SPEAKER ADAPTATION OF A MULTILINGUAL ACOUSTIC MODEL FOR CROSS-LANGUAGE SYNTHESIS

Ivan Himawan[1*], *Sandesh Aryal*[1*], *Iris Ouyang*[1*], *Sam Kang*[1], *Pierre Lanchantin*[1], *Simon King*[2]

[1]ObEN, Inc., Pasadena, California, USA

[2]Centre for Speech Technology Research, University of Edinburgh, UK

{ivan, sandesh, iris, sam, pierre}@oben.com, simon.king@ed.ac.uk

ABSTRACT

Several studies have shown promising results in adapting DNN-based acoustic models as a mechanism to transfer characteristics from pre-trained models. One such example is speaker adaptation using a small amount of data, where fine-tuning has helped train models that extrapolate well to diverse linguistic contexts that are not present in the adaptation data. In the current work, our objective is to synthesize speech in different languages using the target speaker's voice, regardless of the language of their data. To achieve this goal, we create a multilingual model using a corpus that consists of recordings from a large number of monolingual and a few bilingual speakers in multiple languages. The model is then adapted using the target speaker's recordings in a language other than the target language. We also explore if additional adaptation data from a native speaker of the target language improves the performance. The subjective evaluation shows that the proposed approach of cross-language speaker adaptation is able to synthesize speech in the target language, in the target speaker's voice, without data spoken by the target speaker in that language. Also, extra data from a native speaker of the target language can improve model performance.

Index Terms— Multilingual, DNN, TTS, cross-language, speaker embedding, subjective evaluation

1. INTRODUCTION

There is a widespread demand for multilingual text-to-speech (TTS) services, in which one TTS engine can synthesize natural and intelligible speech in different languages. One of the important features in multilingual TTS applications is to generate high-quality speech of a speaker in different languages while still being perceived as spoken by the same speaker. For some TTS applications such as chatbot avatars, navigation systems, and speech-to-speech translation, it is important to maintain the voice identity of the original speaker, as voice switch is not desirable when we switch from one language to another [1, 2]. There is also a growing need for customized voices in multilingual TTS services, where some users prefer celebrity voices while some others prefer voices of their loved ones. Such use cases require the personalization of multilingual TTS systems using a small amount of data spoken in one language.

Recently, the neural TTS system is capable of producing very high-quality speech samples with human-like prosody. Typically, several specialized modules in the conventional statistical parametric speech synthesis (SPSS) pipeline are combined together in the neural TTS framework to form a single neural network. For example, neural vocoders such as WaveNet [3] and SampleRNN [4] directly convert linguistic features into the waveform. Other systems such as Char2Wav [5] and Tacotron [6] directly map the input text into acoustic features. In DeepVoice [7, 8], the entire TTS pipeline is implemented using a similar structure as the traditional TTS system by replacing all modules with neural networks. Most of these methods aim to disentangle speaker specific characteristics such as timbre, style, accents from speech so that it is possible to manipulate these characteristics [9, 10, 11]. However, a large amount of high-quality recordings is usually required to adapt neural TTS systems to a new voice, especially for cross-language synthesis (i.e., synthesis in a language different from the target speaker's language, where both languages are included in the training data) [12, 11, 13]. In contrast, the speech synthesis techniques based on multi-layer feed-forward neural networks can be easily adapted using a small amount of data while maintaining naturalness as well as similarity to the target speaker [14]. Furthermore, a simpler neural network such as feed-forward architecture is suitable for real-time implementation.

There are few examples of multilingual TTS systems that employ neural network architectures in the literature. The multilingual TTS approach proposed in [15] uses long short-term memory (LSTM) architecture and adopts cluster adaptive training (CAT) to model all variations of training languages and speaker variations. They show that a multi-language multi-speaker (MLMS) model (trained on six European languages) can be adapted to new languages (i.e., Polish and Portuguese) using limited training data and the performance is better compared to building the models from scratch. However, no result is reported for distant, unrelated languages such as East Asian languages. In [16], a single-speaker model is built using an LSTM architecture that avoids voice switch when synthesizing different languages. The model is built using bilingual data of a speaker who speak both languages. Hence, it is not possible to synthesize languages that the target speaker does not speak. Another method, employed in [17], factorizes deep neural networks (DNN) using speaker-specific layers and language-specific layers to model speaker and language characteristics in the data. They show that polyglot synthesis is possible without using speech data from a bilingual speaker, albeit the lower speaker similarity compared to DNN-based monolingual systems.

In this work, speakers in different languages are pooled together to build a speaker-independent multilingual acoustic model. This enables the model to learn the average statistics of what speech sounds like from training data of multiple speakers including male and female, various speaking styles, languages, and recording conditions [12, 18]. As a result, the model parameters are particularly well-suited for adaptation to speakers speaking in different languages or accents. The contributions of this paper include: (1) we propose a transfer learning approach from a multilingual model via DNN adaptation to the target speaker and show that it is possible to synthesize different languages in the voice of the speaker regard-

[*]Equal Contribution.

Authorized licensed use limited to: University of Wollongong. Downloaded on May 31,2020 at 18:26:42 UTC from IEEE Xplore. Restrictions apply.

less of the languages they speak, and (2) we investigate if model adaptation can be improved by using additional data from the target language spoken by a different but native speaker of that language. The rest of the paper is organized as follows. Section 2 discusses related work. The cross-language speaker adaptation technique is presented in Section 3. Section 4 describes features, databases, and DNN configurations. Experimental results are presented and discussed in Section 5 and 6. Finally, Section 7 concludes the paper.

2. RELATED WORK

Many techniques have been developed in the past to synthesize speech in different languages with the same voice for HMM-based synthesis [1, 19, 2, 20]. The cross-language synthesis can be achieved using mixed-language methods [1], phone mapping methods [21], and state sharing mapping methods [19]. For example, [1] trained an average voice model (AVM) using data from several speakers in multiple languages and adapted to a target speaker that spoke one of the languages used in the training. Even though similar techniques have been proposed for DNN-based speech synthesis [15, 17, 16], most works in the literature have focused on monolingual systems that can generate multiple voices [22, 23, 24, 25]. Recently, a multilingual TTS able to transfer voices across languages based on Tacotron is proposed in [11]. The model to estimate acoustic features was conditioned on linguistic representations, speaker embeddings and language embeddings. They used adversarial loss to disentangle the correlation between language and speaker identity in the training datasets. In the current work, the disentangled representations of speaker identity and linguistic information are realized using a separate speaker encoding network, and hand-crafted linguistic features and language codes. Thus, we can achieve similar results but with a simpler and more efficient frame-by-frame mapping using a feed-forward network.

3. CROSS-LANGUAGE SPEAKER ADAPTATION

Our cross-language speaker adaption approach is designed for the conventional text-to-speech system that uses a simple feed-forward DNN for acoustic modeling. In this particular case, the TTS is multilingual and it consists of: a) a text analyzer that converts input text from multiple languages to linguistic feature vectors, b) language-specific duration models that estimate phoneme duration from linguistic features given by the text analyzer, c) an acoustic model to estimate acoustic features from linguistic features, and d) a vocoder to generate audio from estimated acoustic features. We use WORLD [26] as the vocoder. There are a few critical differences worth mentioning between our implementation of multilingual TTS and a conventional monolingual TTS. First, the linguistic features given by the text analyzer includes language code (one hot vector for the languages included). While we still include a few language-specific features, we convert critical language-specific features such as phoneme ID to descriptors common across languages such as places and manner of articulation, voicing, etc. Linguistic features also include positional information of phoneme/syllable. Second, the input to the acoustic model comprises not only linguistic features (including the language code) but also speaker embeddings extracted from the neural networks (e.g., d-vectors) [27, 9]. Speaker embeddings are necessary to train an acoustic model with multi-speaker datasets [28]. Figure 1 shows our training strategies for the base multilingual acoustic model and speaker adaptation. The base model is trained using a corpus that consists of recordings by a wide range of speakers in the four languages listed in 4.2. Most

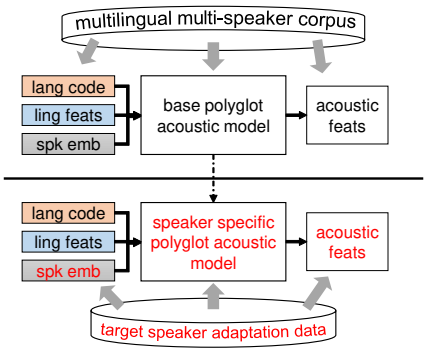

Fig. 1. Illustration of the acoustic model adaptation technique for multilingual TTS.

speakers only recorded in one language; a few speakers recorded in two of the four languages. Our speaker adaptation strategies involve fine-tuning of the base acoustic model using speech data from the target speaker. We propose two adaptation strategies. In the first strategy (baseline strategy), we fine-tune the base model only using the target speaker's recordings in the source language. In the second strategy, we include a large amount of data spoken by a native speaker of the target language, in addition to the target speaker's recordings in the source language. This is a regularization strategy intended to avoid the model from over-fitting to the source language and help the model to generalize better across different languages.

4. EXPERIMENTAL SETUP

4.1. Features

In our experiments, all systems were trained using 60 Mel-Cepstral coefficients (MCCs), 3 band aperiodicities (BAPs), and fundamental frequency F_0 on log scale with their delta and delta-deltas features, and an additional voiced/unvoiced binary feature. These acoustic features were extracted from 48 kHz waveform at 5 millisecond interval. Speaker embeddings are used as an input to encode speaker identity. In our case, identity of the speaker was represented by fixed-dimensional speaker embeddings from a speaker encoder network. We concatenated 200-dim embedding vectors to the linguistic features to form a total of 913-dim input feature vectors that included language codes as augmented features into the DNN.

4.2. Databases

We pooled multi-speaker databases in four languages to build the base multilingual acoustic model. A detailed breakdown is as follows: about 87 hours from 210 speakers in English (EN), about 50 hours from 65 speakers in Korean (KO), about 37 hours from 39 speakers in Japanese (JA), and about 56 hours from 28 speakers in Mandarin Chinese (ZH). These databases were composed of high-quality speech, including studio recordings by voice actors and regular speakers, as well as audiobooks.

The goal of this study is to investigate whether we can synthesize speech when the speakers themselves do not speak the languages. Multiple strategies were devised and compared in the experiments to achieve our objective. In this paper, we evaluated the performance of cross-language adapted acoustic models in ZH-EN and KO-EN language pairs. For this purpose, we recorded speech from two male bilingual target speakers, one fluent in ZH and EN, and another fluent in KO and EN. This bilingual corpus allowed us to compare the performance of cross-language speaker adaptation in four source-target language pairs, namely ZH→EN, EN→ZH, KO→EN, and

Authorized licensed use limited to: University of Wollongong. Downloaded on May 31,2020 at 18:26:42 UTC from IEEE Xplore. Restrictions apply.

EN→KO. For each language pair, we trained three models based on the type and amount of adaptation data. First, we used data (spoken by the target speaker) in the source language only, to train the baseline cross-language adapted acoustic model ($model_1$). Next, we added data spoken by a different, monolingual speaker but in the target language, as an alternative strategy ($model_2$). The purpose of this mixed-speaker data set is to investigate whether additional data from any native speaker of the target language can help the cross-language adaptation performance, especially in terms of accent and quality. Lastly, we used bilingual data that amounted to the same size as the monolingual data set ($model_1$) to train a bilingual reference model ($model_3$). While $model_1$ and $model_2$ represent our two cross-language adaptation strategies, $model_3$ represents the ground truth case for performance comparison. Table 1 lists the amount and type of adaptation data used in cross-language adaptation for all source-target language pairs used in our experiments.

Table 1. Adaptation data used for fine-tuning the base multilingual acoustic model. The additional data from monolingual non-target speakers are indicated by *.

Lang. pair		$model_1$	$model_2$	$model_3$
ZH→EN	type	ZH	ZH + EN*	ZH + EN
	dur. (*min.*)	45	45 + 132	23 + 24
EN→ZH	type	EN	EN + ZH*	EN + ZH
	dur. (*min.*)	45	45 + 132	24 + 23
KO→EN	type	KO	KO + EN*	KO + EN
	dur. (*min.*)	26	26 + 132	12 +13
EN→KO	type	EN	EN + KO*	EN + KO
	dur. (*min.*)	26	26 + 132	13 + 12

4.3. DNN configurations

A feed-forward DNN with ten hidden layers was employed to train multilingual acoustic model. For each hidden layer, a hyperbolic tangent was used as the activation function, followed by a linear activation at the output layer. The weight parameters were initialized randomly using samples drawn from the normal distribution ($\mu = 0, \sigma = 1/\sqrt{hidden\_layer\_input\_size}$), and the models were trained to minimize mean square error using stochastic gradient descent, and a batch size of 1024. We applied batch-normalization to each hidden layer except the first input layer. Learning rate was fixed at 0.001, warm-up momentum was 0.4, drop-out rate was 0.02, and the number of training epochs was 100. In order to fine-tune the model to a specific language in bilingual corpus, all hidden layers from the base polyglot model were adapted. We used Merlin toolkit for training acoustic models [29].

5. SUBJECTIVE EVALUATION

To evaluate the different cross-language speaker adaptation strategies, we conducted subjective listening tests to assess four aspects of model performance: intelligibility, nativeness, acoustic quality, and voice similarity to the target speaker. For each source-language pair, we generated 10 test sentences in the target language using each of the three adapted models [1]. Every audio sample was evaluated by at least 20 unique participants. In all tests, participants were asked to wear headphones and were allowed to listen the sound clips multiple times. In order to focus on performance of the acoustic models, we did not use duration models from the TTS system. Instead, we used phoneme duration given by forced alignment of natural recordings spoken by a different native speaker.

Note that $model_1$ and $model_2$ were trained using our proposed cross-lingual adaptation strategies, and $model_3$ represented the ground truth case where recordings from the target speaker in the target language were available. Using speaker vectors and language codes as inputs into DNN provided a method to control the voice and linguistic characteristics in the output speech of DNN-based speech synthesis. In case of $model_2$, we used data from a different speaker in the target language. We found that it is necessary to use the source language code instead of the target language code when synthesizing the target language. Without this adjustment, we observed voice inconsistency in the model output. Most likely, during the adaptation process, the model learned the correlation between the speakers and the languages present in the adaptation data.

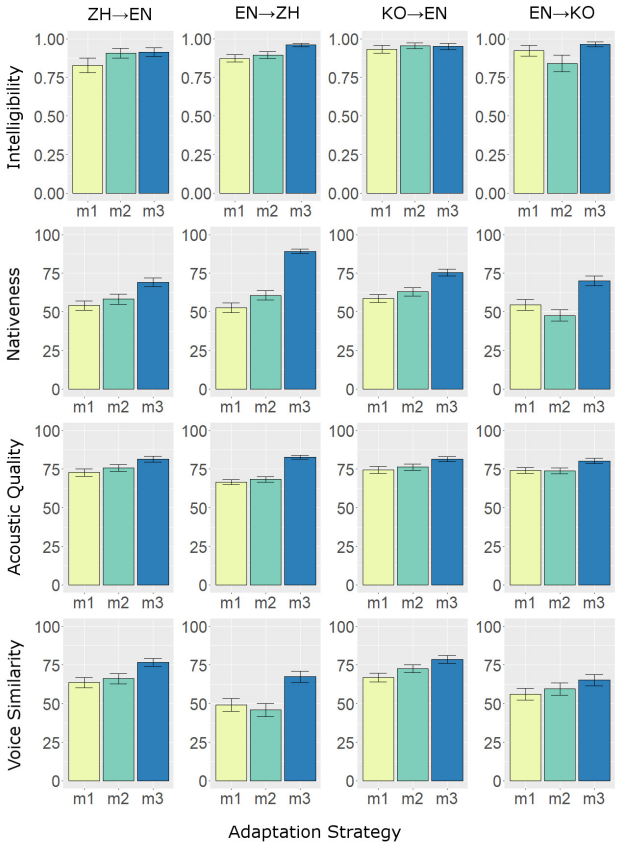

Fig. 2. Results of four subjective evaluations (intelligibility, nativeness, acoustic quality, and voice similarity) presented in each row using three sets of adaptation data (i.e., $model_1$ ($m1$), $model_2$ ($m2$), and $model_3$ ($m3$)) for fine-tuning the base multilingual model. In the figure, different language pairs are presented in each column. The error bars represent the 95% confidence interval (CI) of a mean.

5.1. Evaluation results

In the first test, we evaluated intelligibility of the samples by asking native speakers to transcribe them as accurately as possible. For each source-target language pair, 3 questionnaires were created; each consisted of 10 sentences pseudo-randomly drawn from the three models such that each participant would hear each of the 10 sentences only once. We calculated the word error rate of their transcription. Figure 2 (first row) shows that using target language data

[1] Samples are available in https://oben-ssw10.github.io/icassp2020/

Authorized licensed use limited to: University of Wollongong. Downloaded on May 31,2020 at 18:26:42 UTC from IEEE Xplore. Restrictions apply.

from another speaker can improve intelligibility. The improvement is statistically significant in ZH→EN (z=4.271, p<0.05), where the monolingual model ($m1$) produced less intelligible speech than its mixed-language counterpart ($m2$). The opposite trend is observed in EN→KO (z=-3.387, p<0.05), however. For the other two language pairs, the differences between $m1$ and $m2$ are not statistically significant (z's<1.523, p's>1.000).

In the second test, we evaluated how native-like the synthesizes speech were. Native speakers rated the audio samples using a number between 0 "not native at all" and 100 "completely native". From Figure 2 (second row), using the bilingual adaptation set ($m3$, i.e. the ground truth case) produced samples closest to native speech (scores between 69-89; z's>6.656, p's<0.05 in all pairwise comparisons). Nevertheless, the other two adaptation strategies resulted in somewhat native speech (scores between 48-67). Overall, using extra data from the target language ($m2$) produced speech samples that were less accented compared to the monolingual model ($m1$), despite the mismatch due to different prosody from the two speakers, except for one language pair (EN→KO: z=-4.327, p<0.05; the other three: z's>2.630, p's<0.05).

In the third test, we evaluated the acoustic quality of the TTS utterances. The samples were presented to native speakers in a MUSHRA-like questionnaire [30]. Participants were asked "How is the acoustic quality of the audio (e.g., clean/noisy, clear/muffled, any odd, non-speech sounds)?" and rated the audio samples between 0 "bad" and 100 "excellent". We found all models produce good acoustic quality (scores between 67-78), where using bilingual data yielded the best quality ($m3$; z's>5.705, p's<0.05 in all pairwise comparisons). Furthermore, using extra data from the target language ($m2$) improved acoustic quality compared to the monolingual model ($m1$) (EN→KO: z=-0.323, p=0.746; the other three: z's>2.003, p's<0.05).

In the last test, we evaluated whether the personalized models were able to generate speech in the target speaker's voice. It is important to note that cross-language adaptation models generate speech in languages different from the target speaker's, but it is not straightforward to compare voice across languages. This challenge was overcome by the use of bilingual target speakers for this study, which enabled us to use copy-synthesis speech in the target language as the reference. Participants rated the voice similarity of the TTS utterances with respect to the copy-synthesis speech on a scale between 0 "different person" and 100 "same person". We found that voice similarity can be improved with extra data from the target language. Comparing $m1$ and $m2$, the improvement is statistically significant in KO→EN and EN→KO (z's>2.233, p's<0.05). For the other two language pairs, the differences are not statistically significant (|z|'s<1.790, p's>0.073).

6. DISCUSSION

In this study, we propose the strategy to include speech from a native speaker of the target language during adaptation. Compared to the baseline method, this strategy can help generate more intelligible and native-like speech, with higher quality and in a language that the target speaker does not speak. However, we found that the language code had to be switched to that of the source speaker to maintain voice similarity. We believe this is due to the learned association between language and speaker embeddings. Language code was included in the model to account for allophones and other anomalies between languages that were not fully defined by linguistic features alone. Using source language code while generating target language eliminates this advantage of using language code. Despite this drawback, the results show that our strategy improves quality, intelligibility, and nativeness in most of the cases when compared to the baseline approach, although some differences were not statistically significant. The results also show small improvement in voice similarity between the TTS output and the target speaker. It was unexpected, and we regard this improvement as an effect of improved quality, which brings the TTS output closer to the reference (copy synthesis), not true improvement in timbre.

6.1. Generating speech in a novel language

We evaluated each speaker-adapted model only in one target language, because references were available only in that language. However, it is straightforward to generate target speaker's speech in other languages using the same models without any modification, even for novel languages that are not included in the base model. We designed the linguistic input features for training the multilingual acoustic model in a way that, except for a few language-specific features such as language code and lexical pitch patterns (e.g., ZH), most features such as voicing and places/manner of articulation were language agnostic. Thus, it is possible to build a text analyzer to convert text of a novel language into compatible linguistic features that allow us to generate speech in that language. In fact, we generated audios in Spanish, which was not included in our base model, using the cross-language adapted models from this study. Our internal evaluation reveals that the utterances generated in languages included in the base multi-lingual model (EN, ZH, KO, JA) are more intelligible and natural than the utterances in a novel language (Spanish). It shows the importance of including speech data from the target language in the base model for cross-language adaptation.

7. CONCLUSIONS

This paper investigates speaker adaptation for cross-language synthesis by constructing a multilingual model using speech data from multiple speakers in four languages. Using a feed-forward architecture, the model can exploit the common characteristics among different languages and speakers, and transfer the learned knowledge to a new speaker regardless of the speaker's language. We find speaker embeddings and language codes useful in maintaining speaker identity across languages when synthesizing languages different from the target speaker's language. Overall, using extra data from the target language spoken by other speakers improves cross-language adaptation performance. The performance improvements can be observed in all types of listening tests with most of the language pairs, compared to the model adapted using source monolingual data only. In future work, we plan to incorporate more languages and speakers into the existing system, and experiment with the use of similar speakers from multiple languages to improve the voice similarity of cross-language synthesis. Further, we plan to explore better representations of speaker characteristics to help synthesize cross-language speech that is more native-like, for example, by disentangling accent/style/timbre factors from speakers. We also plan to investigate methods to prevent the model from learning the association between language code and the speaker embedding present in the training data.

8. ACKNOWLEDGEMENTS

We would like to thank our colleagues Kyung-Min Kim, Sung Hah Hwang, and Ethan Sherr-Ziarko for their invaluable support in survey administration. Special thanks also go to Dr. Hsiao-Chi Ho and students at Graduate Institute of Education at Providence University, Taiwan for their help and participation in the surveys.

Authorized licensed use limited to: University of Wollongong. Downloaded on May 31,2020 at 18:26:42 UTC from IEEE Xplore. Restrictions apply.

9. REFERENCES

[1] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[2] H. Zen, N. Braunschweiler, S. Buchholz, et al., "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[3] A. van den Oord, S. Dieleman, H. Zen, et al., "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] S. Mehri, K. Kumar, I. Gulrajani, et al., "SampleRNN: An unconditional end-to-end neural audio generation model," *International Conference on Learning Representations*, 2017.

[5] J. Sotelo, S. Mehri, K. Kumar, et al., "Char2Wav: End-to-end speech synthesis," in *Proc. of International Conference on Learning Representations (ICLR)*, 2017.

[6] Y. Wang, RJ Skerry-Ryan, D. Stanton, et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. of Interspeech*, 2017, pp. 4006–4010.

[7] S. O. Arik, M. Chrzanowski, A. Coates, et al., "Deep voice: Real-time neural text-to-speech," in *Proc. of the 34th International Conference on Machine Learning*, 2017, pp. 195–204.

[8] W. Ping, K. Peng, A. Gibiansky, et al., "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. of International Conference on Learning Representations (ICLR)*, 2018.

[9] Y. Jia, Y. Zhang, R. J. Weiss, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, 2018, pp. 4480–4490.

[10] J. Williams and S. King, "Disentangling Style Factors from Speaker Representations," in *Proc. of Interspeech*, 2019, pp. 3945–3949.

[11] Y. Zhang, R. J. Weiss, H. Zen, et al., "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," in *Proc. of Interspeech*, 2019, pp. 2080–2084.

[12] J. Latorre, K. Iwano, and S. Furui, "New approach to polyglot synthesis: How to speak any language with anyone's voice," in *Proc. of ITRW on Multilingual Speech and Language Processing*, 2006.

[13] E. Nachmani and L. Wolf, "Unsupervised polyglot text-to-speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 7055–7059.

[14] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 4905–4909.

[15] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis.," in *Proc. of Interspeech*, 2016, pp. 2468–2472.

[16] H. Ming, Y. Lu, Z. Zhang, and M. Dong, "A light-weight method of building an LSTM-RNN-based bilingual TTS system," in *International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 201–205.

[17] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5540–5544.

[18] C.-P. Chen, Y.-C. Huang, C.-H. Wu, and K.-D. Lee, "Polyglot speech synthesis based on cross-lingual frame selection using auditory and articulatory features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1558–1570, Oct 2014.

[19] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (mandarin–english) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009.

[20] H. Yang, K. Oura, H. Wang, et al., "Using speaker adaptive training to realize mandarin-tibetan cross-lingual speech synthesis," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9927–9942, 2015.

[21] Y. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *6th International Symposium on Chinese Spoken Language Processing*, 2008, pp. 1–4.

[22] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2015, pp. 4475–4479.

[23] H. T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 4905–4909.

[24] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE Transactions on Information and Systems*, vol. 101, no. 2, pp. 462–472, 2018.

[25] S. O. Arik, J. Chen, K. Peng, et al., "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, et al., Eds., pp. 10019–10029. 2018.

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

[27] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in DNN-based speech synthesis using d-vectors.," in *Proc. of Interspeech*, 2017, pp. 3404–3408.

[28] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *International Conference on Learning Representations*, 2018.

[29] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system.," in *Proc. of the 9th ISCA Speech Synthesis Workshop*, 2016, pp. 202–207.

[30] ITU-R Rec, "Bs. 1534-1 method for the subjective assessment of intermediate quality level of coding systems," Tech. Rep., International Telecommunications Union, 2003.

Authorized licensed use limited to: University of Wollongong. Downloaded on May 31,2020 at 18:26:42 UTC from IEEE Xplore. Restrictions apply.