StrawNet: Self-Training WaveNet for TTS in Low-Data Regimes

Manish Sharma, Tom Kenter, Rob Clark

Google UK

{skmanish, tomkenter, rajclark}@google.com

Abstract

Recently, WaveNet has become a popular choice of neural network to synthesize speech audio. Autoregressive WaveNet is capable of producing high-fidelity audio, but is too slow for real-time synthesis. As a remedy, Parallel WaveNet was proposed, which can produce audio faster than real time through distillation of an autoregressive teacher into a feedforward student network. A shortcoming of this approach, however, is that a large amount of recorded speech data is required to produce high-quality student models, and this data is not always available. In this paper, we propose StrawNet: a self-training approach to train a Parallel WaveNet. Self-training is performed using the synthetic examples generated by the autoregressive WaveNet teacher. We show that, in low-data regimes, training on high-fidelity synthetic data from an autoregressive teacher model is superior to training the student model on (much fewer) examples of recorded speech. We compare StrawNet to a baseline Parallel WaveNet, using both side-by-side tests and Mean Opinion Score evaluations. To our knowledge, synthetic speech has not been used to train neural text-to-speech before.

1. Introduction

With the increasing use of personal assistants in our daily lives, it has become more important than ever to deliver high-quality speech synthesis. The deep neural network revolution has caused a paradigm shift in research and development of text-to-speech systems, outperforming previous statistical and parametric methods [1, 2, 3]. WaveNet [4] is a class of deep neural networks known to generate seemingly realistic speech. The original WaveNet is an autoregressive network that uses dilated convolutions to model the probability distribution of the next speech sample to be generated. With sufficient training examples, it has been demonstrated that this model can generate speech and music with high fidelity. However, a major shortcoming of this network is that it is autoregressive and it generates speech at about 172 timesteps/second [4], which is prohibitively slow for synthesizing speech above 16kHz. Parallel WaveNet [5] was introduced to mitigate these speed concerns, and was shown to generate samples at about 500,000 timesteps/second. Yet, Parallel WaveNet too has limitations, in particular when there is little recorded data to train it on. The synthesis output can contain artefacts, like static noise, which becomes more prominent with fewer training samples of recorded speech.

A single-speaker WaveNet requires about 24 hours of recorded speech [4, 5] for training a good voice. This also holds true for other neural vocoders, for example [6] shows that a single-speaker Tacotron model trained on 25k utterances is substantially better than the ones trained on 15k and 8.5k recordings. In a multi-speaker training scenario, [5] showed that a high-quality voice can be obtained with about 10 hours of recorded speech per speaker if the capacity of the network is increased. However, it can be seen from their results that the quality degrades when the number of recordings is further decreased.

To reduce the voice artefacts observed in WaveNet student models trained under a low-data regime, we aim to leverage both the high-fidelity audio produced by an autoregressive WaveNet, and the faster-than-real-time synthesis capability of a Parallel WaveNet. We propose a training paradigm, called StrawNet, which stands for "Self-Training WaveNet". The key contribution lies in using high-fidelity speech samples produced by an autoregressive WaveNet to self-train first a new autoregressive WaveNet and then a Parallel WaveNet model. We refer to models distilled this way as StrawNet student models.

We evaluate StrawNet by comparing it to a baseline WaveNet student model in side-by-side preference tests, and by performing Mean Opinion Score (MOS) analysis. We show that our approach alleviates the problems observed when distilling student models in low-data regimes. Note that our approach is applicable for a certain segment of the spectrum of available data: too few recordings might not yield an autoregressive teacher network capable of synthesizing good data; too many recordings might not bring out the efficacy of self-training.

2. Related Work

Self-training, or self-learning refers to the technique of using an already trained system to generate outputs on unseen input examples and using these generated outputs as targets for subsequent re-training. Self-training has been shown to be useful in certain applications like linguistic parsing [7], sentiment analysis [8] and unsupervised neural machine translation [9]. Recently, a 2% increase in the top-1 accuracy in the ImageNet classification task was achieved through noisily distilling EfficientNet, using artificial labels on unlabelled images [10]. Our approach is similar to these approaches in that we use self-training, albeit on a synthesis task and not a discriminative one. An additional difference in our setup is that Parallel WaveNet is obtained via knowledge distillation from a different learning paradigm, i.e., an autoregressive WaveNet. This is advantageous because, compared to the samples a feedforward WaveNet student model produces, an autoregressive WaveNet teacher model provides synthetic examples of higher quality for self-training (cf. §5.1). Lastly, synthesized data has been employed before in speech domain in tasks including improving speech recognition [11] and emotion recognition [12]. To the best of our knowledge, training a neural acoustic model on synthesized speech has not been done before.

3. StrawNet

Figure 1 shows a schematic overview of both the baseline and the StrawNet approach. The conventional way of training a Parallel WaveNet [5] is a two-step procedure. In the first step, shown in the top-left corner of Figure 1, an autoregressive

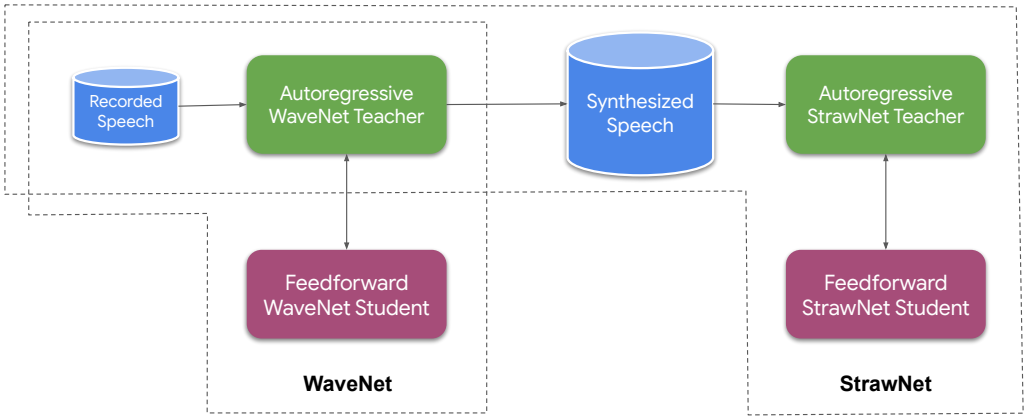

Figure 1: *Overview of the conventional way of training a Parallel WaveNet model and the proposed StrawNet approach.*

WaveNet teacher is trained to model the probability distribution of the next sample given previous samples of the recorded speech. This step is highly parallelizeable during training, as all recorded speech samples are available, so teacher forcing can be used [13]. The second step consists of distilling the density distribution of the autoregressive teacher network trained in the previous step, into a feedforward network. The feedforward student network, unlike the autoregressive network, takes as input, random noise sampled from a logistic distribution, along with other linguistic features and $\log F_0$ conditioning.

Our proposed StrawNet approach uses the autoregressive WaveNet trained in first step of the conventional approach to generate synthetic speech, and further train additional WaveNet models, referred to as StrawNet student and StrawNet teacher. This approach differs from the conventional approach in the crucial additional step of generating synthetic speech from the autoregressive WaveNet model (Figure 1). Because this is a one-time offline processing task, there are no limits (other than practical ones) to the number of speech samples we can generate. It is important that we use the autoregressive WaveNet teacher for generating this synthetic training data, rather than the distilled student model, because the speech it generates is of higher fidelity. The synthetic dataset obtained this way has a more generalized speech distribution than the real recordings, while being plentiful enough for distillation. It does come at a computational cost, as an autoregressive network is used for synthesis. We argue that the overall improvement in voice quality outweighs this computational investment. Furthermore, unlike WaveNet, the StrawNet approach requires a prosody model for the synthesis step. We use a pre-trained CHiVE model [14] to provide this prosody conditioning.

An important component of self-training is the unlabelled data synthetic examples are generated from. For StrawNet, this is the text used to synthesize speech, which is typically readily available. However, script selection with regards to good phonetic coverage is important to prevent the system becoming biased to artefacts of the data. Lastly, noisy/corrupt synthetic speech samples produced by the StrawNet teacher model are pruned from the generated set of synthetic speech samples, as they might affect the StrawNet student model. See §4.3 for further details on the above two steps.

4. Experimental Setup

In this section, we detail the input features, network and its hyperparameters, and the methodology used in our experiments.

4.1. Features

The end-to-end text to speech system employed in the analysis in following sections is composed of either a WaveNet, or a StrawNet acoustic model, conditioned on features encoding linguistic and prosodic information. The linguistic conditioning consists of phoneme, syllable, word, phrase and utterance level features, derived by running a text normalization [15, 16] system followed by rule-base feature computations [17]. The prosody conditioning is provided by a pre-trained hierarchical variational autoencoder as described in [14]. This prosody model uses the linguistic conditioning as described above to output the phoneme durations and $\log F_0$ per frame.

4.2. Baseline WaveNet architecture

The model architecture, loss functions and training recipe for both the autoregressive WaveNet teacher and the feedforward WaveNet student components in Figure 1 is the same as mentioned in [5]. We use a mixture of 10 logistic distributions to model the 16-bit audio samples in the 24kHz recordings. Both the WaveNet teacher and the WaveNet student model have dilated residual blocks as their constituent units. Each dilated residual block consists of 10 layers of dilated convolutions, increasing by a factor of 2 in every layer and reaching a maximum of 512 in the last layer. The WaveNet teacher has a stack of 3 such dilated residual blocks. The student network has 4 inverse autoregressive flows [18], each containing 1, 1, 1 and 3 dilated residual blocks respectively.

The loss function for the WaveNet teacher is the negative log-likelihood of the predicted mixture distribution. For the WaveNet student, the loss function is a weighted sum of four components: the KL divergence between the student and teacher distributions, mean squared error between the predicted and target signal powers, phoneme classification error from a WaveNet like classifier, and finally a contrastive loss that maximizes the difference between KL divergences of student and teacher distributions obtained with correct and incorrect conditionings, respectively. Both the component networks are trained for 1M iterations with Adam optimizer [19] using TPUs.

4.3. StrawNet

To generate synthetic data, we run the pipeline detailed in §4.1 with the WaveNet teacher model trained as described in §4.2. Our objective is to generate as much data as possible. However, it would be detrimental to the performance of the StrawNet stu-

dent and teacher networks for the synthetic dataset to contain a high bias with regards to the phoneme distribution. In order to avoid such a bias, we use a script selection methodology [20]. As the synthesis process can be very time consuming, we employ a parallelizable framework [21] for this computation.

Lastly, we train an HMM-based aligner to generate phoneme alignments for the synthesized audio samples. This aligner outputs a likelihood score for each utterance, where low scores are correlated to bad phoneme alignment. To ensure that the synthetic dataset is of good quality, we reject utterances that produce a bad alignment score because they are likely to be unintelligible or noisy. After pruning the phoneme-aligned utterances, we train a StrawNet by firstly training an autoregressive teacher network on this synthesized dataset, and then distilling it into a feedforward StrawNet student. The configuration and training procedures of the autoregressive teacher and feedforward student networks are as described in §4.2.

4.4. Evaluation

We run two sets of experiments: one on a single-speaker voice and one on a two-speaker voice.

In the experiments on a single-speaker voice, to understand what a "low-data regime" for WaveNet training means, and how StrawNet can help overcome that, we analyze the difference in voice quality between a WaveNet teacher, a WaveNet student and a StrawNet student. We train models on subsets of different sizes of single-speaker recordings, and investigate the variation in voice quality as we increase the subset size. We use an en-US male speaker, referred to as speaker A, for which a total of 24k recorded utterances is available. We generate subsets of these recordings of sizes 3k, 5k, 7k, 10k, 15k and 24k. For the synthesis step in StrawNet, we always generate 40k utterances as this was found to be optimal (cf. §5.2). Note that the effective number of utterances available for training StrawNet teacher and student networks is less than 40k, due to the screening (cf. §5.3).

In the second set of experiments, we train a two-speaker voice. In text-to-speech, it is uncommon to have a large number of recordings ($> 10k$ utterances) from a single speaker. For speakers with an insufficient number of recordings, the preferred approach is to train a voice using a multi-speaker WaveNet, instead of a single-speaker WaveNet, so the speech characteristics common to all speakers can be jointly learnt. In our second set of experiments, we compare the voices of a data-deficient speaker trained using two-speaker WaveNet and StrawNet models. For this task, we use the recordings from another male en-US speaker, referred to as speaker B, for whom 2500 utterances are available, and train it jointly with 24k utterances of speaker A. To determine the optimal amount of synthesized speech required, we analyze the effect of varying the number of synthesized examples on the voice quality.

We employ two types of tests to evaluate the StrawNet approach against the baseline WaveNet. Firstly, MOS (Mean Opinion Score) tests are carried out, where raters are asked to score an audio sample on a scale of 1 to 5. Secondly, side-by-side tests are employed, where raters are presented with two audio files of the same utterance, and are asked to indicate which one sounds better. We test for statistical significance using a two-sided t-test with α of 0.01.

We use 1000 sentence test sets, selected to be typical TTS assistant use cases. Each rater is allowed to rate a maximum of 6 sentences, to reduce personal bias.

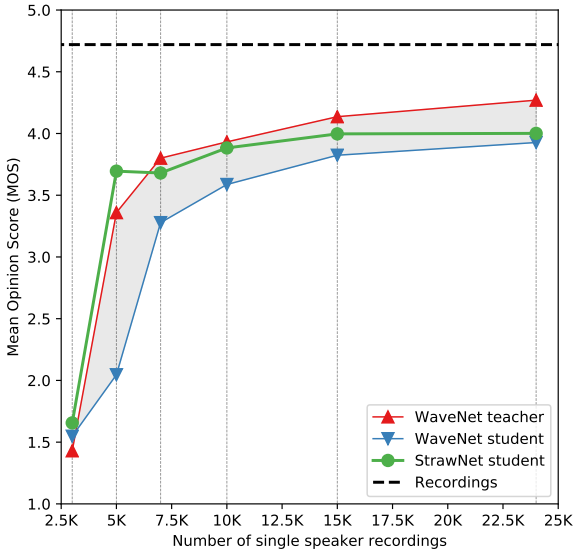

Figure 2: *Variation of MOS with the number of input recordings, for three different models: WaveNet teacher, WaveNet student and StrawNet student.*

Table 1: *Numerical comparison of models for different dataset sizes. Preference scores for StrawNet student over WaveNet student are all statistically significant with a p-value* ≤ 0.01

Subset size	WaveNet teacher	WaveNet student	StrawNet student	% Preference StrawNet
3k	1.43 ± 0.06	1.55 ± 0.04	1.66 ± 0.04	58.5%
5k	3.36 ± 0.07	2.05 ± 0.05	3.69 ± 0.05	98.7%
7k	3.80 ± 0.06	3.28 ± 0.06	3.68 ± 0.05	83.2%
10k	3.93 ± 0.06	3.59 ± 0.05	3.88 ± 0.05	72.7%
15k	4.14 ± 0.05	3.82 ± 0.05	4.00 ± 0.05	69.4%
24k	4.27 ± 0.04	3.93 ± 0.05	4.00 ± 0.05	55.5%

5. Results and discussion

In this section, we present the results of our experiments comparing StrawNet to the baseline WaveNet. We present results for single-speaker training first, followed by the results of the two-speaker training scenario.

5.1. Single-speaker experiments

Figure 2 shows graphically the MOS for models trained on subsets of single-speaker recording data of various sizes. We show the results for the autoregressive WaveNet teacher, feedforward WaveNet student and feedforward StrawNet student. Table 1 shows the same comparison numerically, where we also show these MOS values with 95% confidence interval, and the side-by-side preference scores for StrawNet student vs WaveNet student. We include a WaveNet teacher in our results because that is the data source for StrawNet. As can be seen from the figure, the better the synthetic data generator, the better the resulting StrawNet, as its training dataset is of higher quality. Audio samples from the three models in our analysis are available at `https://google.github.io/StrawNet/`.

From Figure 2, we can see that there is a remarkable difference between the MOS of speech synthesized from the WaveNet teacher and the WaveNet student. This difference is primarily because of different modelling assumptions of an autoregressive network vis-à-vis a feedforward network.

When the number of recordings is below 5k, the WaveNet teacher is incapable of producing intelligible speech and hence

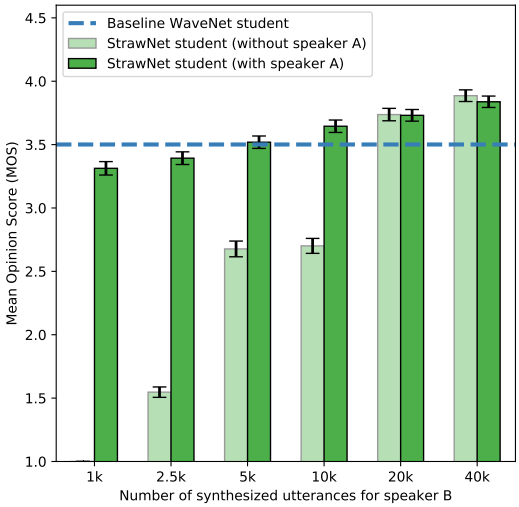

Figure 3: *Variation of MOS for voice quality of speaker B from StrawNet trained with increasing number of synthetic examples. StrawNet can be trained with (light green bar) or without (dark green bar) speaker A.*

Table 2: *Side-by-side preference for speaker B's voice for StrawNet models (trained jointly with speaker A) on different number of synthetic examples. The value in row i and column j represents the percentage of times when model in row i is preferred to model in column j. Preference scores are all statistically significant with a p-value ≤ 0.01, except 20k-vs-10k*.*

vs	1k	2.5k	5k	10k	20k
2.5k	59.8	-	-	-	-
5k	75.0	62.4	-	-	-
10k	77.0	66.2	58.4	-	-
20k	81.1	74.4	66.9	50.1*	-
40k	85.3	77.3	68.6	68.5	61.0

the MOS scores of the WaveNet student, WaveNet teacher and StrawNet student models are all low. The teacher's synthesis quality improves as the number of recordings is increased, whereas the student's voice quality is consistently lower. This reinforces our choice of using the autoregressive WaveNet for synthetic data generation instead of the feedforward student.

From Table 1, we can see that there is a notable preference for StrawNet over the baseline WaveNet student when the number of training recordings is between 5k and 15k. We call this segment of dataset as the "low-data regime" where a Parallel WaveNet fails to learn a good quality voice but the StrawNet approach alleviates that. In this low-data regime, the number of recordings available is plentiful enough to train a good autoregressive teacher network, but not enough to distill a good feedforward student network. Lastly, an interesting observation we can draw from Table 1 is that a StrawNet student model trained on speech synthesized from an underlying AR model with only 5k *recorded* samples can provide a better voice quality than a WaveNet student model trained on 10k *recorded* samples.

5.2. Two-speaker experiments

The comparison of MOS test results for a two-speaker StrawNet with a baseline WaveNet is shown in Figure 3. The autoregressive WaveNet teacher is trained using both speakers. Results are provided for the StrawNet components trained with, and without speaker A. We can see that both the modes are equally effective and can supersede the baseline WaveNet MOS of 3.5, once a sufficient number of synthetic utterances is available. Training a StrawNet on two speakers is beneficial when we have 10k or fewer synthesized utterances. However, for a greater number of synthetic examples, a StrawNet trained with just the synthesized speech from speaker B is found to be slightly better. We hypothesize the reason for this is that mixing synthetized speech with recorded speech creates confusion for the model.

Since the MOS differences in Figure 3 for StrawNet models trained on two speaker subsets of different sizes can appear indecisive, we show side-by-side preference scores in Table 2. We can see from that table, except for the 10k-20k synthetic utterances model pair, a larger synthetic dataset is always significantly better than a smaller synthetic dataset.

We conclude from the findings in this section that StrawNet can achieve superior performance to the baseline WaveNet in a two-speaker training scenario, given a sufficient number of synthetic examples. Although we did not find an upper bound on the amount of synthesized speech beyond which StrawNet's performance starts plateauing, it would be interesting to see the comparison for synthetic datasets beyond 40k examples.

5.3. Additional observations

1. When training StrawNet on a mixture of synthetic speech and recorded speech from the same speaker, we found that the voice quality was slightly worse than when training on only synthesized speech. This is also partly observed when we train a mixture of synthesized speech from speaker B and recorded speech from speaker A in §5.2.
2. Training a different prosody model for each of the subsets in §5.1 causes only a small variation in the MOS (≈ 0.1). To disregard even this little variation due to prosody, we used a common prosody model for all subsets, both for the synthesis step and the final evaluation.
3. The effective number of utterances used for training StrawNet in §5.1 was 40k minus the number of rejected utterances. About 60% utterances were rejected for 5k recordings case, and 5% for 40k, which is consistent with the quality of corresponding WaveNet teachers.

6. Conclusion

We proposed StrawNet – a Self-Training WaveNet that leverages high-fidelity audio generated by an autoregressive WaveNet to generate synthetic dataset used to distill a Parallel WaveNet. We showed that StrawNet can be used to improve the voice quality of speech in low-data regimes, where not enough recordings are available. We argue that this is because the speech synthesized using an autoregressive WaveNet is a good proxy for actual recordings if these are not available. We showed that the voice quality is enhanced when we increase the amount of synthetic training data.

StrawNet could be a useful technique for developing TTS for low-resource languages. Future work would be to compare the performance of StrawNet against a fine-tuned multi-speaker WaveNet model. Finally, an interesting application of StrawNet would be to improve prosodic variety by training on synthetic speech generated with paced or expressive prosody.

7. Acknowledgement

The authors would like to thank the Google TTS Research team and DeepMind, in particular Chun-an Chan, Heiga Zen and Norman Casagrande for their invaluable help and advice.

8. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1996, pp. 373–376.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1315–1318.

[4] 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.

[5] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 3918–3926.

[6] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and V. Klimkov, "Effect of data reduction on sequence-to-sequence neural TTS," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 7075–7079.

[7] D. McClosky, E. Charniak, and M. Johnson, "Effective self-training for parsing," in *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 152–159.

[8] Y. He and D. Zhou, "Self-training from labeled features for sentiment analysis," *Information Processing & Management*, vol. 47, no. 4, pp. 606–616, 2011.

[9] H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, and T. Zhao, "Self-training for unsupervised neural machine translation in unbalanced training data scenarios," *arXiv preprint arXiv:2004.04507*, 2020.

[10] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[11] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[12] B. Schuller and F. Burkhardt, "Learning with synthesized speech for automatic emotion recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5150–5153.

[13] R. Legenstein, C. Naeger, and W. Maass, "What can a neuron learn with spike-timing-dependent plasticity?" *Neural computation*, vol. 17, no. 11, pp. 2337–2382, 2005.

[14] V. Wan, C.-a. Chan, T. Kenter, J. Vit, and R. Clark, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *International Conference on Machine Learning*, 2019, pp. 3331–3340.

[15] P. Ebden and R. Sproat, "The Kestrel TTS text normalization system," *Natural Language Engineering*, vol. 21, no. 3, pp. 333–353, 2015.

[16] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural models of text normalization for speech applications," *Computational Linguistics*, vol. 45, no. 2, pp. 293–337, 2019.

[17] H. Zen, "An example of context-dependent label format for HMM-based speech synthesis in English," *The HTS CMUARCTIC demo*, vol. 133, 2006.

[18] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in neural information processing systems*, 2016, pp. 4743–4751.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, 2015.

[20] M. Podsiadlo and V. Ungureanu, "Experiments with training corpora for statistical text-to-speech systems," in *Interspeech 2018*, 2018, pp. 2002–2006.

[21] C. Chambers, A. Raniwala, F. Perry, S. Adams, R. R. Henry, R. Bradshaw, and N. Weizenbaum, "FlumeJava: easy, efficient data-parallel pipelines," *ACM Sigplan Notices*, vol. 45, no. 6, pp. 363–375, 2010.