

GENERATING MULTILINGUAL VOICES USING SPEAKER SPACE TRANSLATION BASED ON BILINGUAL SPEAKER DATA

Soumi Maiti* and Erik Marchi and Alistair Conkie

Apple

* The Graduate Center, CUNY, New York, NY

ABSTRACT

We present progress towards bilingual Text-to-Speech which is able to transform a monolingual voice to speak a second language while preserving speaker voice quality. We demonstrate that a bilingual speaker embedding space contains a separate distribution for each language and that a simple transform in speaker space generated by the speaker embedding can be used to control the degree of accent of a synthetic voice in a language. The same transform can be applied even to monolingual speakers.

In our experiments speaker data from an English-Spanish (Mexican) bilingual speaker was used, and the goal was to enable English speakers to speak Spanish and Spanish speakers to speak English. We found that the simple transform was sufficient to convert a voice from one language to the other with a high degree of naturalness. In one case the transformed voice outperformed a native language voice in listening tests. Experiments further indicated that the transform preserved many of the characteristics of the original voice. The degree of accent present can be controlled and naturalness is relatively consistent across a range of accent values.

Index Terms— cross-lingual transfer, d-vector, speaker space manipulation, bilingual speaker, text-to-speech synthesis

1. INTRODUCTION

A European survey found that 56% of people can converse in more than one language [1]. Many people in India, China, and African countries often speak more than one language. A person is *bilingual* if he or she is able to speak two languages fluently. A challenging problem for text-to-speech (TTS) systems is to make a synthesized voice bilingual or multi-lingual when recorded speech from only one language is available. A bilingual TTS voice is especially important for pronouncing words from one language embedded within another, i.e., *code switching*.

TTS voices are traditionally synthesized with monolingual speech. Changing the language of the text may result in a change of voice, which can be disconcerting or confusing to the listener. In contrast, it is possible to use hand-crafted mapping of phonemes from one language to another, which may not change the voice characteristics but results in heavily accented speech. The challenge for bilingual TTS is to maintain the same speaker voice characteristics while maintaining fluency or naturalness in both languages.

There are two areas of work related to this paper. The first utilizes data from a bilingual speaker to create a bilingual TTS system. In [2, 3], the authors model English-Chinese TTS with data from one and three bilingual speakers, in [3] one particular case of cross-lingual transfer is shown. Both models achieve relatively low speech

naturalness in terms of Mean Opinion Score (MOS), with $MOS \leq 3$. In the second area of work, multilingual TTS is trained on a large number of speakers. In [4] the authors observe partial sharing of similar sounding phonemes is preferred compared to separate phonemes in multiple Indian-English code switched sentences. In [5], an English-Spanish-German multilingual TTS model is trained with 409 speakers. The authors introduce a speaker-preserving loss to improve speaker similarity while converting a voice to another language. However, since they use the non-neural WORLD [6] vocoder, naturalness of voices is low ($MOS \sim 3$). In another recent paper [7], the authors built a multilingual TTS with 550 hours of data from 92 speakers. They show that, when trained with a large number of speakers, cross-lingual transfer with high naturalness is possible with a loss of speaker similarity.

In contrast to the work listed above, our TTS model is trained on a smaller dataset (56 hours and 7 speakers) and can maintain speaker similarity close to a bilingual speaker. We show that we can use information from the speaker embedding space to transform a monolingual voice to another language thus creating a true bilingual TTS. The key contributions of this paper are as follows: We propose a bilingual speaker TTS model, i.e. one TTS voice that speaks both languages with high naturalness. We also show that bilingual speaker embedding space contains separate distributions for two languages with visualization with Principal Component Analysis (PCA) and by fitting two Gaussian distributions with Linear Discriminant Analysis (LDA) with 99% accuracy. We demonstrate that monolingual TTS voices can be made to speak a second language fluently with a fixed shift (Δ) in the speaker embedding vector. This also gives us a means of controlling the degree of accent present in cross-lingual transfer.

2. TECHNICAL OVERVIEW

This work focuses on cross-lingual transfer using English and Spanish as our example languages. We use Tacotron [8] [9] to generate a mel-spectrogram from text and WaveRNN [10] to synthesize speech from the mel-spectrogram. The Tacotron is conditioned on a speaker embedding d-vector [11] from the speaker encoder, trained separately following [12].

2.1. Tacotron-ML

The baseline Tacotron is a sequence-to-sequence model with attention that takes phonemes as input and generates mel-spectrograms as output. Phonemes are converted to 512 dimensional embeddings through an embedding layer and then processed with a text encoder to generate encoded features of size 1024. Encoded features are passed to an attention layer and then an autoregressive decoder to synthesize mel-spectrograms. With the encoded features we append

*The first author performed the work while at Apple

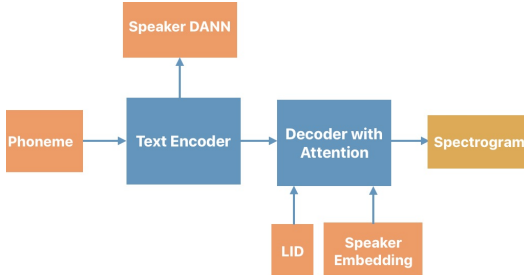


Fig. 1. Modified Tacotron-ML

a 128-dimensional speaker embedding, learned from the speaker encoder for a multispeaker Tacotron. We add an extra one-hot Language ID (LID) that converts to a 32-dimensional language embedding with an embedding layer. The LID for our case is 0 for English and 1 for Spanish.

Following [7], we also add a domain adversarial neural network (DANN) [13] to learn speaker-independent encoded features. The speaker DANN model takes encoded features as input and is trained using a speaker classification loss. The gradient reversal layer in the speaker DANN passes the negative gradient scaled by λ to the text encoder during back-propagation. This helps the text encoder avoid learning speaker information and instead rely on speaker embeddings for speaker information. In informal evaluations we found that without a speaker DANN, the model sometimes generates accented speech, whereas with DANN the Tacotron produces more consistent results. The modified Tacotron for multi-lingual (Tacotron-ML) is shown in Figure 1. During training of the Tacotron-ML, phoneme and LID are computed from the text and the language of the text. A speaker embedding for each training utterances is obtained from the speaker encoder. At synthesis time LID and phonemes are extracted from the text to be synthesized, and mean speaker embedding from training utterances is used for representing the target speaker. The Tacotron was trained for 1.5M steps with a batch size of 12.

We train a single WaveRNN model with multiple speakers data to synthesize speech from mel-spectrograms, with a GRU [14] size of 512 and training for 800K steps with batch size of 8. A better quality model might use a single speaker WaveRNN for each speaker, but the multispeaker WaveRNN is convenient for testing. The single model is also useful for comparing different speaker results with the same decoder.

2.2. Speaker Encoder

The speaker encoder comprises two LSTM [15] layers each with 512 units, followed by a 128-unit linear layer and a softmax layer with 22k units. The input to the LSTM is simply the sequence of MFCC frames (20 MFCCs per frame, 25ms data window, 100 frames per second). For stochastic optimization, we use Adam [16] with an initial learning rate of 10^{-3} and a mini-batch size of 128. The speaker encoder is trained following a curriculum learning procedure which improves both the robustness against various acoustic conditions and the generalizability towards less constrained-text scenarios. More details can be found in [17]. In [18] it was found useful for in-language voice adaptation.

Voice ID	Language	Gen	Locale
0 (<i>ref_{en}</i>)	en	M	US
1	en	F	US
2 (<i>ref_{es}</i>)	es	M	MX
3	es	F	MX
4	en	M	AU
5	en	F	AU
6	es	M	ES
7	es	F	ES

Table 1. 7 speaker English-Spanish dataset.

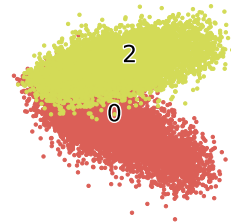


Fig. 2. PCA of a bilingual speaker embedding for each utterance in 7 hours of English (red parts) and 7 hours of Spanish (yellow parts).

2.3. Bilingual speaker embedding

The bilingual voice (Voice IDs 0 and 2 in in Table 1) is used as a *teacher* from which other voices can learn to be bilingual in English and Spanish.

The speaker encoder generates a speaker embedding for each utterance. The speaker embedding space is trained to discriminate between different speakers so that embeddings from the same speaker for multiple utterances have high cosine similarity between each other compared to different speakers. Such discriminative training results in each speaker utterances forming a cluster in the embedding space. We can observe such speaker clusters with PCA [19] or tSNE [20] on the multiple speaker embedding space (One such example can be found in this work [12]).

However, the speaker encoder is generally trained with speakers speaking only one language. There were no bilingual speakers in our training data for the speaker encoder. This is mainly because, speaker encoders are trained with hundreds of speakers, and it is difficult to get data from many bilingual speakers. Thus, it is interesting to learn how the speaker embedding space is distributed when the same speaker is speaking in two different languages, specifically whether it forms one cluster for both languages or two separate clusters. To the best of our knowledge there has been no previous work exploring such bilingual speaker embeddings. Since the speaker embeddings are used to condition for speaker output, it is also important to know how language is represented in these embeddings.

We explore the bilingual speaker embeddings from 14 hours of recorded speech (7 hours in Spanish and 7 hours in English) from the reference speaker. First we visualize the embeddings with PCA and in Section 3.2 we fit two Gaussian distributions and check the accuracy of the fit.

In Figure 2 we visualize language representation in bilingual speaker embedding space with principal component analysis (PCA) [19]. Our observations are:

1. These two languages have distinct clusters, and
2. these two language clusters have overlapping regions and then they diverge. Further investigation shows that the overlapping is due to smaller sentences (fewer than five words).

We hypothesize that any bilingual speaker space will have a similar structure, i.e., two distinct but overlapping clusters. If the monolingual voices could speak another language we would expect them to have similar speaker embedding clusters and so it should be feasible to modify them to speak another language simply by translating the speaker space, i.e. shifting the cluster corresponding to one language towards that of another. We detail this procedure in the following section.

2.4. Speaker embedding translation

We measure the mean speaker embedding cluster as μ_x , where x is voice ID and 8 voices (6 + reference speaker in two languages) are available. The bilingual reference speaker has two language cluster means μ_r^A and μ_r^B for English and Spanish respectively in the speaker embedding space. We compute the modification Δ_{A2B} needed to convert between the average cluster embeddings in both languages as:

$$\Delta_{A2B} = \mu_r^A - \mu_r^B \quad (1)$$

Note that since the speaker is the same, this shift can only contain information about the languages. Then for any different speaker x , who only speaks language A, we can obtain their embedding in language B as,

$$\mu_x^B = \mu_x^A + \epsilon \Delta_{A2B} \quad (2)$$

where ϵ ($0 \leq \epsilon \leq 1$) is a scaling factor to decide the level of the transformation, that is the level of native accent to be expected with the converted embedding. In our experiments it is set to zero or one except for our experiments investigating accent modification.

3. EXPERIMENTS

3.1. Dataset

We use an internal 56 hour dataset of 8 studio recorded voices, with ~ 7 hours of speech from each voice (counting the reference speaker as two voices with a total of ~ 14 hours of recordings). The dataset is balanced in the two languages (four English / four Spanish) and is balanced by gender, four male / four female. Two locales are represented for each language: en-US and en-AU for English and es-ES and es-MX for Spanish. For testing we use total of 400 sentences in English and 400 in Spanish; the sentences are 3 – 15 words long. The reference speaker is bilingual; all other speakers are monolingual. Using Eq. 2 we transfer any monolingual voices to another language.

3.2. Linear Discriminant Analysis

As our first experiment, we fit two Gaussian distributions to the bilingual speaker embeddings from the recorded speech with Linear Discriminant Analysis (LDA). We divide the bilingual speech as 75% / 25% for train-test. This gives 12525 training and 4175 test sentences. Next, we fit an LDA on training data to maximize class separability between English and Spanish bilingual embeddings. On the test set we observe 99% accuracy. This further shows that the two languages of the same speaker can be discriminated with very high accuracy.

	Voice ID	MOS	Std. Dev.
English Speakers	1	4.06	0.88
	0 (<i>ref^{en}</i>)	4.03	0.81
	baseline	2.48	1.16
Spanish Speakers	3 ^{en}	3.39	0.94
	2 ^{en}	4.08	0.73
	6 ^{en}	3.32	0.88

Table 2. MOS naturalness scores with English sentences. Lower section displays cross-lingual transfer following Eq. 2.

3.3. Listening Tests

We carried out five listening tests to gauge the subjective performance of our system. The tests fall into 3 categories. The first category is overall TTS naturalness (two tests, one per language); the second examines cross-language voice similarity (one test); the third examines the effect of accent on overall TTS naturalness (two tests, for two language).

Naturalness: A MOS test (Test 1) compared six variations of synthesis of English, three trained from English recordings (including a concatenative synthesis baseline reference) and three using Spanish voices modified to speak English (see Table 2). 12 test sentences were synthesized and played to 30 listeners, who were native speakers of English. Listeners were asked to rate voice naturalness on a 5-point scale from (1) Bad to (5) Excellent. A second similar experiment, Test 2, reversed the roles of Spanish and English (see Table 3).

Similarity: A similarity experiment (Test 3) examined to what extent speech in a different language can be identified as being from the same speaker. 84 pairs of sentences, with one sentence in English and one in Spanish, using seven different voice combinations (see Table 4) were presented to 30 native speakers of English and 30 native speakers of Spanish. Listeners were asked to rate the voice similarity for each pair of sentences on a scale from (1) Very Different to (5) Very Similar.

Accent: A MOS test (Test 4) asked listeners to give a naturalness rating on a 5-point scale from (1) Bad to (5) Excellent to English sentences generated from 3 Spanish voices for different values of ϵ described in Equation 2. A second MOS test (Test 5) reversed the roles of English and Spanish. Each test had ten sentences and nine voice configurations, and each test part had both English (30) and Spanish (30) listeners.

4. RESULTS

The naturalness results for Test 1 are shown in Table 2. The best scores are for Voice 1 and both versions of the reference bilingual speaker. An ANOVA followed by a Tukey post-hoc test found no statistically significant difference between these three versions. Speaker 3 and speaker 6 fall into a lower-scoring second group. The mapped version of the reference speaker performs as well as the in-language version, an indication that the technique performs well.

The results for Test 2 are shown in Table 3. For this experiment the scores for speaker 1 and both versions of the reference bilingual speaker again showed no statistical difference. Speakers 1, 3 and 4 formed a second group and finally speaker (0^{es}) and speaker 3 were grouped together. We highlight that the cross-language version of speaker 1, where the voice recordings are all for American English, was rated significantly higher speaking Mexican Spanish than speaker 3, a voice custom built to speak Mexican Spanish. Taking the two experiments together it seems likely that some element of

	Voice ID	MOS	Std. Dev.
Spanish Speakers	3	4.14	0.93
	2 (<i>ref_{es}</i>)	4.34	0.81
	baseline	2.29	0.91
English Speakers	0 ^{es}	4.30	0.85
	1 ^{es}	4.19	0.85
	4 ^{es}	4.10	0.92

Table 3. MOS naturalness scores with Spanish sentences. Lower section displays cross-lingual transfer following Eq. 2.

Voice ID pairs	MOS	Std. Dev.
3 ^{en} – 3	3.99	1.08
6 ^{en} – 6	3.84	1.22
1 – 1 ^{es}	3.37	1.28
4 – 4 ^{es}	3.16	1.26
0 – 2 (<i>ref_{en}</i> – <i>ref_{es}</i>)	3.05	1.30
1 – 3 ^{es}	2.50	1.22
0 – 6 ^{es} (<i>ref_{en}</i> – 6 ^{es})	2.43	1.27

Table 4. MOS similarity scores - English and Spanish.

voice preference is factored into the results. We intend to carry out further experiments to control for this effect.

The similarity results for Test 3 are shown in Table 4. The higher the score the higher the perceived similarity between pairs of voices in the two languages. Matched voices with cross-lingual transfer scored higher. Pairs 3 – 3^{en} and 6 – 6^{en} scored highest, showing the model was performing better for Spanish to English cross-lingual transfer than for English to Spanish. The two mixed voice combinations were rated lower than the same-voice combinations and the difference was statistically significant ($p < 0.001$). The bilingual speaker was perceived as less similar than some combinations, probably because the original recordings in each language had slightly different style requirements.

The cross-lingual transfer accent control results for Tests 4 and 5 are shown in Tables 5 and 6 respectively. Generally accent seems to be controllable and the least-accent cases are close to accent-free. For experiment 4, there was no statistically significant difference in terms of naturalness between the 3 reference speaker variants, nor the 3 speaker 3 variants, but the $\epsilon = 0$ version of speaker 6 was rated significantly lower.

For experiment 5, the results again show no significant difference for the reference speaker variants, but for the other two voices all the rating differences were significant. We need to investigate further to try to differentiate accent preference from naturalness.

5. VISUALIZING EFFECTS OF CROSS-LINGUAL TRANSFER IN SYNTHESIZED SPEECH

Finally, on synthesized speech we visualize the effect of speaker space translation: i.e. whether the TTS synthesized speech also mimics the bilingual speaker distribution or not. For this test we have 400 English and 400 Spanish sentences.

Synthesized test sentences are generated as follows: For the

Voice ID	$\epsilon = 0$	$\epsilon = 0.5$	$\epsilon = 1$
2 ^{en} _{ϵ}	4.08	4.22	4.16
3 ^{en} _{ϵ}	3.64	3.76	3.76
6 ^{en} _{ϵ}	3.43	3.69	3.73

Table 5. MOS quality with foreign accent - English

Voice ID	$\epsilon = 0$	$\epsilon = 0.5$	$\epsilon = 1$
0 ^{es} _{ϵ}	3.95	4.23	4.22
1 ^{es} _{ϵ}	3.42	3.89	4.16
4 ^{es} _{ϵ}	3.35	3.86	4.16

Table 6. MOS quality with foreign accent - Spanish

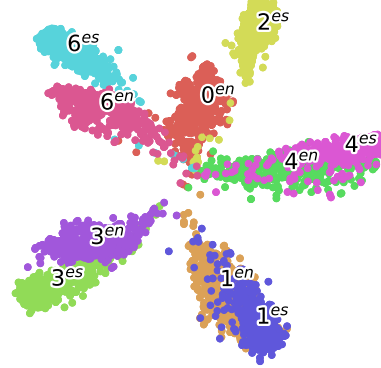


Fig. 3. tSNE plot with speaker embedding with synthesized speech. Monolingual speakers: 0^{en}, 1^{en}, 4^{en}, 2^{es}, 3^{es}, 6^{es}. Cross-lingual speakers: 3^{en}, 6^{en}, 1^{es}, 4^{es}. Same speaker English-Spanish clusters after cross-lingual transfer are close-by.

bilingual speaker we synthesize with their embeddings, English sentences with mean English speaker embedding and Spanish with mean Spanish speaker embedding. For monolingual speakers, we synthesize their native language with no translation and cross-lingual transfer with translation. So this gives us five speakers speaking two languages, one bilingual speaker and two English and two Spanish speakers speaking both languages, 10 voices in total. We visualize with a tSNE [20] plot in Figure 3. Here we can see that 10 voices form five distinct clusters and each cluster contains two subclusters. Here five clusters represent five speaker and two subclusters represent two languages. From the visualization it is easy to see speaker clusters after cross-lingual transfer are close-by and overlapping with their native speech clusters, following the structure similar to the bilingual speaker. Hence, speaker identity is maintained through cross-lingual transfer.

6. CONCLUSIONS

The formalism we have developed leads to high quality TTS in a second language without losing the characteristics of the voice. We found that there is a clustering by language in speaker embedding space for a bilingual speaker and we were able to use the cluster means to help control language and accent at inference time. There are a number of reasons why this technique is extremely interesting: (1) One key point is that in the transformation process the quality remains high. (2) It requires a relatively modest amount of data. (3) Having data from one bilingual speaker helps make other monolingual speaker bilingual, without the complexities of trying to record a monolingual speaker speak a second language. (4) We cite particularly as evidence of the promise of the technique that in one case we demonstrated a transformed voice that performed better than a high quality in-language voice. (5) It is possible to control the degree of accent present in the synthesis. All these are very desirable characteristics for synthesis.

7. REFERENCES

- [1] Special Eurobarometer, “Europeans and their languages,” *European Commission*, 2006.
- [2] Huaiping Ming, Yanfeng Lu, Zhengchen Zhang, and Minghui Dong, “A light-weight method of building an lstm-rnn-based bilingual tts system,” in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 201–205.
- [3] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He, “Speaker and language factorization in dnn-based tts synthesis,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5540–5544.
- [4] Sai Krishna Rallabandi and Alan W Black, “On building mixed lingual speech synthesis systems.,” in *INTERSPEECH*, 2017, pp. 52–56.
- [5] Eliya Nachmani and Lior Wolf, “Unsupervised polyglot text-to-speech,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7055–7059.
- [6] 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*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [7] 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 preprint arXiv:1907.04448*, 2019.
- [8] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgianakis, Rob Clark, and Rif A Sauros, “Tacotron: Towards end-to-end speech synthesis,” *arXiv preprint arXiv:1703.10135*, 2017.
- [9] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
- [10] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, “Efficient neural audio synthesis,” *CoRR*, vol. abs/1802.08435, 2018.
- [11] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5115–5119.
- [12] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., pp. 4480–4490. Curran Associates, Inc., 2018.
- [13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [15] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.
- [17] Erik Marchi, Stephen Shum, Kyuyeon Hwang, Sachin Kajarekar, Siddharth Sigtia, Hywel Richards, Rob Haynes, Yoon Kim, and John Bridle, “Generalised discriminative transform via curriculum learning for speaker recognition,” in *Proceedings 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Calgary, Alberta, Canada, 2018, IEEE, pp. 5324–5328.
- [18] Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik, and Sachin Kajarekar, “Neural Text-to-Speech Adaptation from Low Quality Public Recordings,” in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 24–28.
- [19] Svante Wold, Kim Esbensen, and Paul Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [20] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.