INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Spoken Content and Voice Factorization for Few-shot Speaker Adaptation

Tao Wang[1,2], *Jianhua Tao*[1,2,3], *Ruibo Fu*[1,2], *Jiangyan Yi*[1], *Zhengqi Wen*[1], *Rongxiu Zhong*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing

{tao.wang, jhtao, ruibo.fu, jiangyan.yi, zqwen, rongxiu.zhong}@nlpr.ia.ac.cn

Abstract

The low similarity and naturalness of synthesized speech remain a challenging problem for speaker adaptation with few resources. Since the acoustic model is too complex to interpret, overfitting will occur when training with few data. To prevent the model from overfitting, this paper proposes a novel speaker adaptation framework that decomposes the parameter space of the end-to-end acoustic model into two parts, with the one on predicting spoken content and the other on modeling speaker's voice. The spoken content is represented by phone posteriorgram (PPG) which is speaker independent. By adapting the two sub-modules separately, the overfitting can be alleviated effectively. Moreover, we propose two different adaptation strategies based on whether the data has text annotation. In this way, speaker adaptation can also be performed without text annotations. Experimental results confirm the adaptability of our proposed method of factorizating spoken content and voice. Listening tests demonstrate that our proposed method can achieve better performance with just 10 sentences than speaker adaptation conducted on Tacotron in terms of naturalness and speaker similarity.

Index Terms: speech synthesis, speaker adaptation, few-shot learning, spoken content and voice factorization

1. Introduction

With the development of deep learning, a significant amount of efforts has been made to improve the naturalness of text-to-speech (TTS), such as Tacotron [1–3], wavenet [4]. Apart from the naturalness, a TTS system is also expected to be able to generate an arbitrary speakers voice with few training adaptation data. To respond to this problem, many speaker adaptation solutions have been proposed and get effective results on several benchmark datasets [5–9]. However, when there are fewer data, such as only dozens of sentences, speaker adaptation is still a challenge.

The main challenge faced by few-shot speaker adaptation is that it is easy to overfit [10] when the adaptation database is small. Especially for end-to-end acoustic models, such as Tacotron [1, 2], since the model parameters are large, the loss function can be optimized to a very low level and the model will ignore the inherent laws of text pronunciation. To prevent the model from overfitting, there have been quite a few studies in machine learning [10]. The main idea is to use prior knowledge to constrain the complexity of the model so that it can adapt to fewer samples. When common machine learning models are used to deal with the few-shot learning, they usually choose a small hypothesis space $\mathcal{H}$. A small $\mathcal{H}$ has small complexity, thus requiring fewer samples [11]. Many researchers have tried to make the acoustic model more interpretable so as to adapt new speaker with small hypothesis space [5–7]. It has been experimentally proved that the shared hidden layers of acoustic model can benefit synthesized speech of each speaker from the knowledge of others [8, 12]. However, due to poor interpretability of the model, they often lack theoretical support. If the training of the model fails, it is difficult to find out what the problem is. Therefore, the factorization of the acoustic model is the key to make the model more interpretable. By adapting the parameters related to speaker characteristics, overfitting phenomenon can be alleviated effectively.

From a linguistic point of view, the spoken content and voice are two dominant factors in TTS synthesis. Factorizing the spoken content and voice may make the model more interpretable. Inspired by the voice conversion task [13], which is primarily achieved by modifying spectral features while retaining the spoken content in the given speech signal, phone posteriorgram (PPG) [13] is usually used as the representation of spoken content and it is speaker independent. Therefore, we can use PPG as an intermediate variable from text to speech, which will divide the acoustic model into two smaller subspaces and prevent the model from overfitting when adapting new speaker on small subspaces.

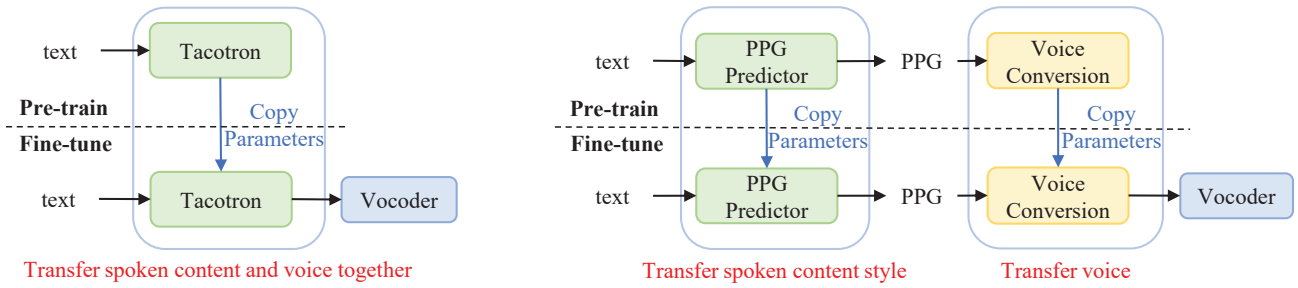
In this paper, we propose a spoken content and voice factorization framework for few-shot speaker adaptation task. The framework decomposes the end-to-end acoustic model into two independent sub-tasks, which corresponding to spoken content predicting and voice conversion respectively. The spoken content is represented by the PPG feature and it is used as prior knowledge to constrain the acoustic model space. By fine-tuning on the two smaller sub-spaces, fewer data can be accommodated and the synthesized speech is more natural and similar. Moreover, our framework can be adapted to data with or without text annotation, thus getting rid of the dependence of text annotation and providing a faster way for speaker adaptation. The main contributions can be concluded into two parts.

- A spoken content and voice factorization framework for few-shot speaker adaptation task is proposed to constrain the acoustic model space and prevent the model from overfitting. Experiments on VCTK datasets illustrate that this framework can deal with fewer data, and can get better natural and similar speech with only 10 sentences.
- We design two adaptive strategies to deal with whether the data for speaker adaptation has a transcript. Experiments show that even if the speaker adaptation is carried out without text annotation, a good effect can be obtained.

2. Background: speaker adaptation based on end-to-end TTS model

A basic acoustic model of end-to-end TTS can be a formula as transferring text information into acoustic features. Given a text sequences x and its target acoustic features y, the model

http://dx.doi.org/10.21437/Interspeech.2020-1745



(a) Speaker adaptation based on Tacotron

(b) Speaker adaptation based on spoken content and voice factorization

Figure 1: *(a) illustrates the speaker adaptation framework based on Tacotron. (b) illustrates the speaker adaptation framework based on spoken content and voice factorization.*

predicting acoustic features y given input x can be expressed as: $P(y|x;\theta)$. θ stands for the model parameters.

In general, speaker adaptation is to reuse the knowledge from other speakers when facing a new speaker. A typically approach is to fine-tune the pre-train multi-speaker acoustic model with the target speaker's corpus. The speaker adaptation method based on Tacotron [1] is shown in the left of Fig. 1. Due to the complex structure of the end-to-end model, which contains encoder, decoder and attention modules, there are some tricks to prevent overfitting. For example, when fine-tuning, the model only updates the parameters of the decoder module. After fine-tuning, we can learn the whole speaker's characteristics from a small amount of data.

3. Spoken content and voice factorization for speaker adaptation

When the amount of data is small, such as only dozens of sentences, it is easy to cause overfitting by adapting directly on the end-to-end acoustic model. In this section, we will introduce the spoken content and voice factorization framework for few-shot speaker adaptation.

3.1. Spoken content and voice factorization

There are mainly two aspects of speaker characteristics. One aspect is the style of spoken content. For example, different people read the same text in different ways. The pronunciation time of different people with the same phoneme are different. Due to this difference, the prosody of each person's pronunciation is greatly different. The other aspect is the voice of the speaker. This difference is mainly due to the different vocal organs of each person. If there is enough data for a speaker, we can learn the speaking style by adapting on the end-to-end model. When we need speaker adaptation with limited data, to prevent overfitting, it is better to reduce the parameter that requiring updating. The updated parameters should be closely related to the speaker's characteristics.

Therefore, to update the model parameters accurately, we introduce spoken content features as prior knowledge to decompose the acoustic model space. The acoustic model are decomposed into two independent parts. One part is responsible for predicting spoken content, and the spoken content is represented by PPG features. The other part is to convert the predicted PPG into the acoustic features. So the TTS model predicting acoustic features y given input x can be expressed as:

$$P(y|x;\theta) = P(y|PPG;\theta_1) * P(ppg|x;\theta_2) \quad (1)$$

$$\theta = \theta_1 + \theta_2 \quad (2)$$

Through the formulas, we can find that the original parameters θ is decomposed into two independent parameter spaces. The parameter θ_1 is responsible for predicting PPG given text information, and θ_2 is responsible for voice conversion. When we need to adapt on a new speaker, we just need to extract the speaker's text-ppg pair and ppg-acoustic feature pair, then adapt the two modules separately. In this way, we will get a more stable voice of the speaker.

3.2. Adaptation strategies

In real life, the corpus we get is often without text annotation, and it is impossible to conduct speaker adaptation on end-to-end model directly. Here, we design two adaptive strategies for whether there is text annotation for our proposed framework.

3.2.1. Adapt the whole style

If the corpus has text annotations, we can transfer all the speaking characteristics. We need to train a multi-speaker PPG predictor and voice conversion model in advance. When it comes to speaker adaptation, we need to prepare text-PPG pairs and PPG-acoustic feature pairs. Then fine-tuning the pre-trained PPG predictor and voice conversion modules.

In this way, the model can learn all characteristics of the speaker including the style of spoken content and the style of voice. However, the corpus obtained in the real scene often doesn't have text annotation, and we have to do text annotation manually, which brings difficulties to adapt a new speaker fastly.

3.2.2. Just adapt the voice

When the corpus doesn't have text annotations, we can only transfer the speaker's voice on our proposed framework. We need to train a multi-speaker voice conversion model in advance and use the data to fine-tune the voice conversion model. This method is easy to practice and can also ensure the similarity and stability of the synthesized speech.

This strategy is more operable in real scenes. With the rapid development of the Internet, we can easily obtain a large amount of speech without annotations. We can use this method to adapt the voice of target speaker quickly, then use the PPG predictor

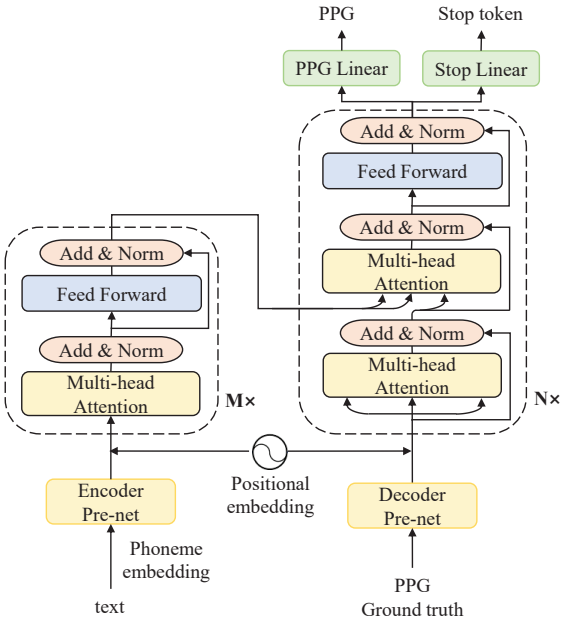


Figure 2: *PPG predictor architecture.*

module to drive the target speaker to say anything. Additionly, we can pre-train a variety of PPG predictors in different styles, so that the target speaker's speaking style is more controllable.

3.3. PPG predictor

The task of the PPG predictor is to predict PPG information given text information. As far as we know, there is no research on this task before. Since the multi-head attention mechanism [14] can model the relationship between two sequences with different lengths well, in this paper, we propose a PPG predictor module based on the transformer framework [14]. Inspired by the structure in paper [3], the structure of PPG predictor is shown in the Fig. 2.

4. Experiments

In this section, we conduct experiments to evaluate our proposed method on VCTK corpus [15]. The audio data were produced by 109 speakers in English with different accents. We randomly select 10 speakers' utterance as testing set, and the rest utterances are split to 90% training set and 10% validation set. All the wav files are sampled at 16KHz.

4.1. Model details

Acoustic features are extracted with 10 ms window shift. LPCNet [16] is utilized to extract 32-dimensional acoustic features, including 30-dimensional BFCCs [17], 1-dimensional pitch and 1-dimensional pitch correction parameter. The 512-dimensional PPGs are extracted from the acoustic model in the speaker independent automatic speech recognition (SI-ASR) system. The SI-ASR system is implemented using the Kaldi toolkit [18] and trained on our 20,000 hours Mandarin corpus. The voice conversion model's structure follows the structure in paper [13]. The numbers of PPG predictor's encoder block and decoder block are 3. The hidden dimensionals of PPG predictor are 512.

In this section, we train four system to illustrate the effectiveness of our proposed model.

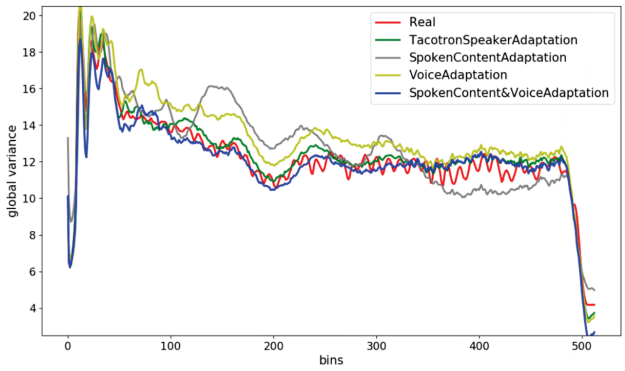


Figure 3: *The global variance of target speaker's real speech and the synthesized speech. 100 randomly chosen utterances are used to calculate the variance.*

- **TacotronSpeakerAdaptation** stands for performing speaker adaptation on tacotron. The structure details of tacotron are same as the tacotron in paper [2].
- **SpokenContentAdaptation** stands for only fine-tuning the PPG predictor module on our proposed framework.
- **VoiceAdaptation** stands for only fine-tuning the voice conversion module on our proposed framework.
- **SpokenContent&VoiceAdaptation** stands for fine-tuning both the PPG predictor and voice conversion modules on our proposed framework.

4.2. Spoken content and voice factorization visualization

To show the effectiveness of spoken content and voice factorization, we use the global variance (GV) [19] as the visualization of spectral distribution. In our experiment, we calculate global variance of each frequency bin for all the utterances from a speaker (or synthesized speaker).

We compare the global variance of the speaker's real speech and the synthesized speech. We choose 200 sentences as adaptive corpus and the results are shown in Fig. 3. We can find that the GV curves of the speech synthesized by **TacotronSpeakerAdaptation** and **SpokenContent&VoiceAdaptation** are more in line with the real curve. However, the result of spoken content adaptation or voice adaptation is obviously different from the real curve. Specifically, **VoiceAdaptation** can make the GV curve shape more similar to the GV curve of real speech. By comparing curve of model **VoiceAdaptation** and **SpokenContent&VoiceAdaptation**, it reveals that spoken content adaptation can further reduce the deviation between the GV curve of synthesized speech and the GV curve of real speech.

4.3. Voice stability

When the model is overfitting, the synthesized speech will be unstable. To show the stability of the synthesized speech, we compared the spectrums of different methods, which is shown in the Fig. 4. When the num of sentences is 200, the synthesized spectrums of both model **TacotronSpeakerAdaptation** and model **SpokenContent&VoiceAdaptation** are clear. However, when the adaptive data is only 10 sentences, the synthesized spectrum of model **TacotronSpeakerAdaptation** becomes rough and unnatural. This phenomenon is mainly appeared in the low-frequencies and shown in the part 1 of the

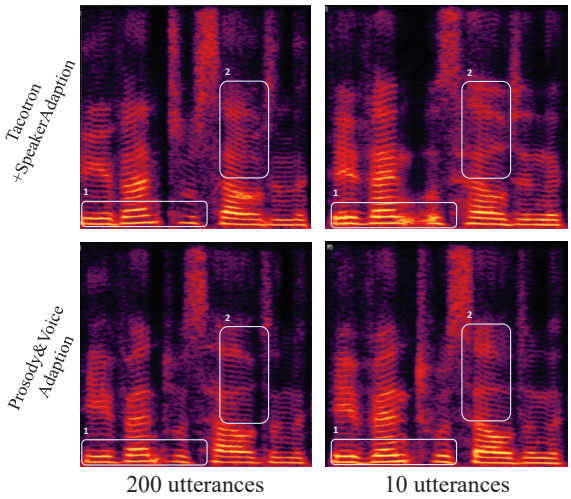


Figure 4: *The comparison of spectrums between TacotronSpeakerAdaptation and SpokenContent&VoiceAdaptation*

Fig. 4. At the same time, there will be more harmonics in the high-frequencies and shown in the part 2. Those proplem will reduce the naturalness of synthesized speech. However, from model **SpokenContent&VoiceAdaptation**, it can be found that the spectrum is clear at both low and high frequencies. This shows that our proposed model is better suited to small amounts of data, even if the num of sentences is only 10.

4.4. Subject evaluation

We conduct Mean Opinion Score (MOS) listening test for naturalness and similarity of speech. 20 listeners participated in the evaluation. In each experimental group, 20 parallel sentences are selected randomly from testing sets of each system. Naturalness and similarity are shown in the Fig. 5 and Fig. 6.

Firstly, as the adaptive data dropped from 200 sentences to 10 sentences, the performance of model **TacotronSpeakerAdaptation** significantly decreased. Especially when the num of sentences is 10, the MOS score of natural and similarity are at a very low level, which indicates that the model is overfitting. By comparing model **TacotronSpeakerAdaptation** and **SpokenContent&VoiceAdaptation**, it can be found that by factorizing spoken content and voice, the performance of the model **SpokenContent&VoiceAdaptation** does not change much after speaker adaptation with fewer data. This phenomenon shows that the decomposition of speech content and voice can effectively prevent the model from overfitting.

Secondly, by comparing models **VoiceAdaptation** and **SpokenContent&VoiceAdaptation**, we can find the difference between the two adaptive strategies proposed in this paper. When the amount of data is large, such as 200 sentences, the result of model **SpokenContent&VoiceAdaptation** is better than model **VoiceAdaptation**. While the data is only 10 sentences, the result of the two is similar. This is because fine-tuning on PPG predictor requires large data to learn spoken content style of target speaker. When the data is few, whether fine-tuning PPG predictor model has little effect on the final effect.

Thirdly, by comparing model **VoiceAdaptation** and **SpokenContentAdaptation**, we can find the magnitude of the impact of spoken content adaptation and voice adaptation on the similarity. Voice adaptation plays an important role in similarity, while spoken content adaptation has little influence on similarity. Therefore, separating adaptation can guarantee the voice adaptation is done well and ensure the similarity of the synthesized speech.

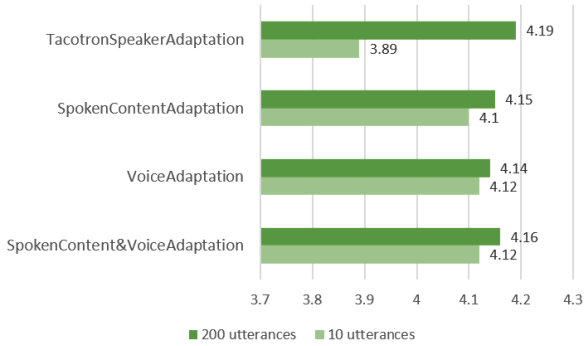


Figure 5: *Naturalness test.*

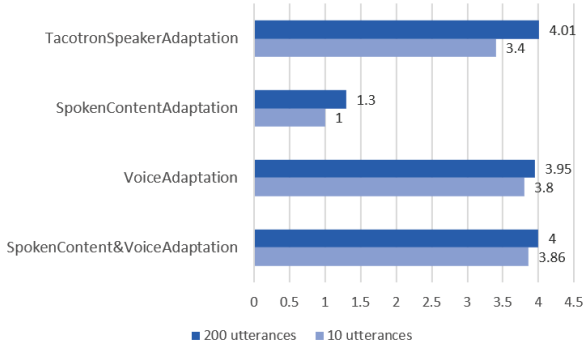


Figure 6: *Similarity test.*

5. Conclusions

In this paper, we propose a novel speaker adaptation method that decomposing the parameter space of the original acoustic model into two parts, with the one on modeling spoken content and the other on modeling speaker's voice. When we conduct the speaker adaptation, fine-tuning on the two subspaces can effectively prevent the model from overfitting. Moreover, we design two adaptive strategies for whether the obtained data has text annotations, and the model can learn the characteristics of speaker quickly. Experiments show that our model can better adapt to a small amount of data, and prevent the model from overfitting effectively. We will continue to explore more elaborate parameter space division in the future.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2018YFB1005003), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379) and Inria-CAS Joint Research Project (No.173211KYSB20170061 and No.173211KYSB20190049). This work is also supported by the CCF-Tencent Open Research Fund.

7. References

[1] 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," *arXiv preprint arXiv:1703.10135*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[5] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for dnn-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] Y. Zhao, D. Saito, and N. Minematsu, "Speaker representations for speaker adaptation in multiple speakers blstm-rnn-based speech synthesis," *space*, vol. 5, no. 6, p. 7, 2016.

[7] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 171–176.

[8] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4475–4479.

[9] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual blstm for speech synthesis of low-resource languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5545–5549.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] T. M. Mitchell *et al.*, "Machine learning," 1997.

[12] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for dnn-based tts synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5135–5139.

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

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

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

[17] T. Gulzar, A. Singh, and S. Sharma, "Comparative analysis of lpcc, mfcc and bfcc for the recognition of hindi words using artificial neural networks," *International Journal of Computer Applications*, vol. 101, no. 12, pp. 22–27, 2014.

[18] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[19] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.