score (MOS) evaluations of speech naturalness via subjective listening tests. As table 3 suggests, introducing a phone-level fine-grained style module could stabilize the training and significantly improve the speech quality. In the model of baseline and $M3w/oFSE$, not all speaker s can synthesize high-quality speech. For example, the quality deteriorated severely for the speaker in LJspeech. Raters also comment that non-authentic accents are prone to be pronounced from baseline and $M3w/oFSE$ model. But M3 mostly gives authentic pronunciation. This heavily relies on the timbre-independent style encoder, which provides existing style patterns at cross-language synthesis. The experiment also shows larger speaker classifier's scale will cause a heavy foreign accent, but a smaller scale will lose the speaker similarity. Introducing the style encoder will ease the pain of finding the proper scale.

Table 2: *Average cosine similarity score for each speaker*

Source language	Identity	CN target	EN target	DE target
CN	spk1	0.8195	0.7681	0.7816
	spk2	0.8176	0.7532	0.7785
	spk3	0.8459	0.7741	0.8319
EN	spk4	0.7393	0.8305	0.8271
	spk5	0.7646	0.8434	0.8567
DE	spk6	**0.7372**	0.7830	0.9059

* Average cosine similarity between English and Mandarin recordings from spk3 are **0.7310**

3.3.2. Speaker Similarity

We perform both objective and subjective evaluations for speaker similarity. We first implement five scores MOS test. Table 3 shows that there is no significant difference in compar-

[2]https://shang0712.github.io/Multilingual/

Table 3: *Naturalness and speaker similarity MOS for multilingual model*

Source Language	Model	CN target		EN target		DE target	
		Naturalness	Similarity	Naturalness	Similarity	Naturalness	Similarity
CN	M3	4.22 ± 0.07	4.17 ± 0.07	4.01 ± 0.11	3.91 ± 0.10	4.15 ± 0.05	3.81 ± 0.09
	M3 w/o FSE	3.90 ± 0.05	4.13 ± 0.07	3.32 ± 0.06	3.84 ± 0.07	3.27 ± 0.07	3.63 ± 0.07
	Baseline	3.75 ± 0.06	4.14 ± 0.09	3.61 ± 0.05	3.96 ± 0.11	3.32 ± 0.11	3.79 ± 0.08
EN	M3	3.97 ± 0.09	3.57 ± 0.08	4.19 ± 0.08	4.00 ± 0.09	4.03 ± 0.08	3.83 ± 0.05
	M3 w/o FSE	3.15 ± 0.09	3.58 ± 0.09	3.44 ± 0.011	4.03 ± 0.05	3.27 ± 0.08	3.71 ± 0.07
	Baseline	3.26 ± 0.05	3.61 ± 0.011	3.46 ± 0.05	4.04 ± 0.08	3.41 ± 0.08	3.69 ± 0.05
DE	M3	3.98 ± 0.10	3.65 ± 0.11	4.05 ± 0.11	3.96 ± 0.07	4.17 ± 0.07	4.01 ± 0.011
	M3 w/o FSE	3.18 ± 0.11	3.61 ± 0.07	3.27 ± 0.05	3.79 ± 0.06	3.28 ± 0.06	3.98 ± 0.09
	Baseline	3.21 ± 0.07	3.56 ± 0.10	3.23 ± 0.05	3.86 ± 0.07	3.34 ± 0.05	4.03 ± 0.07

ison to baseline. That is because the adversarial speaker training module over the text encoder is the only one to determine the realization of cross-language voice cloning. And the degree of adversarial training is determined by the speaker classifier's scale. After carefully tuning this super parameter, each model performs well. But experiment shows a larger scale is needed for $M3w/oFSE$ to exhibit comparable performance.

Then we used a third-party per-trained speaker encoder[3] to evaluate the M3 model. We plot embedding projection of synthesized sentences. As shown in Fig.3, utterances from the same speakers form a tight cluster. Besides, we compute the within-language and cross-language cosine similarity for each speaker. And we use our internal bilingual data from the same person to calculate the cross-language cosine similarity as a benchmark. As shown in Table 2, the cosine similarity of within-language synthesis is slightly higher than the cross-language one. Both of them are greater than the benchmark.

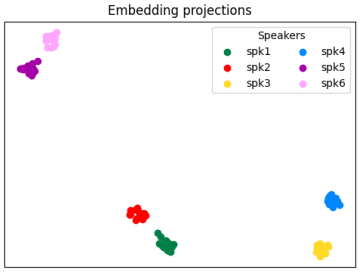

Figure 3: *T-SNE plot of speaker embedding*

3.3.3. Style transfer

We select spk1 as the source speaker and evaluate the style similarity of within-language style transfer and cross-language style transfer. When scoring style similarity, each test sample includes a reference sample and a generated sample. The listeners were asked to score each of the test cases on how closely the sample follows the reference's style. As shown in Table 4, M3 has an excellent performance on style transfer no matter in the cross-gender or cross-language situation. As shown in audio examples, the Indian accent style from spk5 can transfer to other speakers. Here speaker condition and adversarial speaker training are proposed to squeeze out residual speaker information. We perform an ablation study to evaluate which one is more important. The result shows that speaker similarity significantly drops in cross-gender transfer in the no speaker condition model. That means speaker condition contributes more to disentangle residual timbre information than adversarial speaker training. Experiments also show introducing those two methods will more or less cause the loss of style information.

[3]Resemblyzer: https://github.com/resemble-ai/Resemblyzer

Table 4: *Result of ablation study for the fine-grained encoder*

Target style		Score	Ours	-speaker adversarial	-speaker condition
CN	spk1	Style	4.20	4.24	4.26
		Timbre	0.8264	0.8148	0.8263
	spk2	Style	4.11	4.16	4.20
		Timbre	0.8250	0.8278	0.8212
	spk3	Style	3.95	3.91	4.12
		Timbre	0.8050	0.8064	**0.7147**
EN	spk4	Style	3.907	3.89	4.03
		Timbre	0.7965	0.7839	0.7398
	spk5	Style	3.85	3.86	4.08
		Timbre	0.7470	0.7355	**0.6666**
DE	spk6	Style	3.67	3.70	3.96
		Timbre	0.7816	0.7731	**0.6915**

3.3.4. Performance of style predictor

To examine that style sequences are predictable from phoneme sequence, we record the loss on train-set and valid-set. The mean absolute error can reduce to about 0.4 both on the training set and validation set. We also recorded the loss curve on the test set for spk2 when reducing training samples. Results show at least 200 samples are needed to train a decent style.

4. Conclusion

This paper proposed a multi-speaker multi-style multi-language speech synthesis system (M3), which improves the speech quality by introducing a fine-grained style encoder and overcomes the non-authentic accent problem through cross-speaker style transfer. Experiments results show our proposed genuinely avoids foreign accents and generally achieves 4.0 in the MOS-speech-naturalness, which significantly outperform the baseline system.

5. Acknowledgements

We thank Ruimin Wang for discussions and helpful feedback.

6. References

[1] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *Proc. Interspeech 2019*, pp. 2080–2084, 2019.

[2] T. Nekvinda and O. Dušek, "One model, many languages: Meta-learning for multilingual text-to-speech," *arXiv preprint arXiv:2008.00768*, 2020.

[3] Z. Liu and B. Mak, "Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers," *arXiv preprint arXiv:1911.11601*, 2019.

[4] Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6935–6939.

[5] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*, 2018.

[6] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[7] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[8] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6264–6268.

[9] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6699–6703.

[10] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, "Ppg-based singing voice conversion with adversarial representation learning," *arXiv preprint arXiv:2010.14804*, 2020.

[11] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *Proc. Interspeech*, 2020.

[12] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[13] Z. Zeng, J. Wang, N. Cheng, and J. Xiao, "Prosody learning mechanism for speech synthesis system without text length limit," *arXiv preprint arXiv:2008.05656*, 2020.

[14] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[15] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[17] C. BISHOP, "Mixture density networks," *NCRG/94/004*, 1994.

[18] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[19] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.