END-TO-END CODE-SWITCHING TTS WITH CROSS-LINGUAL LANGUAGE MODEL

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, Haizhou Li

Department of Electrical and Computer Engineering, National University of Singapore, Singapore

{xuehao.zhou, grandee.lee}@u.nus.edu, {eletia, rohankd, haizhou.li}@nus.edu.sg

ABSTRACT

Code-switching text-to-speech (TTS) aims to enable a system to speak two languages with a single voice and in the same utterance. In this paper, we propose to incorporate cross-lingual word embedding into an end-to-end TTS system, to improve the voice rendering. The cross-lingual word embedding, generated from a pre-trained cross-lingual language model, is able to encode words of two languages in the same embedding space, therefore, allows words across languages to share each other's contextual information, which is useful for the voice rendering of code-switching content. To investigate the effectiveness of this idea, we conduct studies on two multi-speaker monolingual corpora, namely, THCHS30 Mandarin and LibriTTS English database. The evaluation results show that our proposed framework outperforms the baseline systems when presented with code-switching text input, and achieves state-of-the-art performance.

Index Terms— text-to-speech, code-switching, cross-lingual word embedding, end-to-end

1. INTRODUCTION

Code-switching (CS) refers to the process of switching the linguistic code from one to another, which can occur between two sentences (i.e. inter-sentential) or within one sentence (i.e. intra-sentential). To be truly multilingual, a text-to-speech (TTS) system is expected to be capable of speaking such code-switching content as naturally as monolingual content. The prior work on code-switching speech synthesis can be grouped into three categories: unit mapping, multilingual synthesis, and polyglot synthesis.

The unit mapping approach substitutes linguistic units of one language with the equivalences of another language by using frame mapping [1], state mapping [2], and phone mapping [3, 4], etc. With the unit mapping method, the generated voice may suffer from a strong foreign accent.

A multilingual synthesis system is built on multiple language-dependent systems, which shares the common unit selection module [5], or multilingual synthesizer [6] across languages. Although such method works well for inter-sentential CS content, it doesn't maintain the same voice identity across languages unless language-dependent systems are built on a voice database recorded by the same speaker.

Polyglot TTS systems are capable of synthesizing speech of different languages with the same speaker's voice [7], where one builds a multilingual single speaker diphone speech unit inventory using a phoneset combination approach. In [8], a polyglot average voice is trained with multi-speaker monolingual speech corpora, which can then be adapted to any speaker's voice in one of the training languages. The adapted voice retains the voice identity across different languages. Voice conversion techniques can be also adopted for controlling speaker's voice identity [9].

It is noted that end-to-end (E2E) TTS architecture has achieved the state-of-the-art speech quality, where the joint training mechanism alleviates the need of complex linguistic feature engineering. In the context of CS TTS [10], two kinds of encoders are explored to handle alphabetic inputs from different languages: shared multilingual encoder with explicit language embedding and separated monolingual encoder, both of which produce more natural CS utterances than Tacotron [11]. In [12], Xue et al. present a robust Mandarin-English mixed-lingual TTS system with only monolingual data by exploring speaker embedding and phonetic representations. High quality CS voice is also presented in [13]. In short, E2E TTS represents one of the successful implementations for CS speech content.

A recent study on semi-supervised training [14] of Tacotron benefits from textual and acoustic knowledge obtained from large, publicly available text and speech corpora. In [15], Hayashi et al. show that, the text embeddings computed from a pre-trained BERT model, help TTS systems improve naturalness of generated speech. Motivated by this finding, in this paper, we propose to incorporate cross-lingual word embedding, computed from a pre-trained cross-lingual language model, into Tacotron2-based [16] E2E TTS architecture. As the embedding vectors carry contextual knowledge of the words, syntactically similar words of both languages are able to share each other's contextual information. It is expected that the cross-lingual word vector will help to generate smoother voice when switching between different languages.

This project Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) is supported by the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme. The adaptation and inference data are provided by Data-baker.

2. CODE-SWITCHING TTS

The CS TTS system is expected to generate one homogeneous, high quality voice between two languages. However, it is not easy to find a multilingual, CS speech corpus recorded from a single speaker. As a compromise, the average voice model (AVM) approach is adopted with multi-speaker monolingual speech corpora. There are two common techniques to control the speaker consistency of the AVM, that are speaker embedding and speaker adaptation. Speaker embedding is widely used in multi-speaker TTS [17, 18, 19] by characterizing a speaker with a low dimensional vector. Speaker adaptation is achieved usually by re-training the AVM model with target speaker data.

It has been shown [20] that the combination of these two techniques outperforms the individual techniques. In this paper, we perform adaptation on the entire AVM model that is conditioned on the speaker embedding to benefit from the two techniques. In practice, we use i-vector, extracted from a pre-trained network, as the speaker embedding. Specifically, we combine encoder output with speaker embedding by concatenating the two vectors, in a similar way as [21]. Next, we introduce two CS TTS baseline systems, that represent state-of-the-art performance. We benchmark the performance of our proposed idea against the baselines in the experiments.

2.1. Tacotron2-based approach

We use Tacotron2 [16], as one of our baselines. It consists of an encoder and an attention-based decoder. The phone sequences are taken as the model input. We use Griffin-Lim algorithm [22] to reconstruct the waveform instead of WaveNet vocoder [23], for rapid turn-around, as shown in Fig. 1.

The encoder generates text representations from the input sequence. The phone embedding is taken by three convolutional layers, followed by a bi-directional long short-term memory (BLSTM). The encoder output will be attended by decoder at each decoder time step via a location sensitive attention network, to compute a fixed-length context vector.

The decoder is an autoregressive recurrent neural network, predicting mel spectrograms from encoder output. It is composed of a two layer pre-net, two uni-directional LSTMs, a linear projection layer and a five convolutional layer post-net with residual connection. We use CBHG post-prossing net, presented in [11], to predict linear-scale spectrogram from generated mel-scale spectrogram. Additionally, we implement the guided-attention loss [24], leading to faster alignment results and lower training loss.

2.2. Residual encoder

In [12], Xue et al. presented a residual encoder structure, as shown in Fig. 2. It is a revised version of the encoder in Fig. 1 by adding the encoder input directly into encoder output. The rest of the network is the same as that in Fig. 1.

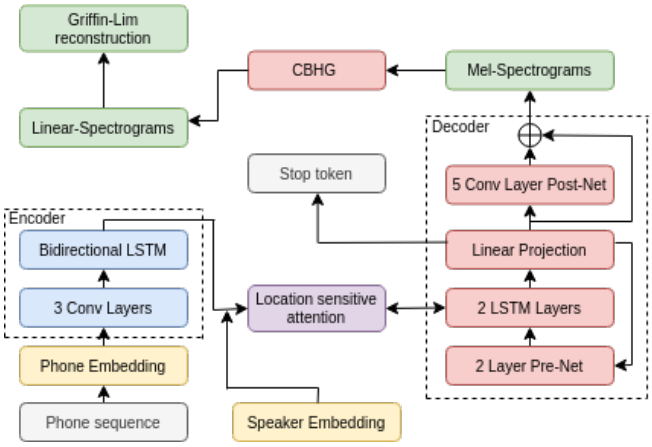

Fig. 1. The Tacotron2-based system with Grifflin-Lim waveform reconstruction.

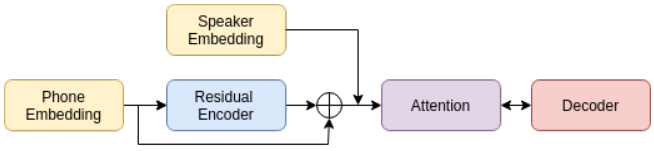

Fig. 2. The architecture of residual encoder

We expect that more phoneme information is retained through the residual encoder, thus, having a direct impact on the whole attention mechanism. As this structure helps generate more natural mixed-lingual speech [12], we adopt it as the second competitive baseline.

3. CODE-SWITCHING TTS WITH CROSS-LINGUAL LANGUAGE MODEL

In [15], it was shown that Tacotron2-based TTS model benefits from pre-trained BERT text embeddings in monolingual case. The text embedding provides the useful information that represents segmental information of speech, such as semantics group of the phrase. Therefore, such additional textual knowledge help improve the naturalness of the speech rendering.

Next, we extend the above idea to the CS TTS system by introducing cross-lingual language model.

3.1. Cross-lingual language model

A language model predicts the next word based on the previous context, while a word token is represented with an embedding. Embeddings derived from fastText [25] encodes the monolingual syntactic and contextual information, while VecMap [26] is able to take in two monolingual embeddings, X, Z, and projects them into a common embedding space, to establish the cross-lingual correspondence. X, Z are mapped according to the mapping functions W_X, W_Z and a learned

dictionary D that is based on cosine distance of the two embedding space. The mapping function are derived to optimize the following equation,

$$W_X, W_Z = \arg \max_{W_X, W_Z} \sum_i \sum_j D_{ij}((X_i W_X) \cdot (Z_j W_Z))$$

We use this embedding to initialize our cross-lingual language model (CLLM) as in [27, 28]. To take care of the domain mismatch, CLLM is trained with the synthetically generated CS corpus to fine-tune the common embedding space [29].

The synthetic CS corpus is augmented using a sentence-level parallel corpus [30] according to the Matrix Frame Language theory [31] whereby the two language will mix in a way that preserves the syntactic structure of the dominant language. The parallel corpus is firstly word-aligned and subsequently phrase-aligned based on the phrase table of commonly occurring phrases. The synthetic CS corpus is sampled from the phrase-aligned text based on a switching probability.

3.2. Encoder with cross-lingual word embedding

The baseline systems, in Section 2, are trained with the parallel monolingual $< text, audio >$ pairs, that doesn't provide code-switching knowledge. Similar to [15] where word embedding is used for TTS, we propose to incorporate cross-lingual word embedding (CLWE) into our residual encoder system. We hypothesize that the phonotactic and prosodic transition between words across languages should follow those within the same language. CLWE allows the corresponding words across two languages to share each other's contextual information, which allows us to implement the idea and validate the hypothesize.

In our work, we augment the encoder network with pre-trained CLWE, since encoder is used to extract text sequential feature representations from the input sequence. As shown in [12], the encoder output has a greater impact on waveform generation than encoder input. Therefore, to maximize the effect of CLLM, we combine the residual encoder output with CLWE by concatenating the embeddings as shown in Fig. 3. Specifically, we concatenate a CLWE to all encoded phone embedding that belong to the same word.

4. EXPERIMENTS

4.1. Experimental setup

We choose 40 Mandarin speakers from THCH30 [32] and 110 English speakers from LibriTTS [33] for average model training. Each Mandarin speaker contributes 200 to 240 utterances, totalling 9,054 utterances with 22 hours of audio.

Each English speaker contributes 50 to 150 utterances, totalling 8,962 utterances with 17 hours of audio. Another 300 Mandarin utterances, 300 English utterances and 300 CS utterances are used during inference time. The matrix language (main langauge) of CS sentences is Mandarin.

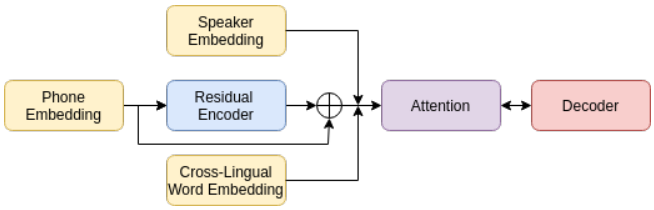

Fig. 3. The architecture of residual encoder with cross-lingual word embedding.

All audios are down-sampled to 16kHz for average model training. We use Mandarin front-end and grapheme-to-phoneme to convert character sequences to phone sequences for Mandarin and English, respectively. The model outputs, log-magnitude linear-scale spectrogram and 80-dim mel-scale spectrogram, are computed from 50ms Hanning window, 12.5ms frame shift and 1024-point Fourier transform.

4.2. System implementation

We use Tacotron2 (T2) and residual encoder structure (RES-ENC) as our baseline systems to investigate the cross-lingual word embedding conditioned on residual encoder structure system (CLWE). T2 system follows the implementations in [16]. For the RES-ENC system, the character-level language identity (LID), determined via orthography-based method, is implemented to model the language difference. We simply concatenate one-hot language vector obtained from LID to all phoneme embedding of the same language.

The CLLM is a two layer LSTM with 650 hidden units and a drop-out rate of 0.3 in-between layers, following the implementation details as outlined in [29]. The 650-dimensional CLWE is extracted from the embedding layer of the CLLM.

The i-vector based speaker embedding is obtained through factor analysis [34]. An i-vector extractor of 400 speaker factors is learned on Switchboard II Corpus to derive the i-vectors [35]. Further, we apply linear discriminant analysis to obtain a 150-dimensional i-vector for each speaker.

4.3. Subjective evaluation

To measure the speech quality, we conduct subjective evaluation on MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) experiments [36]. Ten listeners proficient in both Mandarin and English are invited to each set of the listening tests. We randomly choose twenty utterances from the 300 test utterances for each experiment group.

We implement three types of adaptation, using 200 Mandarin (CH) utterances, 200 English (EN) utterances, 200 Mandarin and 100 English (MIX) utterances, for each of the three systems, respectively. The same optimizing step is taken for all adaptation experiments. All speech corpora used for adaptation and inference are recorded from a unseen

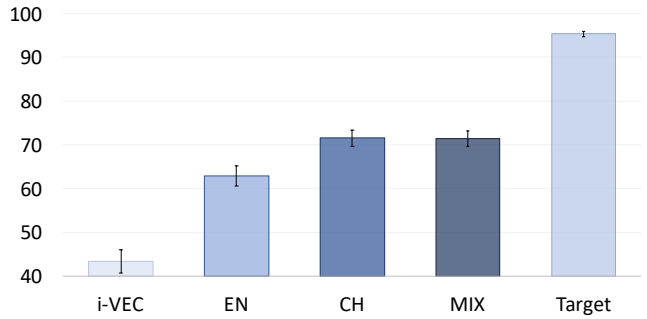

Fig. 4. MUSHRA results of CLWE system (i-VEC) and its English (EN), Mandarin (CH), English and Mandarin (MIX) adapted versions with CS utterance

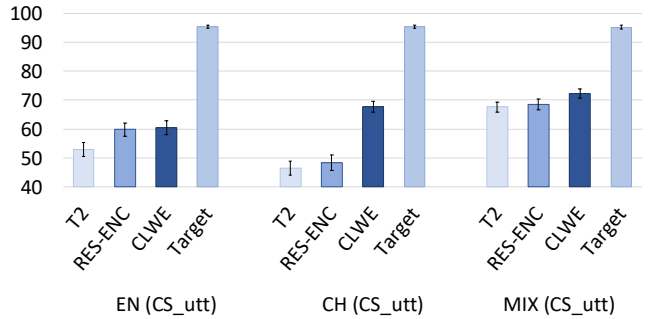

Fig. 5. MUSHRA results of T2, RES-ENC, CLWE systems for English (EN), Mandarin (CN), English and Mandarin (MIX) adapted versions with CS utterance

multilingual single speaker[1]. We investigate the effectiveness of the proposed CLWE system in two experiments. We first compare the CLWE system performance with and without adaptation; we then compare the system performance with and without the cross-lingual word vectors.

4.3.1. CLWE system with and without adaptation

Firstly, we compare the performance of the proposed CLWE system and its adapted versions with target speech for CS input. The MUSHRA results in Fig. 4 show that adapting the CLWE system with Mandarin speech corpora achieves better performance than that with English database, since Mandarin is the matrix language in the CS content. Furthermore, our CLWE system adapted with only Mandarin speech data can achieve almost the same performance as the one adapted with both Mandarin and English speech corpora. Note that all adapted CLWE systems outperform that without adaptation, denoted as i-VEC system.

4.3.2. System performance with and without cross-lingual word vector

We then explore the effectiveness of cross-lingual word vectors on both CS and monolingual cases. To investigate the effectiveness of cross-lingual word vectors on CS input, we compare system performance of three systems in every different adaptation case. The MUSHRA results of synthesizing CS speech in Fig. 5 demonstrate that the CLWE system consistently outperforms RES-ENC, followed by T2. We note that there is a small gap between CLWE and RES-ENC in EN adapted case. As the matrix language in the CS content is Mandarin, it is easy to understand that adaptation with English speech data doesn't significantly improve system performance.

To synthesize monolingual utterance, we generate Mandarin speech (CHutt) from three systems in CH adaptation case, and English speech (ENutt) from three systems in EN adaptation case. As described in Fig. 6, the CLWE system achieves better performance than baseline systems. The results suggest that cross-lingual word vector is effective not only for CS content but also for monolingual input.

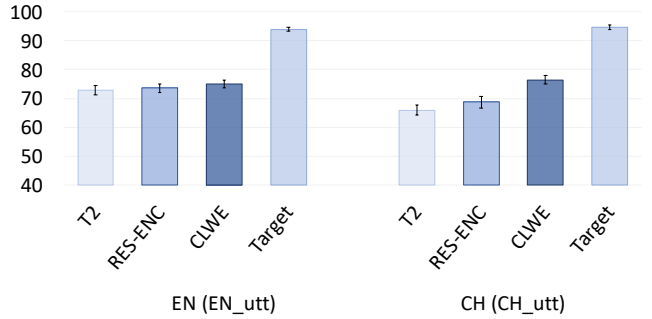

Fig. 6. MUSHRA results of T2, RES-ENC, CLWE systems for English (EN) and Mandarin (CH) adapted versions with monolingual utterance

All the samples can be found in this demo link[2].

5. CONCLUSION

In this paper, we propose to incorporate cross-lingual word embedding into Tacotron2-based TTS system, to achieve better performance on CS content. This embedding is generated from a pre-trained CLLM, enriching cross-lingual contextual knowledge. The conducted studies show that the cross-lingual word embedding contributes to improve quality and smoothness of generated voice when switching between two languages. Our evaluation results verify the effectiveness of the cross-lingual word embedding not only on CS text input, but also in monolingual cases. However, we find that the performance of proposed CS TTS system without adaptation is not good as expected. In the future, we are interested to further understand the influence of speaker embedding on system performance with CS input.

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

[2] https://xuehao-marker.github.io/icassp2020/

6. REFERENCES

[1] Ji He, Yao Qian, Frank K Soong, and Sheng Zhao, "Turning a monolingual speaker into multilingual for a mixed-language tts," in *INTERSPEECH*, 2012.

[2] Yao Qian, Hui Liang, and Frank 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.

[3] Sunayana Sitaram and Alan W Black, "Speech synthesis of code-mixed text," in *LREC*, 2016, pp. 3422–3428.

[4] Sunayana Sitaram, Sai Krishna Rallabandi, Shruti Rijhwani, and Alan W Black, "Experiments with cross-lingual systems for synthesis of code-mixed text.," in *SSW*, 2016, pp. 76–81.

[5] Min Chu, Hu Peng, Yong Zhao, Zhengyu Niu, and Eric Chang, "Microsoft mulan-a bilingual tts system," in *IEEE ICASSP*, 2003, vol. 1, pp. I–I.

[6] Khyathi Raghavi Chandu, Sai Krishna Rallabandi, Sunayana Sitaram, and Alan W Black, "Speech synthesis for mixed-language navigation instructions.," in *INTERSPEECH*, 2017, pp. 57–61.

[7] Christof Traber, Karl Huber, Karim Nedir, Beat Pfister, Eric Keller, and Brigitte Zellner, "From multilingual to polyglot speech synthesis," in *EUROSPEECH*, 1999.

[8] Javier Latorre, Koji Iwano, and Sadaoki Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *IEEE ICASSP*, 2005, vol. 1, pp. I–1.

[9] B Ramani, MP Actlin Jeeva, P Vijayalakshmi, and T Nagarajan, "Voice conversion-based multilingual to polyglot speech synthesizer for indian languages," in *IEEE TENCON*, 2013, pp. 1–4.

[10] Yuewen Cao, Xixin Wu, Songxiang Liu, Jianwei Yu, Xu Li, Zhiyong Wu, Xunying Liu, and Helen Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *IEEE ICASSP*, 2019, pp. 6935–6939.

[11] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv:1703.10135*, 2017.

[12] Liumeng Xue, Wei Song, Guanghui Xu, Lei Xie, and Zhizheng Wu, "Building a mixed-lingual neural tts system with only monolingual data," *arXiv:1904.06063*, 2019.

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

[14] Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *IEEE ICASSP*, 2019, pp. 6940–6944.

[15] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, and Karen Livescu, "Pre-trained text embeddings for enhanced text-to-speech synthesis," in *INTERSPEECH*, 2019, pp. 4430–4434.

[16] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE ICASSP*, 2018, pp. 4779–4783.

[17] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *NIPS*, 2017, pp. 2962–2970.

[18] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural voice cloning with a few samples," in *NIPS*, 2018, pp. 10019–10029.

[19] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf, "Fitting new speakers based on a short untranscribed sample," *arXiv:1802.06984*, 2018.

[20] Andrew Senior and Ignacio Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *IEEE ICASSP*, 2014, pp. 225–229.

[21] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NIPS*, 2018, pp. 4480–4490.

[22] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[24] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *IEEE ICASSP*, 2018, pp. 4784–4788.

[25] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[26] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," *arXiv:1805.06297*, 2018.

[27] Grandee Lee and Haizhou Li, "Word and class common space embedding for code-switch language modelling," in *IEEE ICASSP*, 2019, pp. 6086–6090.

[28] Xianghu Yue, Grandee Lee, Emre Yılmaz, Fang Deng, and Haizhou Li, "End-to-end code-switching asr for low-resourced language pairs," *arXiv:1909.12681*, 2019.

[29] Grandee Lee, Xianghu Yue, and Haizhou Li, "Linguistically motivated parallel data augmentation for code-switch language modeling," in *INTERSPEECH*, 2019, pp. 3730–3734.

[30] Pierre Lison and Jörg Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[31] Carol Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*, Oxford University Press, 1997.

[32] Dong Wang and Xuewei Zhang, "Thchs-30: A free chinese speech corpus," *arXiv:1512.01882*, 2015.

[33] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv:1904.02882*, 2019.

[34] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[35] Rohan Kumar Das and SR Mahadeva Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.

[36] ITUR Recommendation, "1534-1,"method for the subjective assessment of intermediate sound quality (mushra)"," *ITU, Geneva, Switzerland*, 2001.