Submission from CMU for Blizzard Challenge 2018

SaiKrishna Rallabandi, Pallavi Baljekar, Peter Wu, Evangelia Spiliopoulou and Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

{srallaba, pbaljeka, peterw1, espiliop, awb}@cs.cmu.edu

Abstract

In this paper we present the entry from CMU to Blizzard speech synthesis challenge 2018. We begin with a description of build process for our base voice. We then present the following modifications to base voice: (1) Since the data is chosen from children's stories, we employ Rhetorical Structure Theory to obtain relationships between sentences. We specifically model the contrastive relationship between the sentences within a paragraph. (2) The original speaker attempts to use different ways of speaking depending on character and the situation in the story. To model this, we condition our acoustic model on the character and quote type information. (3) For improving the voice quality we present 'segmental wavenet' - a variant of the popular autoregressive framework Wavenet.

Index Terms: speech processing, convolutional neural networks, strength of excitation, classification, emotion

1. Introduction

Blizzard speech synthesis challenges were devised to better understand different corpus driven speech synthesis techniques on a common dataset. The current challenge is a continuation from the previous year and is aimed at synthesizing children‘s audiobooks. Our submission to this year‘s challenge was based on statistical parametric speech synthesis. There have been continuous and significant improvements in all aspects of this framework of speech synthesis from textual representations through post filtering. [1] has developed generic methods to enable the usage of distributional analysis of text at phone, word and character level in an unsupervised fashion. These techniques have been utilized in building systems both for new languages [2] as well as improving the models for existing languages [3, 4]. The world of acoustic modeling has witnessed the advent of neural models [5, 6, 7, 8, 9] while [10, 11] have proposed the use of alternatives. [12] investigates the incorporation of filled pauses and [13, 3, 14] present techniques to accomplish better phrase break prediction. [15, 16] investigate the postfilters.

In this paper, we present our submission to this years ́ evaluation. Broadly, we have investigated the following approaches:

- Since the data is chosen from children's stories, we employ Rhetorical Structure Theory to obtain relationships between sentences. We specifically model the contrastive relationship between the sentences within a paragraph.
- The original speaker attempts to use different ways of speaking for different characters in the story. To model this, we condition our acoustic model on the character and quote type information.
- For improving the voice quality we present 'segmental wavenet' - a variant of the popular autoregressive framework Wavenet [17].

The rest of this paper is organized as follows: We begin with a description of our base voice in section 2. We then present various approaches we investigated in section 3. This is followed by evaluation results and discussion. We present the evolution of our system and conclude this report.

2. BaseVoice

In this section, we describe our base voice built using the CLUSTERGEN statistical parametric framework [18]. In brief, our system predicts acoustic vectors on a per frame basis based on models that use phonetic, metrical and prosodic contexts. The predicted vectors are then passed through an MLSA filter to generate speech. In addition, we also perform the following:

- Improving the labels: We try to improve upon the labels originally obtained using forced alignment by an EHMM.
- Removal of Outliers and pruning frames: We perform data pruning at two stages. First, we remove the outlier examples using duration as the prosody feature so that we obtain reliable prosody models with high correlation. We then also perform pruning using spectral distortion after the base voice is built using a development.
- We employ Random Forest based acoustic model as the averaging effect of individual decision tree based models has been shown to improve the MCD.

2.1. Data

The database used was provided by Usborne Publishing Ltd. and consists of the utterances from childrens audiobooks spoken by a native British female speaker. We are given about 5 hours of speech data. We have removed the 'bell' sounds which were present in the speech and all the other expressions like 'uh..', 'hm..'. The total duration of the audio is approximately 4.5 hours after segmentation.

2.2. Tokenization and G2P

We consider any text entry separated by white space as a token. From the training data, we have observed that there are instances where the calendar entries such as 1859 are represented both in numeric form as well as the expanded form(eighteen hundred and fifty nine). Further, we noticed that tokens like hyphen played a varied role in the pronunciation of the accompanying word(s). For example, here are the different instances where hyphen was used:

- To indicate hesitation in speaking or transient sounds. This can be observed in the pronunciations for words 'S-s-sorry', 'Cr-r-rock', 'W-w-what'.
- To indicate repitition: 'tap-tap-tap', 'glug-glug-glug'.
- As a placeholder joining two words: broken-hearted bulls-eye chinny-chin-chin

- As a phrase break: 'You cant pretend to not know John Canty - your own father'

We have used a decision list based disambiguator built following [19] to tokenize such occurrences appropriately. Once tokens are obtained, we have analysed the usage of two different phonesets: US phoneset which performed G2P mapping based on CMU pronunciation dictionary and UK phoneset which performed G2P mapping based on Unilex pronunciation dictionary. As we did not find striking differences in the final voice quality between the two phonesets, we have continued the system design with US phoneset.

2.3. Pronunciations for OOV words

There were words in the training corpus which were absent from the CMU pronunciation dictionary. Most of these words were observed to be proper nouns and therefore, we concluded that it would be better to build a generic model to predict the pronunciations for such words. For this, we have employed word to phone mapping [20] using automatic epsilon scattering method [21]: We first use epsilon scattering method to align the letters and phones for a set of words in the given database. Each letter is assumed to be specifying a phonetic correspondence to one or more phones and in case a letter is not mapped to any phone then epsilon is used as a placeholder. We first aligned the letter (graphemic) and phone sequences by estimating the probabilities for one letter (grapheme) G to match with one phone P. We then used string alignment to introduce epsilons maximizing the probability of the alignment path of that word. Once all the words have been aligned, the association probability is calculated again and this is repeated until convergence. Once a reliable alignment has been obtained, we use a statistical mapping from letters to phones which can be seen as maximizing the expression:

$$\prod_{i,j \in S,W} Prob(s_i|w_j) \tag{1}$$

for each word w where $w \in W^d$ is the word in the database with a vocabulary(W) of size d.

2.4. Improving the labels

Table 1: *Analysis of Improving the labels*

Pass	No. of Moves	MCD	F0 Error	Duration Error
1	73898	4.671	28.829	0.943
2	63587	4.646	28.736	0.943
3	57072	4.634	28.814	0.943
4	53376	4.641	28.795	0.946
5	51033	4.636	28.835	0.948
6	48881	4.627	28.823	0.946
7	48460	4.626	28.911	0.946
8	46013	4.618	29.022	0.946
9	44776	4.620	28.926	0.947

We perform text segmentation of the utterances at the segment level (phone). However, all our subsequent analyses are carried out at a lower level, which is realized by dividing each phone into three states, corresponding to the begin, middle and end states of a phoneme. Therefore, each frame is labeled as one of these states and these initial labels for the data are obtained using EHMM technique. We then tried to improve the labels using the procedure outlined in [22]: We examine each segment boundary and consider moving it forward or backward (by one frame) and investigate whether this decreases the distance between original and predicted frame. This process is performed over all the labels and then the models are rebuilt. The distance is measured in terms of unnormalized MCD including the energy coefficient but not the deltas. We have performed 10 iterations over the entire database as the improvement in MCD stopped at that point. The results of this procedure have been outlined in the table 1. We have observed that the passes did not necessarily result in an improvement in the prosodic models.

2.5. Acoustic Feature Extraction

For each of the states obtained from segmentation, we extract acoustic feature vectors over a 5ms frames obtained by applying a hamming window. Spectral representation that we use is MCEPs and were extracted using the SPTK toolkit [23]. The order of MCEP was chosen to be 24 with a frequency warping factor of 0.42 and a small value (1.0E-08) was added to the periodogram. For F0, we interpolate between unvoiced section ensuring breaks during silences and then apply a post smoothing using a 25 ms window.

2.6. Outlier Removal

In the context of audiobook synthesis, selection of appropriate examples for building the data driven statistical models is necessary as the statistics may be skewed due to the presence of outliers. We perform this based on the state durations. For each state, we remove the examples that have values farther than 1.5 times the standard deviation of the mean value for the state.

2.7. Acoustic Modeling

For our current submission we used Random Forest [10] as the model for learning a mapping from the linguistic features to the acoustic features. The central idea is based on feature bagging - to replace the original MCEP prediction tree in the CLUSTERGEN framework with multiple prediction trees trained using random linguistic features. For this, we built 20 different trees for each state, by varying the probability of each feature being picked. Then, to form a forest, we average the predicted values from the trees. Based on the observations from [10], we pick the best 3 trees based on the MCD on a held out development set. Predictions at test time are made by averaging the predictions from the selected individual regression trees.

2.8. Pruning Frames

In addition to the outlier removal mentioned in section 2.6 which was performed using state duration, we also perform a frame pruning based on the spectral features and remove the frames that have the predicted values farther than a predetermined threshold value. These frames correspond, in general to the areas where the model consistently makes mistakes. After this, we rebuild the final voice with the pruned frames.

3. Experiments

3.1. Identifying quoted speech type and characters from stories

The data provided for building voices consists of abridged plays such as ‘Androcles and the Lion’. In other words, the data is a continuum of discourse that runs between characters in the play. Subsequently, the original speaker attempts to imitate the persona of characters while recording the content. This attempt is manifested in the form of prosodic variations in the provided recordings. Therefore, we hypothesize that it is beneficial to tag the data with the speech type (quoted vs narrated), character identity [24] and use this information during acoustic modeling [25].

3.1.1. *Quoted vs Non quoted speech*

We define a portion of story as quoted if it is quote annotated. In addition, we have also annotated if the portion of text was a continuation from previous sentence or a new one. We have annotated these segments using SABLE and an example is shown below:

```
<QUOTE TYPE="NEW"> It's my daughter, Hermia,
 </QUOTE> he explained.


 <QUOTE TYPE="CONT"> I want her to marry
 this man, Demetrius. </QUOTE>
```

An informal inspection of the text has not revealed a significant number of nested quotes leading to ‘story within a story’. This might be because the provided content is aimed at children. Therefore, we have not specifically annotated nested quotes. In scenarios where we did encounter them, we have split the sentence into different utterances. During acoustic modeling, we use this information as another label.

3.1.2. *Identifying character type in stories*

Associating each utterance to a character provides a way to render the story mimicking the characters. This is essentially dealt as a Named Entity Recognition task [24]. Additional linguistic information was used to identify the proper names. In our approach, we borrow this idea but confine ourselves to a maximum of three characters. Nominally, we associate the three characters to (1) Narrator (2) Protagonist and (3) Antagonist. Through an analysis of the text, we have come up with the following basic approach for assigning character labels to text:

- Text without ‘quote’ attribute is labeled by the tag ’Narrator’.
- For every quoted utterance after the narrator, we alternate between characters ’Protagonist’ and ’Antagonist’ if the quote is labeled ’NEW’.
- If the quoted text is labeled as continuation (’CONT’), we repeat the label of the character.
- If there is a sequence of more than three utterances tagged as ‘Narrator’, we drop the speaker state and label the next encountered character with the tag ‘Protagonist’.

We have annotated these segments using SABLE and an example is shown below:

```
<CHAR TYPE="NARRATOR">
Just then, an angry man burst into
the Great Hall with
 three other people.
 </CHAR>

 <CHAR TYPE="PRO">
 I must see Duke Theseus now!,"
 </CHAR>
  <CHAR TYPE="NARRATOR">
  he shouted. </CHAR>


 <CHAR TYPE="ANT">
 Is that you, Egeus?"
 </CHAR>
  <CHAR TYPE="NARRATOR">
 asked Theseus.
 </CHAR>
```

We have observed that the stories differ in the consistency of speaker - speech relationships. However, in our current submission we have followed the rudimentary approach described above and have not specifically handled this inconsistency.

3.2. Identifying inter sentential events and intra-sentential relations

Stories are often characterized by flow of emotions. In addition to mimicking the characters, the original speaker has also attempted to render the perceived emotions while recording the content. Similar to characters, the manifestation of these emotions too is in the form of prosodic variations. Therefore, we hypothesize that we can model the flow of emotions in a story by modeling the prosody of reader. In our current submission, we investigate two approaches for realizing this:

- We identify semantic units (or) events that indiciate change of state within the sentences.
- We employ Rhetorical Structure theory to identify contrastive relationship between different sentences within the paragraph.

3.2.1. *Event Detection within sentences*

We define events as the semantic units that express a change of state or an action in world. Events are comprised of the predicate denoting the action (usually a verb or a noun) and a set of arguments: entities that act on the predicate (agent) or that the predicate acts on them (patient, theme). Conforming to this definition, event detection is an information extraction task where, given a sentence, we try to automatically detect the predicate and the event arguments.

Because of their rich structure, events are good semantic representations that provide useful information for the prosody model. Most times, predicates contain the most important information the speaker wants to convey in an utterance, something crucial for the prosody model. Although predicates are the main information carrier, event arguments also provide important information. Mostly prominent in dialogue or story-telling scenarios, event arguments might carry supplemental information for a previous utterance. In such cases, the speaker wants to focus on those entities more than the action, which shows the importance of Event Detection and comparison of events across utterances.

To identify events, we consider each sentence independently. For each sentence we provide a list of actions and a set of participating entities, which we use as features in our prosody model. Given the lack of annotated data, our Event Detection system is a primarily rule-based system. Our system uses the Stanford CoreNLP parser [26] in order to generate a list of candidate verbal and nominal events per sentence. Then we map each of those candidates to a frame provided by FrameNet [27], which represents the semantic type that a word belongs to. Finally, we use a curated subset of FrameNet frames that represent events in order to determine whether or not the mention is an event. In order to extract the event arguments, we use the Dependency Graph in combination with the NER model provided by Stanford CoreNLP.

3.2.2. *Identifying intrasentential relations using Rhetorical Structure Theory*

Discourse theory describes the high level organization of speech and text. Specifically, hierarchical discourse representations such as Rhetorical Structure theory (RST) provide tree shaped parsing of a story that can be used for prosody modeling. In simple words, given spans of text, RST describes the relationship between them. We hypothesize that identifying contrastive rhetoric and emphasizing the contrast leads to soulful synthesis. For this, we use an approach inspired by [28]: We first learn projection function that learns a mapping from surface level representation to the discourse label on a gold set of discourse labels [29]. We then apply the projection function on current data to obtain discourse labels for each utterance in the story. However, our implementation differs from [28] in a couple of ways: (1) We use the latent representation obtained using a Recurrent Neural net based language model instead of lexical features. This allows us to also bypass the shift reduce parse based implementation[30]. (2) The projection function we learn is non linear instead of a linear function. The steps we followed are outlined below:

- We segment the story into different Elementary Discourse Units (EDUs). Instead of using sequential data labeling [31], we mark each utterance as a different EDU while parsing stories.
- We build an LSTM based language model on WSJ corpus. We then pass the entire story corpus through trained LM, reinitializing the hidden state after every story. For each utterance, we obtain the hidden representation. In our implementation, the RNNLM had a single hidden layer with 512 units and achieved a test perplexity of 5.32 on WSJ corpus. Thus for each utterance in the story, we end up with 512 dimensional representation.
- We then learn a projection function that maps the latent representation of the utterance to its discourse label.

Informal listening evaluation of original recordings revealed a stark variation in prosody of contrastive sentences. Therefore, for the current implementation our projection function is a binary classifier that learns to distinguish two relations: Antithesis (contrast) and Not contrast.

3.3. Segmental Wavenet

Autoregressive models such as WaveNet and WaveRNN [32] have proven that neural models can be used as vocoders at the raw waveform level. Consequently, there has been a rise in the interest of research community towards such approaches [33, 34, 35, 36]. However, these models still need long time spans to train and are computationally expensive. Since we would like to experiment rapidly, we have formulated a neural vocoder at the segmental level for this challenge.

3.3.1. *Formulation of Regular WaveNet*

Wavenet [17] is an autoregressive neural model with a stack of 1D convolutional layers that is capable of directly generating the audio signal. It has been shown to generate speech that rivals natural speech when conditioned on predicted mel spectrum [37]. The input to Wavenet is subjected to corresponding gated activations while passing through each dilated convolutional layer and is classified by the final softmax layer into a μ law encoding. The concrete form of the gated activation function is given by following equation:

$$z = tanh(W_f * x) \odot \sigma(W_f * x) \tag{2}$$

where x and z is the input and output to the activation, respectively. W represents a convolution weight. The subscripts f and g represent a filter and a gate, respectively. The joint probability of a waveform **X** can be written as

$$P(X/\lambda) = \prod_{t=1}^{t} P(x_t \| x_1, x_2..x_{t-1}, \lambda) \tag{3}$$

given model parameters λ.

3.3.2. *Formulation of Segmental WaveNet*

An implementation hack we have used to be able to save GPU memory while training regular WaveNet was to limit the number of timesteps being learnt by the model at each update step. An evaluation on Arctic AWB voice revealed that the model was able to generate speech even if it was exposed randomly to 2000 time steps. Therefore, we have investigated the approach of formulating the model at the level of individual phonemes. To strengthen the learning, we have used full context linguistic features as global conditioning to the model.

$$P(X/\lambda) = \prod_{t=1}^{t} P(x_t \| x_1, x_2..x_{t-1}, L, \lambda) \tag{4}$$

where L is the full context label serving as global conditioning. The steps followed for this are outlined below:

- Extract WORLD representation [38] for each waveform at 5 msec.
- Quantize the waveform using mu law quantization.
- Using the aligned labels, obtain the quantized waveforms and WORLD coefficients for each phoneme.
- During training build an autoregressive model for the quantized waveforms with WORLD coefficients as the local conditioning and full context linguistic features as the global conditioning. This can be seen as similar to 'target cost' in unit selection based systems.
- During inference, generate the waveforms for each phoneme using global conditioning and then join them using WSOLA [39].

We have used a batch size of 16 and an upsample convolution block with 4 layers that upsample at {2,2,4 and 5} times respectively. To avoid disproportionate zero padding which might occur (for example, if a fricative and a vowel are in the same batch), we restrict each batch to have only identical phonemes. Alternatively, it might be possible to sort the waveforms by length of segments before batching. However, we have not explored this approach. Although a model trained this way without global conditioning might be able to generate the entire wavefile during inference, we found global conditioning to be helpful in the model training better. In addition, not having to generate whole file at once also led to faster generation during inference. We have also trained a pretrained variant of this model: We first train regular WaveNet without global conditioning for 200K updates and then switch to Segmental version. However, we have not found any particular improvements with the pretrained model. Hence we have not used this model in our submission.

4. Evaluation

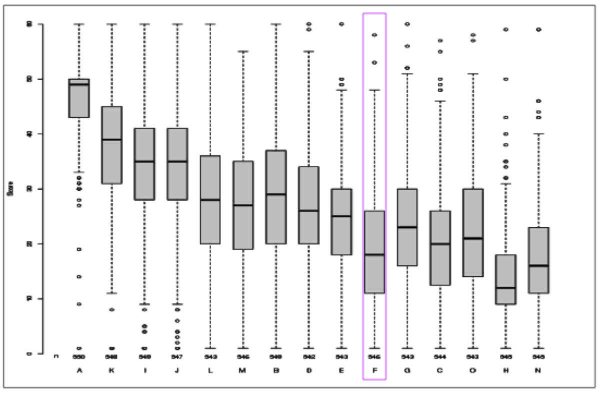

Figure 1: *MOS Scores for all listeners - Overall Impression*

The subjective evaluation was conducted based on various categories: pleasantness, speech pauses, stress, intonation, emotion, listening effort. The identifier of our system is F. Mean opinion score of our system as provided by all listeners is depicted in figure 1.

4.1. Discussion of Results

We are ranked in the last quartile in current evaluation. Informal evaluations of our submission revealed that our prosody model still lacks the necessary rendering required for an interesting automatic story teller. Even though we have used RST based labels for training prosody, our current implementation tracks just simple relation between the utterances: contrast. We believe that incorporating other relations appropriately would result in a much better system. Another possible reason for pretty ordinary performance of our prosody model might be the modular fashion of training. It might be more beneficial to train the models in an end to end fashion instead of a two step approach that we currently employ. An interesting observation was made about our vocoder. While the segmental variant of autoregressive formulation seems to function, and seem to be having an analogue with unit selection based approaches in terms of target cost, we seem to be ending up with power differences for different segments within same utterance during inference. Thus we seem to be ending up with the worst of both autoregressive and unit selection worlds. We plan to investigate this approach in more detail to circumvent this limitation going forward.

4.2. Evolution of our System

We have been consistently in the last quartile in each of the evaluations over the past decade. A postmortem of our performance last year led to the following observations: (1) Our vocoder has not been evolving: We have been using basic mel log filter with mixed excitation as our speech representation. (2) We were using a non neural acoustic model: one based on Random Forests. While decision trees and Random Forests are very robust with respect to acoustic modeling in general, neural approaches have shown to be more flexibility and expressive. (3) Our prosody models were restricted to a single sentence whereas prosody is a suprasegmental feature. In addition, we had a very basic architecture of voice building without modules such as postfilter, etc. This simplicity was a requirement since our voices have been in deployment in real time scenarios. However, we have realized that the price we have been paying for such real time speed was in terms of quality. Therefore, we have started an exclusive research pipeline last year. We have made the following modifications since:

- We have switched our speech representation from basic log mel based representation to WORLD.
- While not deployed for current submission, we have evolved our acoustic models into fully neural based implementations.
- We now have neural prosody models that span across sentences and use richer low dimensional dense representations.
- We are actively experimenting with variants of autoregressive models to strengthen our vocoding.
- We are aggressively investigating end to end based approaches for realizing interesting speech synthesis applications.

5. Conclusion

In this paper we have presented the entry from CMU to Blizzard speech synthesis challenge 2018. We have made these modifications to our previous submission: (1) Since the data is chosen from children's stories, we have employed Rhetorical Structure Theory to obtain relationships between sentences. We have specifically modeled the contrastive relationship between the sentences within a paragraph. (2) The original speaker attempts to use different ways of speaking for different characters in the story. To model this, we have conditioned our acoustic model on the character and quote type information. (3) For improving the voice quality we have presented 'segmental wavenet' - a variant of the popular autoregressive framework Wavenet. The evaluation results highlight that we have a lot of scope to improve our models. We are actively pursuing approaches to strengthen our voice building framework. We believe we will have a much stronger framework and hence a more competitive submission in the coming evaluations.

6. References

[1] O. Watts, "Unsupervised learning for text-to-speech synthesis," Ph.D. dissertation, University of Edinburgh, 2012.

[2] M. Ekpenyong, E.-A. Urua, O. Watts, S. King, and J. Yamagishi, "Statistical parametric speech synthesis for ibibio," *Speech Communication*, vol. 56, pp. 243–251, 2014.

[3] A. Vadapalli and K. Prahallad, "Learning continuous-valued word representations for phrase break prediction," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[4] S. K. Rallabandi, S. S. Rallabandi, P. Bandi, and S. V. Gangashetty, "Learning continuous representation of text for phone duration modeling in statistical parametric speech synthesis," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 111–115.

[5] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural network employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[6] X. Wang, S. Takaki, and J. Yamagishi, "A simple rnn-plus-highway network for statistical parametric speech synthesis," 2017.

[7] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, "Deep neural network context embeddings for model selection in rich-context HMM synthesis," in *Proc. Interspeech*, Dresden, September 2015.

[8] S. Achanta, T. Godambe, and S. V. Gangashetty, "An investigation of recurrent neural network architectures for statistical parametric speech synthesis." in *INTERSPEECH*, 2015, pp. 859–863.

[9] S. Ronanki, S. Reddy, B. Bollepalli, and S. King, "DNN-based Speech Synthesis for Indian Languages from ASCII text," in *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*, Sunnyvale, CA, USA, Sep. 2016.

[10] A. W. Black and P. K. Muthukumar, "Random forests for statistical speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on gaussian process regression," *IEEE journal of selected topics in Signal Processing*, vol. 8, no. 2, pp. 173–183, 2014.

[12] R. Dall, M. Tomalin, and M. Wester, "Synthesising Filled Pauses: Representation and Datamixing," in *Proc. SSW9*, Cupertino, CA, USA, 2016.

[13] O. Watts, S. Gangireddy, J. Yamagishi, S. King, S. Renals, A. Stan, and M. Giurgiu, "Neural net word representations for phrase-break prediction without a part of speech tagger," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 2618–2622.

[14] O. Watts, J. Yamagishi, and S. King, "Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 2157–2160.

[15] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-based post-filter for gmm-based voice conversion," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–4.

[16] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 2003–2014, 2015.

[17] A. Van Den 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.

[18] A. W. Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling," in *Ninth International Conference on Spoken Language Processing*, 2006.

[19] D. Yarowsky, "Homograph disambiguation in text-to-speech synthesis," in *Progress in speech synthesis*. Springer, 1997, pp. 157–172.

[20] N. K. Elluru, A. Vadapalli, R. Elluru, H. Murthy, and K. Prahallad, "Is word-to-phone mapping better than phone-phone mapping for handling english words?" in *ACL (2)*, 2013, pp. 196–200.

[21] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," 1998.

[22] A. W. Black and J. Kominek, "Optimizing segment label boundaries for statistical speech synthesis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3785–3788.

[23] S. W. Group *et al.*, "Speech signal processing toolkit (sptk)," *h ttp://sp-tk. sourceforge. net*, 2009.

[24] J. Y. Zhang, A. W. Black, and R. Sproat, "Identifying speakers in children's stories for speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[25] M. Theune, S. Faas, A. Nijholt, and D. Heylen, "The virtual storyteller," *ACM SigGroup Bulletin*, vol. 23, no. 2, pp. 20–21, 2002.

[26] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[27] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ser. ACL '98. Stroudsburg, PA, USA: Association for Computational Linguistics, 1998, pp. 86–90. [Online]. Available: https://doi.org/10.3115/980845.980860

[28] Y. Ji and J. Eisenstein, "Representation learning for text-level discourse parsing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 13–24.

[29] V. W. Feng and G. Hirst, "Text-level discourse parsing with rich linguistic features," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 60–68.

[30] D. Marcu, "A decision-based approach to rhetorical parsing," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 365–372.

[31] H. Hernault, D. Bollegala, and M. Ishizuka, "A sequential model for discourse segmentation," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2010, pp. 315–326.

[32] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[33] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *Proc. Interspeech*, vol. 2017, 2017, pp. 1118–1122.

[34] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 712–718.

[35] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," in *Proc. Interspeech*, vol. 2017, 2017, pp. 1138–1142.

[36] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, "Fast wavenet generation algorithm," *arXiv preprint arXiv:1611.09482*, 2016.

[37] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[38] 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.

[39] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 554–557.