INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Syllable-level representations of suprasegmental features for DNN-based text-to-speech synthesis

Manuel Sam Ribeiro[1], *Oliver Watts*[1], *Junichi Yamagishi*[12]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]National Institute of Informatics, Tokyo, Japan

m.f.s.ribeiro@sms.ed.ac.uk, owatts@inf.ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

A top-down hierarchical system based on deep neural networks is investigated for the modeling of prosody in speech synthesis. Suprasegmental features are processed separately from segmental features and a compact distributed representation of high-level units is learned at syllable-level. The suprasegmental representation is then integrated into a frame-level network. Objective measures show that balancing segmental and suprasegmental features can be useful for the frame-level network. Additional features incorporated into the hierarchical system are then tested. At the syllable-level, a bag-of-phones representation is proposed and, at the word-level, embeddings learned from text sources are used. It is shown that the hierarchical system is able to leverage new features at higher-levels more efficiently than a system which exploits them directly at the frame-level. A perceptual evaluation of the proposed systems is conducted and followed by a discussion of the results.

Index Terms: speech synthesis, prosody, deep neural networks, suprasegmental representations

1. Introduction

Statistical parametric speech synthesis (SPSS) has seen improvements over recent years, especially in terms of intelligibility [1]. Synthetic speech is often clear and understandable, but it can also be bland and monotonous. Therefore, proper modeling and generation of natural speech prosody is still considered to be a largely unsolved problem [2]. This is relevant especially in the context of expressive audiobook or conversational speech synthesis, where speech is expected to be fluid and captivating.

A clear understanding of prosody is essential for achieving good prediction at synthesis time. In general, prosody can be seen as a layer that lies on top of the segmental sequence. Listeners can perceive the same melody or rhythm in different utterances, and the same segmental sequence can be uttered with a different prosodic layer to convey a different message. For this reason, prosody is commonly accepted to be inherently suprasegmental [2, 3, 4]. It is governed by longer units within the utterance (for example, at syllable, word, or phrase levels), and beyond the utterance, at discourse-level [3].

However, common techniques for the modeling of speech prosody – and speech in general – operate mainly on very short intervals, either at the state or frame level, in both hidden Markov model (HMM) [5] and deep neural network (DNN) [6] based speech synthesis. Although speech parameter generation algorithms ensure smooth speech-like trajectories, each prediction prior to this is performed independently for each short-term unit without access to the longer-term acoustic context.

In an attempt to leverage the suprasegmental properties of prosody for speech synthesis, earlier work has focused either on modeling, or on the input and output features used. In terms of modeling, multi-level approaches have been proposed for HMM-based systems [7, 8]. In the case of DNN-based systems, recurrent [9], hierarchical [10], or mixed [11] approaches have claimed to capture the long-term dependencies of speech. Recent work has also directed its attention to the output features, proposing acoustic representations that are able to capture longer-term information using various wavelet-based decomposition strategies on the *f0* signal [12, 13].

On the input side, it has been shown that prosodic contexts are very poorly understood. Recent work revealed that, in HMM-based synthesis, features above the syllable-level do not improve the naturalness of synthetic speech [14]. In an effort to acquire a better understanding of linguistic contexts, continuous representations of input features have been explored, either at segmental [15, 16] or word-levels [17, 18], with various degrees of success. In fact, it was observed that systems focusing on the linguistic features have been shown to be less effective in terms of improvements of synthetic speech [1].

This investigation adds to the work exploring input features for prosody modeling in text-to-speech synthesis, specifically focusing on continuous representations of suprasegmental contexts. Two main contributions are made: (1) a top-down hierarchical model that can leverage suprasegmental information and represent it compactly at syllable-level; (2) an investigation of this architecture with a bag-of-phones representation at syllable-level and word embeddings learned from a large text database with the skip-gram model [19, 20].

Section 2 introduces the basic and proposed DNN systems, while section 3 describes the expressive databased used. In section 4, we investigate the size of the embedding layer and in section 5 the new suprasegmental features. In section 6, we conduct a subjective evaluation of the systems. We conclude with a discussion of the results.

2. DNN-based speech synthesis

2.1. Basic DNN

The basic deep neural network is a simple feedforward multilayer perceptron. We use a configuration similar to the baseline system described in [16]. A network with 6 hidden layers is used, each layer consisting of 1024 nodes. We set *tanh* as the activation function in the hidden layers and we use a linear output layer. During training, we use a mini-batch size of 256 and we set the maximum number of iterations to 25.

As output features, we use *log-f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicties (BAPs) at 5 ms intervals, with their respective delta and delta-deltas. *Log-f0*

http://dx.doi.org/10.21437/Interspeech.2016-1034

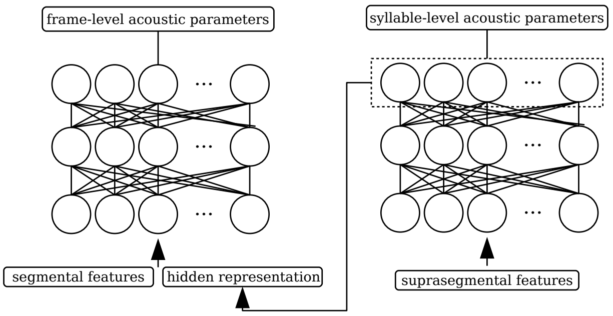
frame-level acoustic parameters
syllable-level acoustic parameters
...
segmental features
hidden representation
suprasegmental features

Figure 1: *Top-down hierarchical deep neural network.*

is linearly interpolated and a binary voiced/unvoiced decision is added to the acoustic feature vector. The complete acoustic feature vector consists of 259 dimensions, which are normalized to zero mean and unit variance.

For these experiments, we use natural duration, which is inferred from the force-alignment given by a pre-trained 5-state left-to-right HMM. As input features, we use the same set of 592 binary questions defined at phone, syllable, and word levels used in [16] plus 2 features defined at state and frame-level. These features are normalized to the range [0.01, 0.99] and, for the basic DNN, segmental and suprasegmental features are concatenated and used as input to the network.

2.2. Hierarchical DNN

Figure 1 illustrates the top-down hierarchical deep neural network. A subset of the above-mentioned 592 binary features which are defined at syllable and word-level (and are here termed *suprasegmental features*) are separated from segmental (phone-level) features. An initial network inputs these features and maps them to acoustic parameters defined at syllable-level. For the current experiments, the acoustic features predicted for each syllable consist of a 259-dimensional vector obtained by averaging the frame-level features over the entire syllable.

The network is set to be a 6 hidden layer triangular network. The lower layers begin with 1024 nodes and this is halved in the next layer such that the top hidden layer reaches the desired dimensionality. Further intuition is detailed in section 4. The syllable network uses the *tanh* activation function in the hidden layers and a linear output layer. Mini-batch size is set to 16 and the maximum number of iterations is set to 25.

After the suprasegmental network is trained, the hidden representation of the bottleneck layer is concatenated with the segmental feature vector. The frame-level network is then trained as described in the previous section.

3. Database

We use expressive audiobook data to conduct these experiments. This type of data is desirable for this type of analysis as the narrator typically records entire chapters sequentially. Higher-level prosodic effects are thus captured in the recorded speech, which allow us to explore suprasegmental effects within the sentence and in the overall discourse. For this work, we have focused on the freely available audiobook *A Tramp Abroad*, available from *Librivox*[1]. The data has been pre-processed according to [21] and [22]. The hand-selected narrated speech described in [22] was used, thus setting aside noisy direct speech data. A training, development, and test set of 4500, 300, and 100 utterances, respectively, was set for the experiments described in this work. A further heldout set of 50 utterances was used for the subjective evaluation described in section 6.

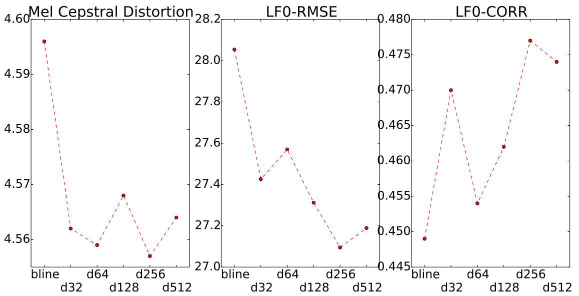

Figure 2: *Objective measures as a function of bottleneck size. 'bline' represents the baseline system, while 'dn' indicates a bottleneck layer of dimensionality n.*

4. Embedding size

In this set of experiments, we observe the effect of bottleneck layer size on objective measures. The main motivation for these experiments, and for the proposed framework, is the hypothesis that *repeatedly adding high-dimensional features to a frame-level network might reduce the impact of suprasegmental effects*. The frame-level network might depend too much on short-term variations and ignore the larger high-dimensional features. We hypothesize that a good balance between segmental and suprasegmental features will return the best results.

We use the framework illustrated in Figure 1, where we replace the suprasegmental features with the learned hidden representations. All experiments use the syllable mean for the acoustic features and the top layer as bottleneck layer. Bottleneck size is varied in powers of 2, from 32 to 512. All syllable-level systems are triangular networks with 6 hidden layers. The lower layer has 1024 nodes and we either maintain that size or we reduce it in half until we reach the desired dimensionality in the top hidden layer. As an example, the triangular network for a bottleneck layer of size 256 has the following structure, in terms of layer size: (1024, 1024, 1024, 1024, 512, 256).

Figure 2 shows layer size effect on mel-cepstral distortion (MCD), *log-f0* RMSE and correlation. All systems using hidden representations outperform the baseline, with the best results occurring with a dimensionality of 256. These results appear to support our hypothesis, as the segmental features use a vector of 350 dimensions. A vector of 256 dimensions for suprasegmental features balances the two types of features and allows the best prediction for all acoustic parameters.

5. Syllable and word-level features

5.1. Syllable bag-of-phones

In this and the following section, we investigate the addition of new features to the frame-level and hierarchical networks. The hypothesis is that *the hierarchical network will be able to leverage the information given by the new features, while the frame-level network will depend mostly on segmental features and ignore the new high-dimensional representations.*

We therefore propose a bag-of-phones representation for syllables, which is essentially an *n-hot encoding*. We use 3 bags of phones, each defined for the onset, nucleus, and coda, and containing phone identity and articulatory features. Tak-

[1] `https://librivox.org`

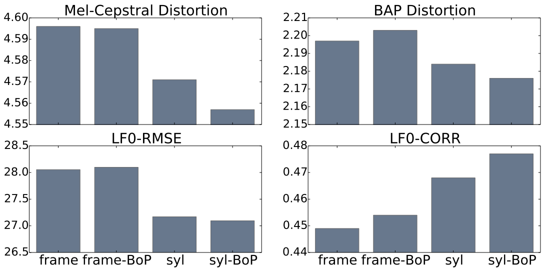

Figure 3: *Objective measures for the four systems testing the proposed* bag-of-phones *representation.*

ing the onset as an example, we define a binary vector where each component indicates either an articulatory features or a phone identity. For all phones belonging to the onset of the current syllable, we activate the respective articulatory and identity component in the onset vector. This approach allows us to define a fixed-size representation of syllables that accounts for the variable number of phones in the onset and coda, while still including all of their defining features. We compare four systems:

frame frame-level deep neural network with all available features

frame-BoP baseline DNN with all features plus bag-of-phones representation for onset, nucleus, and coda of current syllable.

syl suprasegmental features are trained separately at syllable-level using a hidden layer with 256 nodes. Input features to syllable DNN consist of the suprasegmental features used with the baseline system. That is, this system does not use the bag-of-phones for onset, nucleus, and coda.

syl-BoP similar to *syl-noBoP*, except we include the bag-of-phones features in the input to the syllable-level DNN.

Figure 3 plots the results for mel-cepstral distortion, band aperiodicity distortion, *f0* RMSE and correlation. Results indicate that adding the bag-of-phones representation to the frame-level DNN does not affect the results. This follows our initial hypothesis that simply appending more suprasegmental features to a frame-level DNN might limit the impact of those features.

The difference between *frame* and *syl* shows the changes we get by training supra-segmental representations separately. This causes the biggest improvements among the four systems. The difference between *syl* and *syl-BoP* measures the changes caused by the bag-of-phones features when training suprasegmental features separately. In the case of RMSE, we do not see many improvements, but we do notice better results in terms of correlation and MCD. It's quite interesting to observe that, in the BAP case, adding bag-of-phones features to the baseline slightly decreases performance, while adding the same features to a syllable-level DNN slightly increases it. This set of experiments shows that not only is the separate training of suprasegmental feature predictors useful, but it creates a framework that is able to leverage additional features in a more robust manner.

5.2. Word embeddings

We extend the previous investigation to incorporate word-level features. For this task, we use word embeddings learned by a skip-gram model [19, 20] To learn these embeddings, we have used the freely available English Wikipedia data dump from September 2015.[2] This data has been pre-processed and cleaned and we have kept the first 500 million words. Two models were trained on this dataset, one using an embedding size of 100 and another an embedding size of 300. The systems use the publicly available *word2vec* implementation of the skip-gram model with negative sampling and they were trained for 15 epochs with a window of 5 words.

System	MCD	BAP	F0-RMSE	F0-CORR
frame	4.596	2.197	28.054	.449
frame-w100	4.598	2.204	28.048	.448
syl-BoP	4.557	**2.176**	27.095	.477
syl-BoP-w100	**4.55**	2.177	27.086	.463
syl-BoP-w300	4.565	2.178	**26.850**	**.479**

Table 1: *Objective results for word embedding systems.*

We consider five systems, whose identifiers are given in Table 1. The ***frame*** system is the basic frame-level DNN using no additional features. ***frame-w100*** uses 100-dimensional word embeddings and appends them to the input of the basic DNN. The remaining systems use the framework illustrated in figure 1, including the *bag-of-phones* representation described in the previous section. The first model (***syl-BoP***) is trained without word embeddings, and the final two with 100 and 300-dimensional embeddings (***syl-BoP-w100*** and ***syl-BoP-w300***).

Table 1 summarizes the results for each system. Adding word embeddings to the frame-level DNN does not show any improvements over the baseline. This is surprising, as we would expect some improvements, given the work described in [18]. However, the authors in [18] used a carefully annotated non-expressive database for their experiments. They have also used a bi-directional LSTM, while we have used a feedforward DNN. We do not observe any relevant differences in terms of objective measures for the hierarchical systems. However, adding a larger word feature vector allows the system to slightly improve *f0* prediction. This is interesting, as it suggests that higher-dimensional features may be useful to learn more complex relationships between suprasegmental units. In that case, the proposed hierarchical system might be useful in processing them.

6. Subjective evaluation

A listening test was conducted on selected systems described in previous sections. System *syl* pre-processes suprasegmental features separately with bags-of-phones and uses a bottleneck layer with 256 nodes. System *syl-w300* is similar to *syl*, but it adds 300-dimensional word embedding representations to the input of the syllable-level network. The *baseline* system processes suprasegmental features directly. From a held-out set, 50 test utterances were synthesized from the parameters predicted by the frame-level network. 16 native speakers judged randomized utterance pairs for the two conditions in a preference test. Each pair was judged 8 times and each condition received a total of 400 judgements. Results are shown in Table 2. Percentages indicate the overall preference for a hierarchical system over the baseline. Aggregated results for all listeners do not show a preference for either system which is significant with $\alpha = 0.05$ under a binomial test with an expected 50% split.

This is surprising, as listening to the synthesized waveforms informally showed clear differences between systems. Individual participant results indicate that some listeners prefer the hierarchical systems (2, 6, 10, 14) and others prefer the baseline system (7, 11), while some participants do not have a clear preference (1, 12).

[2] `http://dumps.wikimedia.org/enwiki/20150901`

ID	syl	syl-w300
1	48.15%	43.18%
2	60.87%	51.79%
3	59.26%	59.09%
4	56.52%	48.21%
5	53.70%	54.55%
6	63.04%	46.43%
7	38.89%	56.00%
8	56.52%	53.57%
9	44.44%	50.00%
10	65.38%	51.72%
11	42.59%	43.18%
12	52.17%	48.21%
13	55.56%	63.64%
14	65.22%	42.86%
15	46.30%	47.73%
16	45.65%	42.86%
all	53.39%	50.19%

Table 2: *Subjective results.*

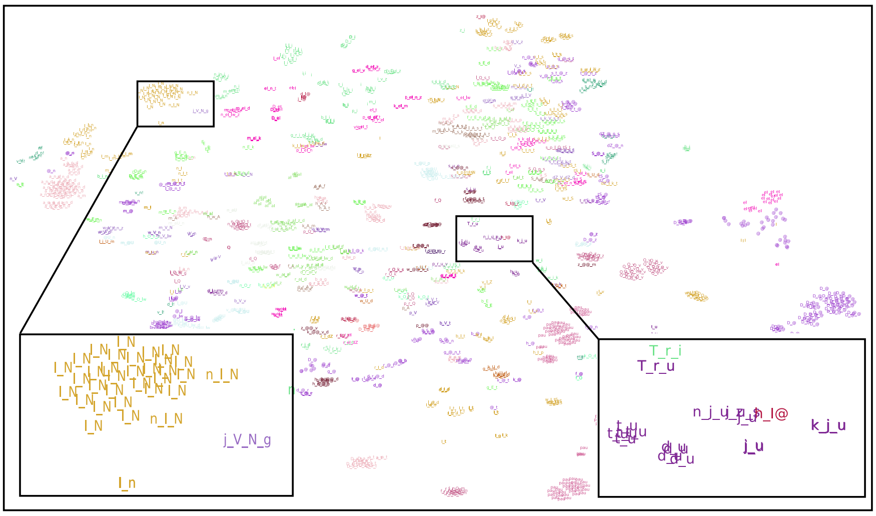

Figure 4: *2-dimensional vizualization of suprasegmental embeddings at syllable-level using t-SNE [23].*

7. Discussion

The results observed in the listening tests were surprising. Objective measurements for the two proposed systems showed statistically significant improvements over the baseline. Clear differences were also perceived when listening to the synthesized speech samples informally.[3] The subjective evaluation, however, showed no overall significant preference, although some participants clearly prefer the proposed method.

In order to understand what listeners are responding to when submitting their judgements, we compare their preferences and the differences between the two systems in terms of objective measures. We compare the *baseline* and *syl* systems. Correlating objective measures with perceptual scores, we observe that there is no significant correlation in terms of mel-cepstral distortion (r=-0.0075, n=50, *ns*) and band-aperiodicity distortion (r=-0.174, n=50, *ns*). However, we do observe a significant correlation in terms of *f0* rmse (r=-0.355, n=50, p<.01) and *f0* correlation (r=0.323, n=50, p<.05). These results show that listeners are judging the utterances in terms of the *f0* signal. This is reassuring, as when learning representations of suprasegmental context, we are essentially focusing on a better understanding of prosody. The proposed systems, therefore, do modify the *f0* signal in a way that affects listeners' preferences.

For a better insight into how the syllable-level DNN manipulates the suprasegmental features, we vizualize the hidden representations learned by the network. For this task, we use t-SNE [23], an efficient technique for dimensionality reduction. We randomly sample 1500 syllable embeddings from the *syl* system. These are then reduced with t-SNE to a two dimensions. The results are plotted in Figure 4. In the figure, two sections are enlarged for clarity. Colors are assigned according to syllable nucleus. We observe that some syllables, which could potentially be different, are closer in the embedding space. This is the case, for example, for *T_r_i* and *T_r_u*. The onset of the syllable seems to be the main similarity between the two samples. On the left-hand section, it appears to be the nucleus and coda the main point of similarity for *I_N* and *n_I_N*.

The syllable-level network can be thought of as a feature extractor for suprasegmental features. Appending new representations of context to a frame-level feature vector could introduce more noise than useful information. This could be the reason why we failed to observe improvements when adding word embeddings to the frame-level network. But pre-processing such features separately, we can learn useful and compact representations of suprasegmental context for frame-level prediction. The *syl* system achieves good accuracy when computing objective measures on the syllable-level parameters (MCD: 4.699, BAP: 1.252, F0-RMSE: 26.717). Although we failed to observe an overall significant preference for the proposed systems, the hierarchical systems are still capable of learning meaningful embeddings of suprasegmental features.

In [10], parallel and cascaded networks are proposed to model *f0* components separately. The top-down hierarchical system described in figure 1 can be thought of as a cascaded network. Future work could explore a parallel integration of segmental and suprasegmental features, rather than a cascaded integration. Similarly, extending this investigation to recurrent systems might be useful.

8. Conclusions

We have investigated a hierarchical top-down system for DNN-based speech synthesis. We observed the best results when using a good balance between segmental and suprasegmental features. A bag-of-phones representation for syllables was proposed, which was tested with word embeddings as additional features in the hierarchical model. When evaluating the hierarchical systems, participants mostly reacted to the *f0* signal, which suggests that we learn representations that could help us predict and control prosodic variations.

Acknowledgements: This research was supported by the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS), the EPSRC Standard Research Studentship (EP/K503034/1), and the EPSRC Programme Grant (EP/I031022/1) - Natural Speech Technology (NST). We thank Toshiba Research Europe Limited (Cambridge Research Laboratory) for the data used in this work and Zhizheng Wu for contributions and insights into DNNs in this and earlier work.

[3] Speech samples can be found in: `http://homepages.inf.ed.ac.uk/s1250520/samples/interspeech16.html`

9. References

[1] S. King, “Measuring a decade of progress in text-to-speech,” *Loquens*, vol. 1, no. 1, 2014.

[2] K. Yu and S. Young, “Continuous F0 modeling for HMM based statistical parametric speech synthesis,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1071–1079, 2011.

[3] A. Wennerstrom, *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press, 2001.

[4] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[5] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden markov models,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[6] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[7] C.-C. Hsia, C.-H. Wu, and J.-Y. Wu, “Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1994–2003, 2010.

[8] Y. Qian, Z. Wu, B. Gao, and F. K. Soong, “Improved prosody generation by maximizing joint probability of state and longer units,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1702–1710, 2011.

[9] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, “Prosody contour prediction with long short-term memory, bidirectional, deep recurrent neural networks,” in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

[10] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, “Modeling F0 trajectories in hierarchically structured deep neural networks,” *Speech Communication*, vol. 76, pp. 82–92, 2016.

[11] S.-H. Chen, S.-H. Hwang, and Y.-R. Wang, “An RNN-based prosodic information synthesizer for mandarin text-to-speech,” *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 3, pp. 226–239, 1998.

[12] M. S. Ribeiro and R. A. J. Clark, “A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Brisbane, Australia, April 2015.

[13] M. S. Ribeiro, O. Watts, J. Yamagishi, and R. A. J. Clark, “Wavelet-based decomposition of f0 as a secondary task for dnn-based speech synthesis with multi-task learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Shanghai, China, March 2016.

[14] M. Cernak, P. Motlicek, and P. N. Garner, “On the (un) importance of the contextual factors in HMM-based speech synthesis and coding,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8140–8143.

[15] H. Lu, S. King, and O. Watts, “Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis,” *Proc. ISCA SSW8*, pp. 281–285, 2013.

[16] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, “Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4460–4464.

[17] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2599–2603.

[18] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, “Word embedding for recurrent neural network based TTS synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2015. ICASSP 2015.*, 2015.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[21] N. Braunschweiler, M. J. Gales, and S. Buchholz, “Lightly supervised recognition for automatic alignment of large coherent speech recordings.” in *INTERSPEECH*, 2010, pp. 2222–2225.

[22] N. Braunschweiler and S. Buchholz, “Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality.” in *INTERSPEECH*, 2011, pp. 1821–1824.

[23] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.