11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Improving Polyglot Speech Synthesis through Multi-task and Adversarial Learning

Jason Fong[1*], *Jilong Wu*[2], *Prabhav Agrawal*[2], *Andrew Gibiansky*[2], *Thilo Koehler*[2], *Qing He*[2†]

[1]The University of Edinburgh
[2]Facebook AI

jason.fong@ed.ac.uk, {jilwu, prabhavag, gibiansky, tkoehler, qinghe}@fb.com

Abstract

It is still quite challenging for polyglot speech synthesis systems to synthesise speech with the same pronunciations and accent as a native speaker, especially when there are fewer speakers per language. In this work, we target an extreme version of the polyglot synthesis problem, where we have only one speaker per language, and the system has to learn to disentangle speaker from language features from just one speaker-language pair. To tackle this problem, we propose a novel approach based on a combination of multi-task learning and adversarial learning to help the model produce more realistic acoustic features for speaker-language combinations for which we have no data. Our proposed system improves the overall naturalness of synthesised speech achieving upto 4.2% higher naturalness over a multispeaker baseline. Our qualitative listening tests also demonstrate that system produces speech which sounds less accented and more natural to a native speaker.

Index Terms: TTS, speech synthesis, multilingual, multi-task learning, generative adversarial networks

1. Introduction

The holy grail of Multilingual TTS is to build a truly 'polyglot' system, which can synthesise native-sounding speech in multiple languages using *any* of its voices. This polyglot capability would enable simple sharing of voices from high resourced languages to low resourced ones, resulting in an overall improvement of synthesis quality for low-resourced languages due to transfer learning. However, existing systems are far from this goal, since existing systems either require using a parallel multilingual corpora, which is expensive to record, or fail to fully disentangle speaker from language in synthesised speech if trained on a dataset with only monolingual speakers. In this paper we pursue model-based improvements to multilingual TTS in the extreme scenario where only one speaker per language is available.

It is important to tackle the limitations of existing systems since doing so would enable applications previously not possible that are both inclusive and key to connecting people across the globe. For example, polyglot TTS systems can allow the creation of personal voices, friends & family voices, and even celebrity voices in languages not spoken by each respective person. This is very exciting in the case of voice assistants, where it allows users to receive the same voice experience while maintaining speaker identity across multiple languages. In the scenarios above, we are usually familiar with the speaker, which makes us skilled at recognising a speaker's identity, subsequently this makes the problem of maintaining speaker similarity even more challenging[1].

Both unit-selection [2] and deep neural network [3] based Multilingual TTS approaches have shown good results leveraging large parallel corpora, consisting of 1000s of utterances per language per speaker. Parallel corpora improve polyglot synthesis by providing a wide coverage of how a speaker identity would pronounce phones in each target language. However, such parallel corpora are costly or sometimes impossible to procure since voice talents speaking multiple languages are rare and almost non-existent if we go beyond the most-spoken languages. Furthermore, even if multilingual voice talents are available, their proficiencies in their languages are unlikely to all be at a native level as the authors of [3] found.

Subsequently newer approaches to polyglot TTS have focused on lessening the need for native-level parallel corpora. Some approaches have tried using cross-lingual voice cloning to augment monolingual recordings thereby creating artificial parallel datasets [4] but these approaches require explicit voice cloning models, which have faced issues with producing good quality cross-lingual output.

More recent approaches have sought to train using only monolingual corpora. The difficulty of training on only monolingual corpora however is that of speaker and language factor entanglement. Since speakers only speak one language, there is perfect correlation between speaker identity and language in the data, making factorisation difficult, and potentially resulting in the model ignoring the language conditioning feature. This is problematic however as an acoustic factorisation [5] of speaker identity and language must be obtained in order for a model to be able to then generate arbitrary combinations of speaker and language. The approaches of [6, 7, 8] attempt to achieve factorisation by representing speaker and language factors as distinct transformations that are then applied sequentially to input linguistic features. [9, 10] alternatively use speaker and language features to condition the decoder of seq2seq TTS systems and then attempt to achieve a speaker-language factorisation by training with multiple speakers from multiple languages. The advantage of this approach is that it forgoes the need for adding separate modules or layers for different speakers and languages.

In this paper, we focus on the problem of building polyglot TTS systems using solely monolingual corpora. Our main contribution is to further improve cross-lingual voice quality through the use of additional training losses and tasks. Our approach, in a similar vein as [9], uses an adversarial loss to improve multilingual performance, however we apply it to predicted acoustics to improve the realisation of acoustics in general and phones in particular.

Our model architecture is novel but is slightly similar in concept to [11] that uses a loss term to preserve speaker identity,

[*]Work performed while interning at Facebook AI.
† Correspondence to Qing He

10.21437/SSW.2021-30

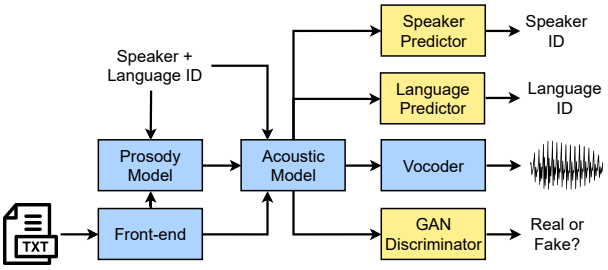

Figure 1: *Proposed model overview. Our baseline acoustic model is coloured blue, and proposed model additions are coloured in yellow.*

but this is performed over speaker embeddings only, whereas we do it using a multi-task speaker-language prediction task over predicted acoustics. They do this to avoid the problem of speaker embeddings also encoding language information. Their model doesn't use language-based conditioning features, and instead relies on using language specific text encoders. They train in a scheduled manner, first training the network to synthesise multilingual speech and then optimising the speaker embedding space for polyglot synthesis using unseen speaker-language combinations. In our model we do not perform such scheduling. [12] similarly tries to resolve language dependency in the speaker space by viewing cross-lingual TTS as a domain adaptation problem and attempt to learn a language independent speaker space.

The main contribution of this work is in improving the naturalness and quality of speech in a language foreign to the original voice talent. We demonstrate that multi-task learning over speaker and language features combined with a GAN inspired adversarial loss can be fruitful when little data is available but polyglot systems are required.

2. Proposed Multilingual Acoustic Model

At the core of our proposed model is a seq2seq acoustic model (AM) that predicts output vocoder features from input linguistic features concatenated with speaker/language one-hot vectors. To improve the AM's Speaker Language Factorisation (SLF), its core acoustic loss function is augmented with additional losses obtained from supplementary tasks. An overview of the losses and tasks in our proposed model is found in Figure 1 and the following subsections detail the AM and each of its augmentations. To keep the notation of our various losses clear we use the following notation: a loss L's subscript denotes which model it is used to help train, and its superscript denotes where it is obtained from.

Since the augmentations detailed in the subsequent subsections work with the AM's outputs and are not tied to our particular AM architecture they are subsequently likely usable with other AM architectures, such as transformers or feed forward networks for example.

2.1. Multilingual Acoustic model (AM)

The AM receives as input a series of T frame-wise linguistic features ($\mathbf{x}_{1:T}$) and is trained to output a corresponding series of frame-wise 'vocoder' features ($\hat{\mathbf{y}}_{1:T}$), such as MFCCs, f_0, and periodicity features that can be fed to a signal processing based vocoder [13, 14], or mel-spectrograms that can be used to condition a neural vocoder [15, 16, 17]. We define a forward pass through the acoustic model as follows: $\hat{\mathbf{y}}_{1:T} = AM(\mathbf{x}_{1:T})$. Further description of the linguistic and acoustic features used to train models for our experiments is deferred to Subsection 3.1.

The AM uses the encoder-decoder with multi-rate attention architecture of [18]. The encoder and decoder are both unidirectional single-layer LSTMs with 512 hidden dimensions. The decoder additionally uses a multi-rate attention mechanism to attend over the hidden states of three encoders, providing contextual information relevant to a particular decoder timestep by attending over frame, syllable, and word-level features.

The AM, is *primarily* trained using an acoustic L^2 loss between ground truth and predicted vocoder features. We denote this primary loss component as $L_{AM}^{acoustic} = \sum_{t=1}^{T} L^2(\mathbf{y}_t, \hat{\mathbf{y}}_t)$ where $\mathbf{y}_t$ is a particular frame of ground truth acoustics, and $\hat{\mathbf{y}}_t$ is its corresponding predicted frame. Additional loss components detailed in the following subsections are used with $L_{AM}^{acoustic}$ in order to obtain Equation 1 which is the final loss function used to update the AM's weights during training:

$$L_{AM} = L_{AM}^{acoustic} + \alpha L_{AM}^{MT} + \beta L_{AM}^{adv} \quad (1)$$

where $\alpha = 0.025$ and $\beta = 20.0$ are weights for each loss component discovered from hyperparameter search.

2.2. Speaker & Language Multi-task prediction heads

Along with the AM we train speaker and language multi-task (MT) prediction heads for one core reason: such prediction tasks serve as an inductive bias [19] that can encourage the AM to utilise speaker and language features. Both prediction heads use the *same* series of T frame-wise acoustic predictions from the AM to make a *downsampled* series of U categorical predictions over k classes (k_S speaker classes or k_L language classes).

The architecture of each prediction head consists of 5 1D convolutional layers each with 256 filters, a stride of 3, kernel size 5, and a padding of 2. Book-ending the 5 convolutional layers are two linear projection layers: an input layer projects features from Dim_{in} to Dim_{hid} dimensions, and an output layer projects features from Dim_{hid} to Dim_{out} dimensions. Additionally we apply a dropout of 0.2 to the input features before the first linear projection layer and each convolutional layer uses the Leaky ReLU [20] activation function with a leakiness of 0.2 and slope of -0.1.

We train the prediction heads' weights using a Cross-Entropy loss between their output logits and ground-truth one-hot targets. The loss for each prediction head is $L_{MT} = \sum_{u=1}^{U} CE(\mathbf{c}_u, \hat{\mathbf{c}}_u)$ where $\hat{\mathbf{c}}_{1:U} = MT(\hat{\mathbf{y}}_{1:T})$ represents speaker or language predictions obtained from passing predicted acoustics through the multi-task heads, $\mathbf{c}_{1:U}$ represents a corresponding series of one-hot ground truth classes, and $CE(.)$ is the Cross-Entropy loss function.

Note that we train the prediction heads using predicted acoustics $\hat{\mathbf{y}}_{1:T}$ rather than ground truth acoustics $\mathbf{y}_{1:T}$ in order to avoid train-test mismatch that can be caused by teacher-forcing.

By default we do *not* detach the acoustic predictions from the computation graph before feeding them to the multi-task heads so that L_{MT} also updates the AM's weights during training. Therefore we also refer to L_{MT} as L_{AM}^{MT}. We also experimented with detached multi-task losses, in which case L_{MT} does not update the AM, but found that in doing so our model does not improve over our baseline.

2.3. Adversarial training of AM

To complement the multi-task prediction heads we introduce a GAN discriminator that is trained to predict whether a series of acoustic features are either ground truth (*'real'*) or predictions generated by the AM (*'fake'*). We use the GAN discriminator to help ensure that the AM uses speaker/language inputs in a perceptual way rather than *cheating* by minimising L_{MT} in non-perceptual ways. That is, by encoding speaker and language information into the predicted acoustics in a acoustically non-perceivable way.

The architecture of the discriminator follows that of [21], consisting of 10 1D convolutional layers each with 128 filters, a stride of 1, kernel size 3, and a linearly increasing dilation rate (dilation increases by 1 per layer). Identical to the multi-task prediction heads detailed in Subsection 2.2 the discriminator's convolutional layers are each followed by LeakyRELU activation functions and are book-ended by linear projection layers. The final projection layer which projects from Dim_{hid} to Dim_{out}, where Dim_{out} is equal to 1, is followed by a Sigmoid activation function, collapsing the model's output to the range $[0, 1]$ and as such its output can be interpreted as the probability that the discriminator's input is real acoustic data.

To train the discriminator to differentiate between real and fake acoustics we adopt a two component loss $L_D = L_D^{real} + L_D^{fake}$. We train the discriminator to output 1 when it recognises real acoustics with $L_D^{real} = \sum_{t=1}^{T} L^2(\mathbf{r}_t, 1)$, and train it to output 0 when it recognises fake acoustics with $L_D^{fake} = \sum_{t=1}^{T} L^2(\mathbf{f}_t, 0)$ where $\mathbf{r}_{1:T} = D(\mathbf{y}_{1:T})$ and $\mathbf{f}_{1:T} = D(\hat{\mathbf{y}}_{1:T})$ are generated from the discriminator by feeding it ground truth and predicted acoustics respectively.

Finally we obtain from the discriminator an adversarial loss $L_{AM}^{adv} = \sum_{t=1}^{T} L^2(\mathbf{f}_t, 1)$ that is incorporated into the AM's loss function to help ensure its predicted acoustics are high quality and perceptually synthesise speaker and language. This loss is minimised when the AM successfully generates acoustics that fool the discriminator into believing that they are real.

2.4. Training loop

In this Subsection we define one iteration of the training loop for our proposed model.

1. Use inputs $\hat{\mathbf{x}}_{1:T}$ to get AM predictions $\hat{\mathbf{y}}_{1:T}$.
2. Use $\hat{\mathbf{y}}_{1:T}$ to a) get the acoustic loss $L_{AM}^{acoustic}$, b) get the GAN discriminator adversarial loss L_{AM}^{adv}, and c) calculate speaker and prediction losses through the multi-task heads to obtain L_{MT} and use this loss to train the heads.
3. Combine all of the AM's losses to get Equation 1 and use it to update the AM.
4. Use the inputs $\hat{\mathbf{x}}_{1:T}$ again to get a *new* set of AM predictions and use them to obtain L_D and train the GAN discriminator.

3. Experimental setup

To evaluate the efficacy of our proposed model, we perform a subjective listening test to compare its performance against two baseline models. A monospeaker baseline and a multispeaker baseline. This section describes the details of these experiments.

3.1. Input representations

The framewise input features used by our AM are obtained by up-scaling the output of our linguistic front-end. This up-scaling is performed using durations obtained from a prosodic model that predicts both the duration and f_0 of each phone aligned frame of contextual linguistic features.

In order to improve multilingual TTS performance by encouraging the model to share language-independent acoustic knowledge across languages, our front-end produces a *shared* phone representation common to all our languages. Previous work has approached this by using a phone set that is common across all languages [11]. Recent work [22] however uses 'phonological features' (PFs) as input to a neural TTS system. These PFs features have been shown to enable zero-shot multilingual TTS to unseen languages, and [23] also show that using PFs improves intelligibility and naturalness for low-resourced languages due to pooling of data, and pervasive sharing of encoder parameters across languages. Our model similarly uses multidimensional PFs to represent each phone. We start with a phone-set, which represents phonetic identity using the various dimensions for speech production such as place of articulation, and manner of articulation. This ensures that our baseline system can produce multilingual output of reasonable quality , without requiring an explicit mapping between phone-sets.

3.2. Modelling

Both our baseline and proposed acoustic models share the same core multi-rate attention architecture [18]. Acoustic or prosodic features are predicted for every frame by a recurrent LSTM module. Additionally contextual information at different levels relevant to producing a particular timestamp of acoustics is summarised from the entire input sequence by the multi-rate attention module. Previous experiments have found that the usage of multiple attention alignments overall improve prosody realisations from input linguistic features.

The acoustic models predict spectrum features, which is a 19-dim feature vector consisting of 1-dim f_0 vector, a 13-dim MFCC vector along with a 5-dim periodicity vector. Our conditional neural vocoder is a WaveRNN [16] model, with hidden dimension 1024. It takes in the 19-dim spectrum features and generates the audio waveform at 24kHz.

Our AMs and vocoder are additionally made multilingual via the use of speaker and language one-hot conditioning features. In this work we use one-hot features rather than speaker embeddings as in this study we focus on improving polyglot synthesis, rather than enabling multilingual synthesis for new unseen speakers, which we leave as potential future work.

3.3. Training setup & Data

Our acoustic models are trained with the Adam optimizer with a learning rate of 1e-4. We implemented them using Pytorch and conduct the training with distributed GPU clusters. After some fine-tuning, we decide to train at 500K steps with a training time of approximately 2 days using batch size of 32.

The TTS datasets were recorded in a voice production studio by contracted professional voice talents. Our multilingual dataset `5lang-5speaker` contains five voices each speaking a different language: English (30 hours), Spanish (23 hours), Italian (9 hours), German (8 hours) and French (10 hours) and the data was collected at a 24kHz sampling rate. `5lang-5speaker` is used to train both the baseline and proposed multilingual AMs, and our multilingual multi-

speaker WaveRNN. We additionally use each individual voice in `5lang-5speaker` to train monospeaker baseline AMs that can still perform some level of multilingual TTS due to our use of language-independent phonological features.

3.4. Evaluation

We have designed our listening tests to answer one question regarding our proposed AM vs baseline AMs: does adding speaker and language prediction tasks along with adversarial training improve the overall naturalness of speech when synthesising polyglot 'non-native' speech.

We synthesised each language's test set conditioning using a *non-native* speaker, that is a monolingual speaker whom has no data in that particular language. In other words in our experiment we examine how well *each* of our dataset's speakers perform at 'non-native' polyglot synthesis. We use the following speaker and language combinations for generating our *non-native* test sets: S_{ES}-T_{EN}, S_{DE}-T_{ES}, S_{IT}-T_{FR}, S_{FR}-T_{DE}, S_{EN}-T_{IT}. For clarification EN is English, ES is Spanish, FR is French, IT is Italian, and DE is German. Also S_{ES} refers to our Spanish speaker and T_{EN} refers to our English test set.

Using a crowdsourcing platform we recruited the following number of participants for each test set language: 349 English, 214 Spanish, 39 French, 300 Italian, and 61 German. Participants are all native speakers of the language that they are rating. Each participant is shown 50 stimuli from that language and are asked to rate them from 1 to 5 in terms of naturalness as a voice assistant. We use these ratings to obtain an averaged naturalness MOS for each system.

3.5. Voice training and inference

We trained a total of 7 AMs for submission to listening tests: 5 monospeaker baselines, 1 multispeaker baseline, and 1 proposed multispeaker model. They are each trained with the following data and hyper-parameters:

- $\texttt{B}_{mono}$: We train 5 monospeaker baselines each one trained using a single native dataset as described in Subsection 3.3.
- $\texttt{B}_{multi}$: We train a single multispeaker baseline using the `5lang-5speaker` dataset.
- $\texttt{P}_{multi}$: We train a single multispeaker proposed model using the `5lang-5speaker` dataset. It differs from $\texttt{B}_{multi}$ with its use of speaker and language prediction tasks with adversarial loss during training.

To generate listening test stimuli for our subjective evaluations we use the *non-native* speaker and language combinations discussed in Subsection 3.4 to condition each multispeaker model in order to generate the test set that matches the language. The monospeaker models generate a non-native language for which it never saw any training data. For example $\texttt{B}_{EN}$ is used to generate the Italian test set, even though it used only English data during training. Again this is made possible by our model's use of phonological features rather than language specific phone-sets. A selection of samples used in our listening test can be found on our webpage for this paper[1].

4. Results

A summary of our MOS listening test results can be found in Figure 2. We observe several clear trends across the three types of systems: First of all $\texttt{B}_{multi}$ consistently out-performs $\texttt{B}_{mono}$, suggesting that training acoustic models with data from multiple speakers and languages is beneficial even given we only have one speaker per language. Secondly, except from the S_{EN}-T_{IT} stimuli, $\texttt{P}_{multi}$ consistently out-performs $\texttt{B}_{multi}$, suggesting that our proposed model modifications make an improvement in both quality and naturalness. The largest gains from using our proposed model are seen with S_{FR}-T_{DE} where naturalness is improved by 4.2 % over the multispeaker baseline. When listening to the test set stimuli we discovered that our proposed model also makes improvements in how native each utterance sounds and in phone intelligibility. We include examples on our samples page reflecting these findings.

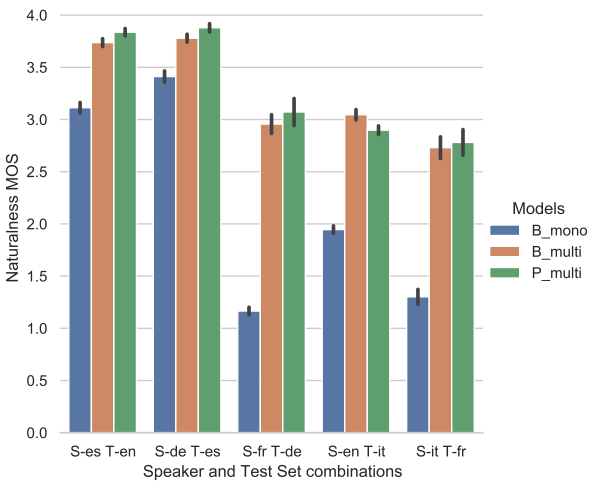

Figure 2: *Mean opinion scores obtained from our subjective listening test described in Subsection 3.4. 95% confidence intervals are depicted as black lines. Colours of the bars refer to one of three model types: monospeaker baseline, multispeaker baseline, and multispeaker proposed model. Further details regarding these models can be found in 3.5.*

5. Conclusions

In this work, we have proposed a novel way to improve polyglot speech synthesis across five languages through adversarial learning and multi-task training. According to a MOS study using on average 200 native raters per language, our proposed model achieved better overall quality compared with a multispeaker and multilingual baseline. As for future work, we plan to extend the proposed idea to prosodic modelling and combine it together with our proposed acoustic model. Another direction we also would like to pursue is using data augmentation methods to further improve the overall quality using synthetic polyglot data.

[1] https://multilingual-tts.github.io/samples/

6. References

[1] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, no. 3, pp. 265–275, 1991.

[2] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft mulan-a bilingual tts system," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[3] H. Ming, Y. Lu, Z. Zhang, and M. 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.

[4] S. Zhao, T. H. Nguyen, H. Wang, and B. Ma, "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," *arXiv preprint arXiv:2010.08136*, 2020.

[5] M. Gales, "Acoustic factorisation," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.* IEEE, 2001, pp. 77–80.

[6] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, "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.

[7] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis," 2016.

[8] Y. Fan, Y. Qian, F. K. Soong, and L. 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.

[9] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *arXiv preprint arXiv:1907.04448*, 2019.

[10] Amazon, "English-language Alexa voice learns to speak Spanish," 2021. [Online]. Available: https://www.amazon.science/blog/english-language-alexa-voice-learns-to-speak-spanish

[11] E. Nachmani and L. 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.

[12] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space," *Proc. Interspeech 2020*, pp. 2947–2951, 2020.

[13] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[14] 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.

[15] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[16] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[17] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[18] Q. He, Z. Xiu, T. Koehler, and J. Wu, "Multi-rate attention architecture for fast streamable text-to-speech spectrum modeling," *arXiv preprint arXiv:2104.00705*, 2021.

[19] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1. Citeseer, 2013, p. 3.

[21] A. Polyak, L. Wolf, Y. Adi, and Y. Taigman, "Unsupervised cross-domain singing voice conversion," *arXiv preprint arXiv:2008.02830*, 2020.

[22] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological features for 0-shot multilingual speech synthesis," *arXiv preprint arXiv:2008.04107*, 2020.

[23] A. Gutkin, "Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages," 2017.