

UNSUPERVISED POLYGLOT TEXT-TO-SPEECH

Eliya Nachmani^{1,2}, Lior Wolf^{1,2}

¹Facebook AI Research, ²Tel-Aviv University

ABSTRACT

We present a TTS neural network that is able to produce speech in multiple languages. The proposed network is able to transfer a voice, which was presented as a sample in a source language, into one of several target languages. Training is done without using matching or parallel data, i.e., without samples of the same speaker in multiple languages, making the method much more applicable. The conversion is based on learning a polyglot network that has multiple per-language sub-networks and adding loss terms that preserve the speaker’s identity in multiple languages. We evaluate the proposed polyglot neural network for three languages with a total of more than 400 speakers and demonstrate convincing conversion capabilities.

Index Terms—TTS, multilingual, unsupervised learning

1. INTRODUCTION

Neural text to speech (TTS) is an emerging technology that is becoming dominant over the alternative TTS technologies, in both quality and flexibility. On the quality front, multiple parallel lines of work, provide, in addition to increasing capabilities over time, an improved Mean Opinion Scores (MOS) from one work to the next: (i) From WaveNet [1] to Parallel WaveNet [2], and its alternative [3], (ii) from DeepVoice [4] to DeepVoice2 [5] and DeepVoice3 [6], (iii) from Tacotron [7] to Tacotron2 [8], and (iv) from VoiceLoop [9] to its recent feed-forward-fitting successor [10].

There is a parallel effort to improve the ease in which new voices are sampled. The first systems were able to reproduce a single voice, as captured from long high-quality recordings. Subsequent systems provided multi-speaker capabilities [9, 11, 12, 13], then multi-speakers sampled from challenging sources, such as audiobooks, and even multi-speakers sampled “in the wild” in noisy and uncontrolled recordings [10].

Our work continues this line of evolution and takes the voice sampling capabilities a step further. We would like to take a sample of a speaker talking in one language and have her voice-avatar speak, as a native speaker in another language. This direction is worth pursuing for several reasons. First, one of the most attractive applications of TTS is automatic translation of spoken words. This translation is most authentic, when maintaining the voice identity of the original

speaker. Second, it is a challenging AI problem that requires the disentanglement of voice characteristics from spoken language characteristics.

The proposed method is based on training a single TTS engine, with many components that are shared between the languages. In addition, there are two types of language-specific components. The first is a per-language encoder, that embeds the input sequences of phonemes in a language-independent vector space. The second component is a network that, given a sample of the speaker’s voice, encodes it in a shared voice-embedding space. In other words, while there is one network per dataset that extracts the speaker’s embedding from a speech sample, this embedding space is shared for all languages. This universality of the embedding is enforced by a novel loss term that preserves the speaker’s identity under language conversion. The results show convincing conversions between English, Spanish, and German.

2. RELATED WORK

The recent neural TTS systems are sequence to sequence methods, where the underlying methods differ. Wavenet [1] employs CNNs with dilated convolutions. Char2Wav [14] employs RNNs, the original Tacotron method contains multiple RNNs, convolutions and a highway network [15]. The subsequent Tacotron2 [8] method replaced the highway networks with RNNs and directly predicts the residuals. Deep Voice (DV) [4] and DV2 [5] employ bidirectional RNNs, multilayer fully connected networks and residual connections. DV3 [6] switched the architecture to a gated convolutional sequence to sequence architecture [16] and also incorporated the key-value attention mechanism of [17]. The VoiceLoop [9] method is based on a specific type of RNN, in which a shifting buffer is used to maintain the context.

Out of this list, the multi-speaker systems are DV2, DV3, the modified Tacotron by the authors of DV2, and the two versions of VoiceLoop. Chat2Wav has a public multi-speaker implementation, with an added speaker embedding.

The input text is represented by one of a few possible alternative types of features. Tacotron, Tacotron2 and DV3 employ the input letters. Phonemes are used by Char2Wav, VoiceLoop, and some of the DV3 experiments. Learned linguistic features, including duration and pitch of phonemes, are used by WaveNet, Parallel WaveNet, DV1, and DV2. Our

work, which requires the modeling of multiple languages, relies on phonemes. Phoneme dictionaries are available for many languages. While we employ the same phonemes for English, Spanish and German, the method is general enough to support incompatible phoneme representations.

The best audio quality is produced by methods that employ WaveNet decoders in order to generate wave forms, including DV1, DV2, DV3, Tacotron2. Char2Wav and VoiceLoop employ World Vocoder Features [18] in order to represent the output, which has to go through a separate decoding process. Our work also produces these vocoder features, which allows us to efficiently train on multiple corpora, even if these are insufficient for training a WaveNet decoder.

The literature has a few examples of polyglot TTS systems. These systems are mostly concatenative or HMM-based. The approach given in [19] is a concatenative method that is based on unit selection. Polyglot synthesis is achieved by mapping the phonemes from the source language into the target language. An earlier system that uses matching samples of a single speaker [20] uses a concatenative system that is applied to a recording of a single speaker speaking in four languages. Polyglot TTS based on HMM, has been proposed in various papers [21, 22, 23].

Recently, neural multilingual TTS was proposed in [24]. The method uses an LSTM architecture and is trained on a single speaker, who is recorded speaking two languages in the training data. Another related LSTM-based work [25] is trained on six different languages, with eight speakers in total. While the model is multi-language, it is not Polyglot – every speaker speaks the language of the training set.

In [26], a neural multilingual TTS is proposed. A deep neural network is trained on samples from a pair of languages – English and Mandarin. Three speakers are recorded, each speaking both languages. In the last section of that paper, a neural polyglot TTS is discussed and demonstrated in a specific setting: the Mandarin samples of one of the speakers were removed from the training set, and a polyglot synthesis in Mandarin of this speaker was generated. Our work does not employ shared speakers among the corpora used for the different languages.

3. POLYGLOT NEURAL SYNTHESIS

The technique we employ in order to achieve a multi-language network, transcends any specific TTS architecture. We recommend the following modifications of the existing systems: (i) replacing the speaker embedding with a learned network, as was done in [10], in order to verify via a loss that the output audio is compatible with the speaker’s embedding, (ii) using multiple such speaker embedding networks, one per language, (iii) using multiple language (e.g., text or phoneme) embeddings, and (iv) using a loss term for speaker preserving. At test time, conversion is performed by a mixed input that consists of a sample of the speaker’s voice speaking in the source language and a text in the target language.

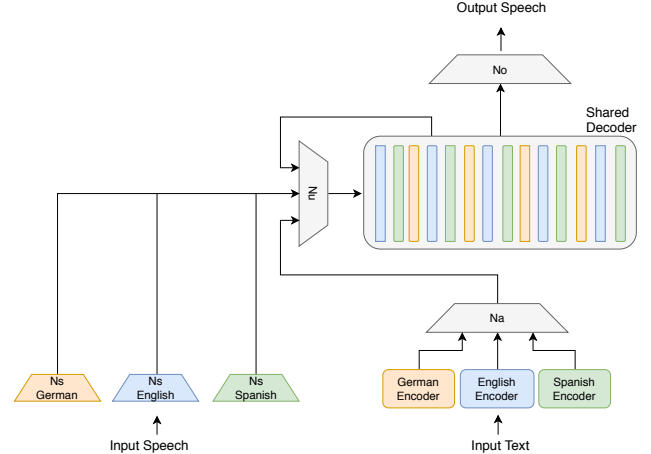


Fig. 1: The polyglot neural Text To Speech architecture.

3.1. The Architecture

The specific implementation of our method is based on the VoiceLoop architecture [9], which is a neural multi-speaker Text To Speech (TTS) system. There is no specific reason to use this system, except that it was shown to outperform Char2Wav [14], which is the only other multi-speaker TTS system, whose open code is currently available.

The VoiceLoop is a biologically motivated architecture, which employs a “phonological” buffer S_t , which combines both auditory and lingual information. The model contains multiple networks (N_u , N_a , N_o), whose input includes this buffer. N_u creates the new vector u_t that is pushed to the buffer at time t , N_a is an attention network that produces weights over the input text, and N_o generates the next audio frame, o_t . At every time step, u_t is inserted into the buffer, and another vector is discarded in a FIFO manner.

The model also employs two LUTs, LUT_p and LUT_s . The first LUT is the embedding of the input phonemes, one vector per each possible phoneme. LUT_s stores a different embedding vector z_s to each speaker encountered during training.

We modify the architecture of [10], by training a separate fitting network N_s and a separate language embedding LUT_p for each language. In order to achieve polyglot synthesis, we enforce a shared representation for all of the languages, and use only one copy of the attention network N_a , the buffer update N_u and the output network N_o (see Fig. 1).

The architecture of the various networks follows [10]. The N_s network has five conventional layers with ReLU activation. Each layer has 3×3 filters, with 32 channels. Average pooling over time is performed after the conventional layers. The N_s network ends with two fully connected layers with ReLU activation (each layer has size of 256). The networks N_o , N_u and N_a have the same architecture: two fully connected layers with ReLU activations.

3.2. The Loss Terms

We propose a new loss function to achieve polyglot synthesis. Given voice sample \mathbf{y}^a of the source language a , and sequence of phoneme \mathbf{s}^b from the target language b . We would like to preserve the identity of the speaker in the voice sample created at the target language i.e., the embedding of the original speaker \mathbf{z}_s^a should be the same as the embedding of the polyglot speech \mathbf{z}_s^b . An $L1$ loss term is used:

$$L_{poly} = \sum_a \sum_{b \neq a} \sum_{\mathbf{y}^a, \mathbf{s}^b} \|\mathbf{z}_s^a - \mathbf{z}_s^b\|$$

$$\mathbf{z}_s^a = N_s^a(\mathbf{y}^a), \mathbf{z}_s^b = N_s^a(\mathbf{o}^b), \mathbf{o}^b = G(LUT_p^b(\mathbf{s}^b), \mathbf{z}_s^a)$$

where $G(LUT_p^b(\mathbf{s}^b), \mathbf{z}_s)$ is the proposed polyglot network (a mapping between text in one language and voice sample in another to a new voice sample) N_s^a is the sample embedding network of the source language a , and LUT_p^b is the embedding of the text of the target language. Since we do not have matching parallel data, for this loss, unlike the training losses of VoiceLoop and VoiceLoop2, the network G generates speech without teacher forcing. Note that for each speaker in the source language, we generate samples in the other languages, i.e. for each English speaker, we generate the converted voice in Spanish and German.

In addition to the loss above, we use the loss terms of VoiceLoop2 [10]: (i) Reconstruction loss, given an audio sample \mathbf{y} , $L_{MSE} = \frac{1}{d_o} \sum_{\mathbf{y}} \sum_{t=1}^l \|\mathbf{y}_t - \mathbf{o}_t\|^2$, where the ground truth at time t is \mathbf{y}_t and the network's output $G(LUT_p^a(\mathbf{s}), N_s^a(\mathbf{y}))$ at time t is \mathbf{o}_t , \mathbf{s} being the phoneme sequence of \mathbf{y} (both \mathbf{y}_t and \mathbf{o}_t are vectors in space of world vocoder features \mathbb{R}^{d_o}). (ii) A contrastive loss term that compares three samples from the same language a : $L_{contrast} = \frac{1}{2} \sum_{\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3} (\|N_s^a(\mathbf{y}^1) - N_s^a(\mathbf{y}^2)\|^2 + \max(0, \Delta - \|N_s^a(\mathbf{y}^2) - N_s^a(\mathbf{y}^3)\|)^2)$, where Δ is the margin, and $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3$ are voice samples, such that $\mathbf{y}^1, \mathbf{y}^2$ are from the same speaker, and \mathbf{y}^3 is a sample of another speaker of the same language. (iii) A cycle loss $L_{cycle} = \sum_{\mathbf{y}} \|N_s^a(\mathbf{y}) - N_s^a(\mathbf{o})\|^2$, where \mathbf{y} is the input audio sample of language a , and \mathbf{o} is the output audio sample which is generated with $N_s^a(\mathbf{y})$ as embedding.

The overall loss is $L = L_{MSE} + \alpha L_{contrast} + \beta L_{cycle} + \gamma L_{poly}$, where we set $\alpha = \beta = 10$, and $\gamma = 1000$

3.3. Training and Inference Details

The L_{poly} term is only effective when the network G is able to generate speech in each one of the trained languages. Since it does not use teacher forcing, the signal it uses would diverge quickly for networks that are insufficiently trained. Therefore, the proposed network is trained with three phases, where the first two phases train the network to synthesize multilingual speech, and the third phase is for optimizing the speaker embedding space to enable convincing polyglot synthesis. Specifically, the first two phases follow the procedure, as described in [10], except that the batches contain examples

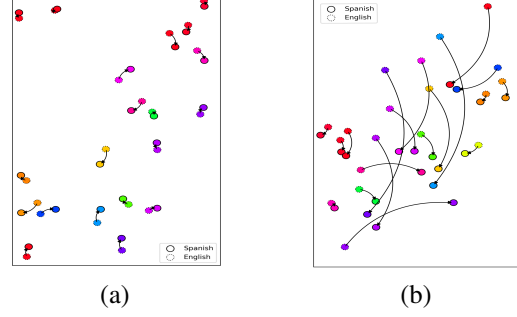


Fig. 2: TSNE visualization of the learned embedding for VCTK speakers in English and Spanish with (a) and without (b) the loss term L_{poly} . See text for details.

from multiple languages. The first phase uses white noise augmentation with standard deviation (SD) equal to 4.0 and sequences of ground truth vocoder features of length 100. The second (and the third) phases use SD of 2.0 and sequence length of 1000 vocoder features. The phases are run until convergence. The entire architecture is trained end-to-end on matching samples of inputs and outputs, where the input is the speaker embedding and the spoken text and the output is the recorded audio. The minimization of the MSE loss of the output frame, which is given as vocoder features, relies on the generated output to be perfectly aligned with the target, and is sensitive to drift. Teacher-forcing is, therefore, used, i.e., during training, the network receives the ground-truth target of the previous frame, in lieu of the predicted output from that frame. The third phase of training adds the L_{poly} term to the overall loss. Since the purpose of the third phase is to optimize the speaker embedding space to achieve polyglot synthesis, during this phase we only update the N_s networks. At inference, polyglot synthesis is obtained by using N_s^a of the source language a , and LUT_p^b for the target language b . For example, we can synthesize an English speaker in the Spanish language, by using the speaker embedding from English N_s^a and use as input a sequence of Spanish phonemes, as encoded by the Spanish LUT_p^b . Note that the speaker does not need to be from the training set and samples from unseen speakers can be used for synthesis. Note also, that the transcript of the voice sample is not needed. However, the information on the language being spoken in that sample is required.

4. EXPERIMENTS

We focus on three large multi-speaker datasets. For the English language, we use the VCTK dataset [27], which contain 109 speakers. For the Spanish language, we use the DIMEx100 dataset [28], which contain 100 speakers. For the German language, we use the PhonDat1 dataset [29], which contain 201 speakers. Audio samples can be found online <https://ytaigman.github.io/polyglot/> and in the supplementary. It was shown in [10] that by em-

bedding averaging the embeddings of many short samples, we improve the synthesis performance. We follow this technique, and for the speaker embedding during inference, we take a mean of 20 training samples for each speaker, each of five seconds in length.

In order to evaluate the quality of the generated audio, we employ Mean Opinion Score (MOS). All samples were presented at the 16kHz sample rate. The raters were told that they are presented with the results of the different algorithms. At least 10 raters participated in each such experiment. In addition, we measure the speaker similarity with two methods. The first method, evaluates similarity MOS. We provide pair of audio samples - a ground truth audio sample of the source language and a synthesized audio sample in the target language. The raters were asked to evaluate the similarity of speakers on a scale with 1 – 5, where 1 and 5 scores corresponding to "different person", and "same person" respectively. The second method, uses a multiclass speaker identification deep network, employing the same identification neural network, as described in [9].

In Tab. 1, we evaluate the naturalness of the generated speech sentences for the three languages. As can be seen, we get good results for MOS score quality (above 3.0). However, it should be noted that the mean MOS score on the original samples from these datasets was about 4.46. Following encoding and decoding with the world vocoder, this drops slightly to 4.19.

In Tab. 2, we evaluate the similarity of the generated speech sentences for the three languages. As can be seen for the baseline, we get good self-similarity for the English language. Moreover, for English language with polyglot synthesis, we get good MOS similarity scores results, i.e. above 3.40. Spanish and German obtains lower MOS similarity scores than English. We conjecture that this is a result of the fact that the VCTK dataset is larger than DIMEx100 and PhonDat1 (VCTK has $\sim 40k$ voice samples, DIMEx100 has $\sim 5.5k$ voice samples and PhonDat1 has $\sim 15k$ voice samples). For the Spanish and German languages, we get good results for self-similarity (above 3.15), and the Spanish to German (and vice versa) MOS scores polyglot synthesis results are close to the self-similarity MOS scores on these languages.

In Tab. 3, we perform ablation analysis to show the contribution of the L_{poly} term to the polyglot synthesis. Without this term, the polyglot behavior arises only from the shared embedding. As can be seen, we get acceptable MOS similarity scores results (above 3.0). However, there is a noticeable gap in performance, in comparison to the results of the full system, as presented in Tab. 2.

In Tab. 4 and 5, we evaluate the contribution of the L_{poly} term on the automatic identification accuracy of the generated sentences. As expected, the best identification results are obtained for the self-language synthesis sentences (located along the diagonal). The polyglot synthesis accuracy (off di-

agonal) is good, in terms of both Top1 and Top5 accuracy (above 70% and 91%, respectively). Where Top1 and Top5 refer to the accuracy of correctly classify the right speaker in the top one and the top five guesses, respectively. We can observe that adding L_{poly} to the loss, improves the polyglot synthesis dramatically. For example, conversion from German to English improves in terms of Top1 accuracy from 82.71% to 92.29% when adding L_{poly} . Moreover, we can see that adding it also gets better or comparable results for same-language synthesis.

In Fig. 2(a) (and 2(b)), we provide t-Distributed Stochastic Neighbor Embedding (TSNE) visualization of the learned embedding for VCTK speakers in English and Spanish with (resp. without) the loss term L_{poly} . Each data point is the embedding, using N_s^a of one voice sample, where a is the sample's language. Data points with fragmented circles represent an original VCTK speaker in English, while data points with continuous circles represent the corresponding speaker in Spanish. As can be seen, adding L_{poly} improves the invariance of the speakers' embedding to the language spoken.

Table 1: MOS naturalness scores (Mean \pm SE)

<i>Source\Target</i>	English	Spanish	German
English	3.05 \pm 1.30	3.01 \pm 1.34	3.30 \pm 1.15
Spanish	3.05 \pm 1.07	3.33 \pm 1.17	3.08 \pm 1.19
German	3.35 \pm 0.99	3.12 \pm 1.08	3.02 \pm 1.00

Table 2: MOS similarity scores (Mean \pm SE)

<i>Source\Target</i>	English	Spanish	German
English	3.51 \pm 1.21	3.41 \pm 1.37	3.19 \pm 1.36
Spanish	3.52 \pm 1.25	3.27 \pm 1.36	3.22 \pm 1.38
German	3.40 \pm 1.25	3.15 \pm 1.35	3.16 \pm 1.39

Table 3: No L_{poly} - MOS similarity scores (Mean \pm SE)

<i>Source\Target</i>	English	Spanish	German
English	3.51 \pm 1.16	3.02 \pm 1.34	3.03 \pm 1.43
Spanish	3.00 \pm 1.37	3.01 \pm 1.43	3.13 \pm 1.33
German	3.09 \pm 1.37	3.14 \pm 1.34	3.27 \pm 1.42

Table 4: Identification accuracy top1[%] (top5[%])

<i>Source\Target</i>	English	Spanish	German
English	98.08 (99.66)	87.17 (94.57)	84.32 (93.63)
Spanish	70.74 (91.62)	89.88 (98.59)	70.02 (92.34)
German	92.29 (97.76)	90.67 (99.11)	94.40 (99.63)

Table 5: No L_{poly} - Identification accuracy top1(top5)

<i>Source\Target</i>	English	Spanish	German
English	97.96 (99.66)	83.33 (91.67)	80.44 (89.93)
Spanish	66.47 (87.85)	89.91 (98.61)	67.21 (91.32)
German	82.71 (94.15)	88.44 (97.78)	94.28 (99.63)

Acknowledgements

This work was carried out in partial fulfillment of the requirements for the Ph.D. degree of the first author.

5. REFERENCES

- [1] Aaron van den Oord, Sander Dieleman, Heiga Zen, et al., “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [2] Aäron van den Oord, Yazhe Li, Igor Babuschkin, et al., “Parallel wavenet: Fast high-fidelity speech synthesis,” *ICML*, 2018.
- [3] Wei Ping, Kainan Peng, and Jitong Chen, “Clarinet: Parallel wave generation in end-to-end TTS,” *arXiv preprint arXiv:1807.07281*, 2018.
- [4] Sercan O Arik, Mike Chrzanowski, Adam Coates, et al., “Deep voice: Real-time neural TTS,” in *ICML*, 2017.
- [5] Sercan Arik, Gregory Diamos, et al., “Deep voice 2: Multi-speaker neural TTS,” in *NIPS*, 2017.
- [6] Wei Ping, Kainan Peng, et al., “Deep voice 3: 2000-speaker neural TTS,” in *ICLR*, 2018.
- [7] Y Wang, RJ Skerry-Ryan, et al., “Tacotron: A fully end-to-end TTS synthesis model,” *Interspeech*, 2017.
- [8] Jonathan Shen, Ruoming Pang, Ron J Weiss, et al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” *ICASSP*, 2017.
- [9] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani, “VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop,” in *ICLR*, 2018.
- [10] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf, “Fitting new speakers based on a short untranscribed sample,” *ICML*, 2018.
- [11] Sercan O Arik, Jitong Chen, Kainan Peng, et al., “Neural voice cloning with a few samples,” *arXiv preprint arXiv:1802.06006*, 2018.
- [12] Younggun Lee, Taesu Kim, and Soo-Young Lee, “Voice imitating TTS neural networks,” *arXiv preprint arXiv:1806.00927*, 2018.
- [13] Ye Jia, Yu Zhang, Ron J. Weiss, et al., “Transfer learning from speaker verification to multispeaker TTS synthesis,” *NIPS*, 2018.
- [14] Jose Sotelo, Soroush Mehri, Kundan Kumar, et al., “Char2wav: End-to-end speech synthesis,” in *ICLR workshop*, 2017.
- [15] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber, “Training very deep networks,” in *NIPS*. 2015.
- [16] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, “Convolutional Sequence to Sequence Learning,” in *ICML*, 2017.
- [17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention is all you need,” in *NIPS*. 2017.
- [18] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, “World: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, 2016.
- [19] Alan W Black and Kevin A Lenzo, “Multilingual TTS synthesis,” in *ICASSP*, 2004.
- [20] Christof Traber, Karl Huber, Karim Nedir, Beat Pfister, Eric Keller, and Brigitte Zellner, “From multilingual to polyglot speech synthesis,” in *Sixth European Conference on Speech Communication and Technology*, 1999.
- [21] Yi-Jian Wu, Simon King, and Keiichi Tokuda, “Cross-lingual speaker adaptation for HMM-based speech synthesis,” in *Chinese Spoken Language Processing*, 2008.
- [22] Xianglin Peng, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, “Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices,” in *ICSP*, 2010.
- [23] Matthew Gibson, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, and William Byrne, “Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction,” in *ICASSP*, 2010.
- [24] Huaiping Ming, Yanfeng Lu, Zhengchen Zhang, and Minghui Dong, “A light-weight method of building an LSTM-RNN-based bilingual TTS system,” in *Asian Language Processing*. IEEE, 2017.
- [25] Bo Li and Heiga Zen, “Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis,” in *INTERSPEECH*, 2016.
- [26] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He, “Speaker and language factorization in DNN-based TTS synthesis,” in *ICASSP*, 2016.
- [27] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2017.
- [28] Luis A Pineda, Hayde Castellanos, Javier Cuétara, Lucian Galescu, Janet Juárez, Joaquim Llisterri, Patricia Pérez, and Luis Villaseñor, “The corpus dimex100: transcription and evaluation,” *Language Resources and Evaluation*, 2010.
- [29] Ludwig-Maximilians-Universitt München Bavarian Archive for Speech Signals, “Bavarian archive for speech signals (bas),” 2017.