A Study on Acoustic Parameter Selection Strategies to Improve Deep Learning-Based Speech Synthesis

Hyeonjoo Kang*, Young-Sun Joo*, Inseon Jang†, Chunghyun Ahn† and Hong-Goo Kang*
* Yonsei University, Seoul, Republic of Korea
E-mail: {volleruhe, disfruta}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr
† ETRI, Daejon, Republic of Korea
E-mail: {jinsn, hyun}@etri.re.kr

***Abstract*—In this paper, we investigate the variation in the performance of a deep learning-based speech synthesis (DLSS) system based on the configuration of output acoustic parameters. Our method is mainly applicable for vocoding-based statistical parametric speech synthesis (SPSS), which has advantages in low-resource scenarios. Given the independence assumption of the source-filter model for the spectral and fundamental frequency F0 parameters, we propose a reliable network architecture for training acoustic parameters. Particularly, the F0 parameter suffers from high fluctuation and an extremely low number of dimensions. To relieve these problems, we introduce a context-window approach. Furthermore, we apply data augmentation to the proposed structure to overcome a lack of training data, which is a frequent issue with multi-speaker TTS systems. Experimental results confirm the superiority of the proposed algorithm over conventional ones in both single-speaker and multi-speaker TTS setups.**

I. Introduction

Recently, research on deep learning-based speech synthesis (DLSS) systems has advanced significantly thanks to the superior capability of deep neural networks in modeling the relationship between linguistic and acoustic parameters [1], [2], [3]. In the training stage, a typical DLSS system estimates the optimal network parameters by learning a relationship between the linguistic and acoustic parameters. When input text is given, i.e. during the synthesis stage, acoustic parameters with the highest likelihood are predicted from the trained networks, then a speech waveform is generated using a vocoding process using these predicted acoustic parameters [1], [4].

DLSS systems can be categorized into two types depending on acoustic parameter that is to be estimated. The first type is a family of Tacotron-like end-to-end structures [5], [6], [7] that estimates the mel-spectrum for acoustic parameters as a condition vector for a neural vocoder, i.e. WaveNet [8]. The second type estimates vocoding parameters which can be derived as spectrum- and excitation-related parameters. These estimated vocoding parameters are fed into statistical vocoders to generate speech waveforms.

In this paper, we choose the second system type referred as statistical parametric speech synthesis (SPSS), as a baseline because of its high flexibility in terms of speaker adaptation and its reasonably good synthesized speech quality in a small footprint system setup [9]. In addition, the SPSS approach has a much shorter training time than the other DLSS system type.

DLSS systems do not consider the independence assumption of the source-filter parameters [10] or the dimensional imbalance between these parameters meaning training results are often biased [11], [12]. Note that the number of dimensions for spectral parameters, *e.g.*, line spectral pairs (LSPs), is typically larger than 20, while the number of dimensions for excitation parameters, *e. g.*, fundamental frequency (F0), is only 1 per frame.

Several papers have concluded that these acoustic parameters are required to be trained separately [13], [14], [15], but they do not clearly explain the reason for this conclusion. In addition, the authors do not consider the important characteristics of the F0 parameter during the training process, *i.e.*, its severe fluctuation in consecutive frames due to high variance. In [11], [16], [17], the authors attempted to unravel the complicated characteristics of F0 using wavelet transforms [18] for text-to-speech (TTS) and voice conversion systems. Although they insisted that wavelet transforms enhanced modeling performance by decomposing the F0 parameters into a hierarchical structure, but still they could not fully analyze the fluctuating characteristics in the training process.

To further improve acoustic modeling performance, we investigate various configurations of acoustic parameters for a deep learning-based framework. Based on the results, we propose a reliable architecture under the assumption of a source-filter vocoding structure. The proposed architecture successfully resolves the issues surrounding F0 modeling, *i.e.*, low dimensionality and high fluctuation. We apply the structure of the proposed system to a data augmentation task. It is well-known that DLSS systems require a large volume of training data to obtain high-quality synthesized speech, which requires a long time and is expensive to record. The proposed method [1] is very useful when a network is trained using databases from multiple speakers. We could get benefit even if the characteristic of each speaker's database is different.

The remainder of this paper is structured as follows. Section II investigates the variation in the performance for different combinations of acoustic parameters in a source-filter model framework. In Section III, the proposed approach is described in detail, and the effectiveness of using data from multiple

[1]The proposed approach is useful even when the database for each speaker is not large.

Table I: Spectral modeling performance [dB]

	LSP only	LSP and BAP	All features combined
LSD [dB]	5.00	**4.71**	4.79

speaker to improve performance is investigated. The superiority of the proposed system is verified with experiments in Section IV, while Section V summarizes and concludes the paper.

II. Impact of different acoustic parameter combinations on a DLSS system

This section investigates various strategies to increase the modeling accuracy of acoustic parameters, *e.g.*, spectral and F0 parameters, in a deep learning-based training process. First, we seek an effective combination of acoustic parameters by examining he interactions between these parameters. Even though the experimental results in [13], [14] showed that it would be better to train spectral and F0 parameters separately because of the independence assumption of the source-filter model, there was no clear explanation for this. Second, we investigate a method to increase F0 modeling accuracy by solving the imbalance in dimension issue.

A. Experimental environment

The systems in this section use a 160-minute speech dataset. The speaker is male and a native speaker of Korean. The input features are linguistic parameters consisting of phonetic and syntactic information and phoneme duration, with a total of 210 dimensions. The output features are acoustic parameters extracted from the STRAIGHT vocoder [19]. The acoustic parameters consist of LSP, F0, band aperiodicity (BAP), and a voiced/unvoiced (V/UV) flag including their dynamic parameters with a total of 139 dimensions.

The network architecture chosen for the experiment is a stack of vanilla fully-connected (FC) layers that consist of three hidden layers with 1024 nodes. We use stochastic gradient descent (SGD) with momentum [20] as the optimization criterion [21], [22]. The weights of the networks are initialized by applying a layer-wise pre-training method [23] and different learning rates (0.001, 0.005, and 0.01) and momentum values (0.5, 0.5, and 0.9) are utilized at each step.

Log-spectral distance (LSD) and root mean-square error (RMSE) are used to evaluate the estimation accuracy of the spectral and F0 parameters, respectively. Both LSD and RMSE measure Euclidean distance between generated acoustic parameters and the original one extracted from recorded speech.

B. Analysis of spectral parameter modeling performance

We compare spectral modeling performance for following three configurations as follows.

- **LSP-only system**: LSP and V/UV
- **LSP and BAP system**: LSP, V/UV and BAP
- **All features combined system**: LSP, V/UV, BAP and logarithmic F0 (LF0).

All three configurations include LSP and V/UV because generating LSP is a target task and V/UV is strongly correlated to the 0th LSP coefficient, which represents energy of observation [24]. The final two configurations have the spectral-related parameter BAP, while only the final configuration has the excitation-related parameter, LF0. For better trainability, LF0, which has lower variance, is adopted instead of using F0 directly.

All of the network architectures and settings are exactly the same except for the output parameter settings. Table I shows LSDs obtained in the three configurations.

Because LSP and LF0 are assumed to be independent, it may not be beneficial to train all acoustic parameters together within a single output layer. Training with the spectral-related parameters *LSP and BAP* produces the best performance.

C. Analysis of F0 parameter modeling performance

In this subsection, we focus on the influence of the spectral parameters when modeling F0. We also attempt to resolve the dimension problem that occurs in the F0 parameter estimation process. First, we compare the following two configurations.

- **LF0-only system**: LF0 and V/UV
- **All features combined system**: LF0, V/UV, BAP, and LSP.

Table II presents the RMSE values obtained from each configuration. Modeling F0 solely produces a lower performance than the configuration with all features combined, which does not match with the results of the spectral parameter modeling in the previous subsection. However, it is unclear whether the cause of the poor performance is because spectral parameters are helpful for training F0 parameters or because the dimensional imbalance between the spectral and F0 parameters is a problem. Therefore, we compare two additional configurations:

- **Context LF0 system**: Concatenated LF0 with previous and succeeding frames and V/UV
- **Context LF0, LSP and BAP system**: Concatenated LF0 with previous and succeeding frames and of all the spectral parameters (LSP, BAP and V/UV)

Because utilizing F0 only from the current frame makes it difficult to predict the change in F0 over time, we introduce a context LF0 framework. Context LF0, widely utilized in the automatic speech recognition (ASR) field [25], involves the concatenation of succeeding LF0 parameters from previous and subsequent frames. This context window enables to provide temporal information using consecutive frames, thus it is also expected to resolve the low-dimensional data problem.

However, it is not clear whether spectral parameters will be beneficial for the modeling of F0 even if the dimensional mismatch issue is solved using the context-window method. By comparing the *context LF0 system* with the *context LF0, LSP, and BAP system*, the effectiveness of the use of spectral

Table II: F0 modeling performance [Hz]

	LF0 only	All features combined	Context LF0	Context LF0, LSP and BAP
RMSE [Hz]	14.38	11.91	**11.52**	19.95

parameters for F0 modeling can be verified. The network architecture and settings are the same as described in Subsection II-A. The context-window size is set to 44, *i.e.*, previous and 22 subsequent frames.

As shown in Table II, F0 modeling accuracy is improved by introducing the context-window method. On the other hand, context LF0 modeling with spectral parameters produces a lower performance than the context LF0 system. This means that the separate training of acoustic parameters improves modeling accuracy, thus it is clear that the reason for the poor performance of the *LF0-only system* is the dimensional imbalance issue. It can be concluded that the context-window approach benefits the F0 modeling process by introducing temporal patterns and resolving the dimensional problem, but the spectral parameters have a negative influence on the F0 modeling process.

III. Proposed system

Based on the experiments conducted in Section II, we summarize the key factors related to source-filter model based processing:

- It is better to separately train spectral and excitation parameters.
- It is useful to introduce a context window approach for F0 training.

Based on these conclusions, we propose a DLSS architecture that improves modeling accuracy by enhancing the trainability of the acoustic parameters. Figure I depicts the conventional and proposed architecture, respectively. The proposed architecture on the right side of the figure is referred to as a feature-type dependent architecture (FD-DLSS) and the conventional DLSS architecture on the left is referred to as feature-type independent architecture (FI-DLSS). The proposed architecture separates the hidden and output layers for the spectral and F0 parameters. This enables the independence assumption of the source-filter model to be kept. Specifically, the context window exploits temporal patterns to overcome the considerable fluctuation of the F0 parameter.

Even though the proposed architecture improves modeling performance, a sufficient volume of speech data is required to reliably train the architecture. To cope with the lack of data, one solution is to utilize data from other speakers. However, we should be very careful not to distort the key characteristics of the target speaker while generating synthesized speech. Therefore, we extend the proposed architecture to ensure that is suitable for data augmentation.

Originally, data augmentation was designed to obtain more generalized models by seeking transform-invariant characteristics within a diverse range of data, and this technique has been

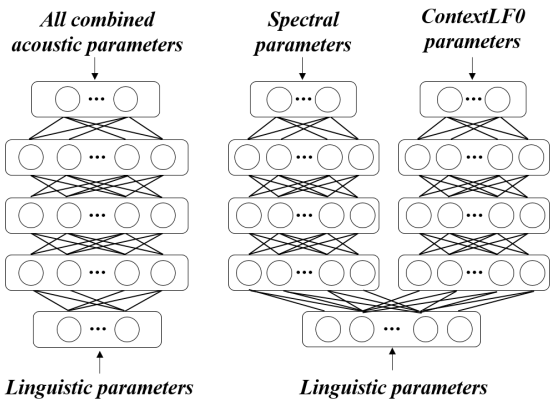

Figure I: Conventional feature-type independent DLSS (FI-DLSS) structure (left) and proposed feature-type dependent DLSS (FD-DLSS) structure (right)

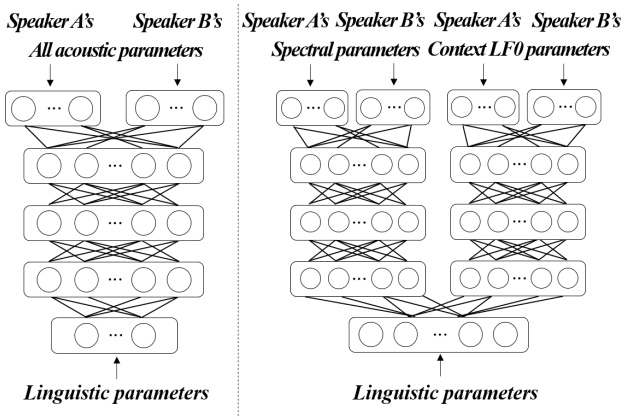

Figure II: Conventional SHL-based feature-type independent DLSS (SHL-FI-DLSS) structure (left) and proposed SHL-based feature-type dependent DLSS (SHL-FD-DLSS) structure (right)

successfully employed in a wide range of tasks such as speech enhancement and speech recognition [26], [27], [28]. The architecture with data augmentation is is depicted in Figure II. We introduce a shared hidden layer (SHL) structure [29] that shares hidden layers and only separates the output layers to define the corresponding tasks. Though the objective of finding a mapping relationship between linguistic and acoustic parameters is the same because this is type of TTS system, the acoustic characteristics of the target speaker should not be changed with the use of different speakers, so the use of an SHL structure is a desirable option.

The generation phase is the same as for the conventional DLSS system. The acoustic parameters are estimated from the linguistic parameters extracted from the given input text. Maximum likelihood parameter generation with global variance [4] is used to avoid the over-smoothing and then synthesized speech us obtained using a vocoder.

Table III: Experimental result for a single speaker's database. FI-DLSS = feature-type independent-based system ; FD-DLSS = feature-type dependent system

	FI-DLSS	FD-DLSS
LSD [dB]	4.79	**4.71**
RMSE [Hz]	11.92	**11.52**

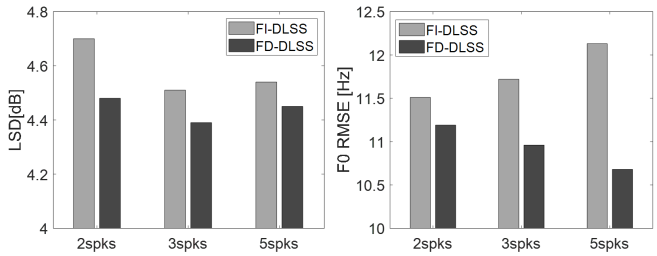

Figure III: Performance measurement of data augmentation scenario

IV. Experiment

A. Corpus construction and system notation

In the following experiments, we use a Korean speech database recorded by five male speakers. The linguistic and acoustic parameters are the same as those described in Section II with the only difference being the output feature configurations.

The output feature configuration of the spectral feature estimation network consists of LSP, BAP, and V/UV and their dynamic information with 136 dimensions in total. For the F0 estimation network, a combination of context-window LF0 size of 22, their dynamic information and V/UV parameters is the output feature. The hidden layer is a stack of four FC layers with 1024 nodes in both networks, optimized using the SGD algorithm. For initialization, a layer-wise back-propagation method is used. The learning rate and momentum are set to 0.01 and 0.9, respectively.

B. Analysis of the proposed structure

To verify the superiority of the proposed FD-DLSS architecture, we compare the conventional FI-DLSS. The experimental results are presented in Table III. The proposed architecture produces a better performance than the conventional architecture for both spectral and F0 modeling.

C. Analysis on proposed structure with data augmentation

In this subsection, we further demonstrate the suitability of our model for use with data augmentation to resolve the lack of data, which is inevitable when the training database is small. We conduct experiments in which we increase the number of speakers when training the structure. The speech database size of each speaker is set at 80 minutes to allow a fair comparison. As described in Section III, the SHL-based structure is used to separate different speaker characteristics in separate output layers. The network settings are the same as described in the previous subsection.

The experimental results are presented in Figure III. Overall, spectral modeling accuracy improves as the size of the database increases, which confirms that data augmentation also works for speech synthesis. Furthermore, the proposed parameter-separated training approach (SHL-FD-DLSS) improves the modeling accuracy compared to the conventional method. The use of three speakers (3spks) produces a slightly better performance compared to the use of five speakers (5spks) regardless of the type of structure employed.

For F0 modeling, the SHL-FD-DLSS architecture improves model accuracy as the volume of data increases, whereas the accuracy of the SHL-FI-DLSS is worsens. Because the data is augmented using speech data from several speakers, the significant inter-variability of the F0 parameters prevents the modeling of the F0 parameter for the target speaker within the SHL-FI-DLSS architecture. However, our proposed architecture overcomes this problem by including long-term temporal features via the concatenation of adjacent frames. In addition, half of the SHL-FD-DLSS structure is designed to fully concentrate on modeling F0, which results in improved performance.

D. Subjective evaluation

To evaluate the perceptual quality of the proposed system, ABX preference listening tests were conducted. We compared our proposed FD-DLSS system with the conventional FI-DLSS system during two sessions, one with a single-speaker and one with multiple-speakers (SHL-*-DLSS). Thirteen native Korean listeners were asked to choose their preferred synthesized speech sample based on speech quality after listening to 20 sentences per session.

Table IV presents the scores for the ABX preference test. The fourth column of Table IV indicates that the proposed architecture is clearly better than the conventional architecture for both the single- and multiple-speaker sessions. In particuler, for the single-speaker data 95.4% of the listeners selected the proposed architecture, which means that our proposed system is robust when faced with a laparticularck of data.

V. Conclusions

In this paper, we proposed a reliable network architecture that allows feature-dependent training and considers the independence assumption of the source-filter model for the spectral and excitation parameters. The proposed DLSS architecture also solves the dimensional imbalance issue for the F0 parameter by introducing the context-window method. The superiority of the proposed architecture was verified through objective and subjective tests, with the proposed method producing a strong performance in a multiple-speaker TTS environment.

Acknowledgment

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2019-0-00447, Development

Table IV: Subjective preference test results (%)

	FI-DLSS	no preference	FD-DLSS	p-value
Single-spk	1.5	3.1	95.4	$< 10^{-13}$
Multi-spk	25.0	20.1	54.9	$< 10^{-10}$

of emotional expression service to support hearing/visually impaired)

REFERENCES

[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7962–7966, May 2013.

[2] E. Song and H. Kang, "Deep neural network-based statistical parametric speech synthesis system using improved time-frequency trajectory excitation model," *in Proc. Interspeech*, September 2015.

[3] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3829–3833, May 2014.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1315–1318, 2000.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *in Proc. Interspeech*, pp. 4006–4010, 2017.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *in Proc. Interspeech*, pp. 4779–4783, 2018.

[7] S. O. Arik, G. F. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *in Proc. Neural Information Processing Systems (NIPS)*, 2017.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuogluu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[9] Z. Liu, Z. Ling, and L. Dai, "Statistical parametric speech synthesis using generalized distillation framework," *IEEE Signal Processing Letters*, pp. 695–699, May 2018.

[10] G. Fant, *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Walter de Gruyter, 1960.

[11] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with different temporal scales of F0 based on Wavelet transform," *The 9th ISCA Speech Synthesis Workshop*, pp. 140–145, 2016.

[12] C. Wang, Z. Ling, and L. Dai, "Asynchronous f0 and spectrum modeling for hmm-based speech synthesis," *in Proc. Interspeech*, pp. 404–407, 2009.

[13] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," *The 9th ISCA Speech Synthesis Workshop*, pp. 181–186, 2016.

[14] X. Wang and S. Takaki and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," *Speech Communication*, pp. 1–9, 2018.

[15] H. Luong and J. Yamagishi, "Scaling and bias codes for modeling speaker-adaptive DNN-based speech synthesis systems," *2018 IEEE Workshop on Spoken Language Technology (SLT)*, 2018.

[16] M. S. Ribeiro and R. A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4909–4913.

[17] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," *in Proc. IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–5, 2016.

[18] Y. Meyer, *Wavelets and Operators*. Cambridge University Press, 1992.

[19] H. Kawahara, I. M. Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communications*, vol. 27, pp. 187–207, 1999.

[20] G. D. I. Sutskever, J. Martens and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML)*, 2013, pp. III–1139–III–1147.

[21] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, pp. 462–466, 1952.

[22] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *arXiv preprint arXiv:1606.04838*, 2006.

[23] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," *IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 24–29, 2011.

[24] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation," *in Proc. International Conference on Spoken Language Processing (ICSLP 94)*, September 1994.

[25] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, pp. 82–97, 2012.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *in Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[27] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, March 2017.

[28] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," 2013.

[29] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7304–7308, 2013.