Parallel and cascaded deep neural networks for 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

An investigation of cascaded and parallel deep neural networks for speech synthesis is conducted. In these systems, suprasegmental linguistic features (syllable-level and above) are processed separately from segmental features (phone-level and below). The suprasegmental component of the networks learns compact distributed representations of high-level linguistic units without any segmental influence. These representations are then integrated into a frame-level system using a cascaded or a parallel approach. In the cascaded network, suprasegmental representations are used as input to the frame-level network. In the parallel network, segmental and suprasegmental features are processed separately and concatenated at a later stage. These experiments are conducted with a standard set of high-dimensional linguistic features as well as a hand-pruned one. It is observed that hierarchical systems are consistently preferred over the baseline feedforward systems. Similarly, parallel networks are preferred over cascaded networks.

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

1. Introduction

Over the last decade, statistical parametric speech synthesis (SPSS) has improved considerably in terms of intelligibility [1]. Synthetic speech is often clear and fairly easy to understand. However, the same speech can appear to be bland and monotonous, indicating that how to handle prosody still remains a largely unsolved problem [2]. This is especially relevant when dealing with expressive audiobook or conversational speech data. In these scenarios, synthetic speech is expected to be fluid and natural.

Prosody is a fundamental aspect of human communication. It allows speakers to convey information of a linguistic, non-linguistic, and para-linguistic nature. That information can express dependencies between the various units within the utterance and link them to the overall discourse. Due to these characteristics, it is generally agreed that prosody is inherently suprasegmental [2, 3, 4].

This implies a conceptual division between a segmental layer, operating mostly at the phone-level, and a suprasegmental layer, operating over longer temporal spans, such as the syllable, word, phrase, utterance, and discourse [3]. It should however be noted that this division is not entirely clear-cut. Fundamental frequency (or *f0*), for example, which is often associated with prosodic variation, can be affected at various linguistic levels. Phones can be voiced or unvoiced or have higher or lower *f0* [5]. Syllables can be stressed or unstressed and words may carry different prominence. Utterances may assume *f0* patterns which depend on where they fit in the discourse as a whole [3].

To achieve natural synthesis of speech prosody, a good understanding and representation of higher-level linguistic units is required. Furthermore, the model which generates speech parameters must not only be given useful representations of context at the various linguistic levels, but also be able to exploit them. This is not the case for current techniques in statistical parametric speech synthesis. Most approaches based on hidden Markov models (HMM) [6] or deep neural networks (DNN) [7] still operate over very short intervals, at the level of either the state or the frame. Although speech parameter generation algorithms ensure speech-like trajectories, they in effect give a short-term smoothing rather than global prosodic coherence; predictions from text of neighbouring units are still performed independently without exploiting any good representations of long-term units.

Earlier work has proposed several techniques that attempt to leverage the suprasegmental properties and the long-term dependencies of speech prosody. Various multi-level systems have been proposed for HMM-based speech synthesis. Common approaches focus mostly on modeling the *f0* signal with superpositional [8, 9] or joint [10, 11, 12] systems. For DNN-based synthesis, recurrent [13], hierarchical [14], or mixed [15] approaches have been proposed.

In terms of linguistic features, most work has shown that prosodic contexts are not clearly understood. For HMM-based speech synthesis, traditional features above the syllable-level do not significantly affect the naturalness of synthetic speech [16]. In an effort to acquire a better understanding of linguistic contexts, continuous representations of input features have been explored, either at the segmental [17, 18] or the word [19, 20] level. For

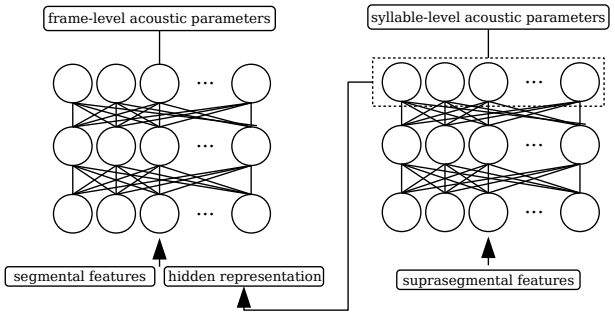

Figure 1: *Hierarchical cascaded deep neural network.*

example, [20] used distributed representations of words learned with the skip-gram model [21, 22] and other variants. It was found that the unsupervised embeddings can be good substitutes for manual annotation of a database.

This work proposes a hierarchical system that learns compact distributed representations of suprasegmental features. An initial higher-level network learns embeddings at the syllable level. These are then integrated into a second lower-level network for the prediction of acoustic parameters at the frame level. Two methodologies for the integration of segmental and suprasegmental features are evaluated: cascaded and parallel.

The architectures we use in this work are closely related to those described in [14] and, for consistency, the same terminology is adopted. However, the work detailed here is different as it focuses on distributed representations of suprasegmental features rather than the superpositional modelling of the *f0* signal. Our core system still operates at the frame level, and jointly models source and spectral parameters.

This paper is organized as follows: section 2 introduces the basic and hierarchical DNN systems, as well as the linguistic features used. Section 3 describes the experiments that were conducted, stating hypotheses and detailing objective and subjective evaluations. We conclude with a discussion of the results in section 4.

2. DNN-based speech synthesis

2.1. Basic network

The basic deep neural network is a simple feedforward multilayered perceptron. We use a configuration similar to the baseline system described in [18]. A network with 6 hidden layers is used, each layer containing 1024 nodes. The hidden layers use *tanh* as the activation function and the output layer uses a linear activation function. For training, a mini-batch size of 256 is set and the maximum number of iterations is set to 25.

For output features, we use *log-f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicities (BAPs) at 5 ms intervals. To these features, we append their respective dynamic features (deltas and deltadeltas). The *log-f0* signal is linearly interpolated and a binary voiced/unvoiced decision is appended to the acoustic feature vector. Therefore, the complete output vector has a total of 259 dimensions, which are then normalized to zero mean and unit variance.

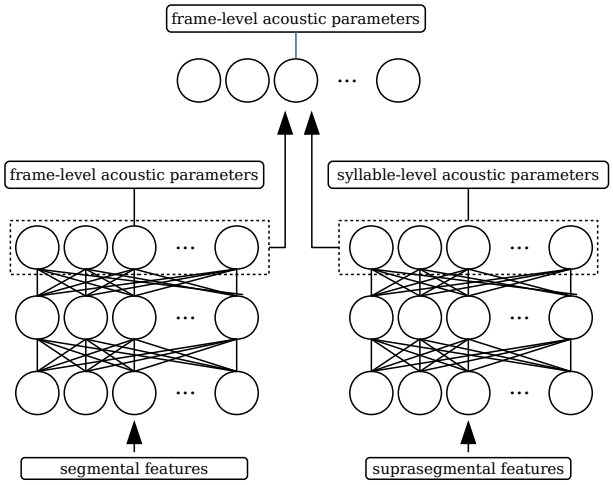

Figure 2: *Hierarchical parallel deep neural network.*

2.2. Cascaded and parallel networks

We define *segmental features* to be those that describe the input at the level of the segment and below, at the phone and frame level. We term features that represent the input at linguistic levels above the segment *suprasegmental features*: features at the syllable, word, phrase, and utterance levels.

In the cascaded and parallel approaches, segmental and suprasegmental features are decoupled and processed separately. In both systems, distributed representations of suprasegmental contexts are learned and later integrated into a frame-level system. An initial suprasegmental network is defined at the syllable-level. This network inputs representations of context at the syllable level and above levels and maps them to acoustic parameters defined at the syllable level. For the current experiments, the output of this network consists of a 258-dimensional vector obtained by averaging the frame-level acoustic features over the entire syllable. The network is set to be a 6 hidden layer triangular network. In terms of layer size, it is defined as (1024, 1024, 1024, 1024, 512, 256). That is, the top hidden layer is a bottleneck layer with 256 dimensions. The hidden activation function is set to be *tanh* and the output layer uses the linear activation function. Mini-batch size is set to 16 and the maximum number of iterations is set to 25.

Figure 1 illustrates the cascaded deep neural network [14], which can be thought of as a top-down hierarchical network. The distributed representation of suprasegmental features is concatenated with the segmental fea-

linguistic level	hand-selected	standard
state	2	
phone	350	
syllable	152	426
word	92	184
phrase	-	211
utterance	-	300

Table 1: *Dimensionality of input features per linguistic level.*

ture vector. A second network is then trained to generate source and spectral parameters at the frame level. Figure 2 illustrates the parallel deep neural network. In this integration strategy, segmental and suprasegmental features are joined at a later stage. The second network inputs only segmental features and its architecture is similar to that of the suprasegmental network. The distributed representation learned from both networks, each with 256 dimensions, is used to drive a single layer network that generates acoustic parameters at the frame level.

2.3. Linguistic features

As input to the deep neural networks, we use a *standard* set of linguistic features. This is the full question set used for tree clustering in HMM-based synthesis. Linguistic contexts obtained through a common front-end such as the one distributed with Festival[1] are defined at phone, syllable, word, phrase, and utterance levels. Questions are defined in terms of quinphone identity, syllable stress or accent, part-of-speech, predicted phrase ToBI labels, or positional information in words, phrases, and utterances. A detailed description of this set can be found in [6]. To these, we add two additional features defined at state-level. These refer to the state number (absolute and relative position) within the current phone after forced alignment of the data.

A major concern with the standard set of linguistic features is its high dimensionality. There is an imbalance between the segmental and suprasegmental features and many components may not be useful for frame-level prediction. This is acceptable in tree-based acoustic modelling, as features are not included in the final model if they are not useful. In DNN-based acoustic modelling, the system is forced to account for all features in the input. Therefore, the question set was pruned and various features were discarded. Features at phrase and utterance levels were removed. Various features within the syllable and word level sets were ignored, such as forward or backward context and several positional features. This smaller pruned set of linguistic features is here termed the *hand-selected* feature set.

[1] `http://www.cstr.ed.ac.uk/projects/festival`

Binary representations of these question sets were used and all features were normalized to the range of [0.01, 0.99]. Table 1 summarizes the dimensionality at each linguistic level of each of the feature sets. Segmental features were kept constant for the standard and hand-selected sets. Thus, only suprasegmental features vary between the two sets.

3. Experiments

3.1. Database

These experiments were conducted on expressive audiobook data. It is desirable to use this type of data for these analyses as the narrator typically records entire chapters instead of isolated sentences. This ensures that higher-level prosodic variation is captured in the recorded speech, thus making it ideal for investigating effects of suprasegmental units within the utterance and discourse.

We use the audiobook *A Tramp Abroad*, which is freely available from *Librivox*[2]. The data has been preprocessed according to [23] and [24]. The hand-selected narrated speech described in [24] was used, thus excluding highly variable direct speech data. Training, development, and test sets of 4500, 300, and 100 utterances, respectively, were defined for the experiments described in this work. The data used for the listening test was randomly drawn from a held-out set.

3.2. Systems and hypotheses

Given three network architectures and two sets of linguistic features, six systems were trained. Two systems employed the basic feedforward deep neural network architecture (*feedfwd-**), two systems the cascaded deep neural network architecture (*cascaded-**), and two systems the a parallel network architecture (*parallel-**). Within each of these system pairs, we vary the input feature vector, either using the standard set (**-std*) or the hand-selected subset (**-hsel*). These systems were constructed to test the following hypotheses:

Addition of noisy suprasegmental features: Adding more (suprasegmental) features to a frame-level DNN will degrade the performance of the model. It is expected that the baseline system with the standard feature set will perform worse than the baseline system with the hand-selected features, as saturating a subsegmental model with noisy suprasegmental inputs is likely to be harmful.

Hierarchical systems: Hierarchical architectures will outperform non-hierarchical systems. Previous investigations have suggested that exploiting various linguistic levels tends to be beneficial for speech synthesis systems. We expect cascaded and parallel deep neural networks to outperform the basic feedforward network.

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

system	MCD	BAP	F0-RMSE	F0-CORR
feedfwd-std	4.68	2.22	28.23	.43
cascaded-std	4.60	2.19	27.43	.45
parallel-std	**4.59**	**2.17**	**26.97**	**.45**
feedfwd-hsel	4.61	2.20	27.66	.44
cascaded-hsel	**4.57**	2.19	27.48	.45
parallel-hsel	4.59	**2.17**	**27.16**	**.45**

Table 2: *Objective results for trained systems. MCD is mel cepstral distortion, BAP is band aperiodicity error, and F0-RMSE and F0-Corr are the root-mean-squared error and correlation between the predicted and original* f0 *signal on voiced frames.*

Parallel and cascaded DNNs: Parallel architectures will be preferred over cascaded architectures. Although using a different setup, previous work using these methodologies has found that parallel systems tend to outperform cascaded systems [14]. One of the disadvantages of processing suprasegmental information directly with a subsegmental network is that the system might learn to depend highly on segmental features and ignore long-term unit information. In a cascaded approach, even though segmental and suprasegmental feature sets are decoupled, a frame-level network still has to account for them. In a parallel architecture, this may not be the case, as the system processes the two feature sets separately and only concatenates them in the top hidden layer.

3.3. Objective results

Table 2 shows objective measures on the test set for all six systems. The first block in the table denotes the three networks operating with the full set of linguistic features. The second identifies those that use the hand-selected feature set. Observing only the baseline feedforward networks, we note a small improvement when moving to the hand-selected feature set, especially in terms of mel-cepstral distortion. All hierarchical systems outperform their respective baselines, although the impact appears to be less for the systems using hand-selected features.

The parallel architecture gives the best results. It is interesting to observe that we achieve performance that is comparable to that when using the hand-selected features. In terms of *f0* RMSE, the system with the full feature set gives the lowest error. This is reassuring, as we provide the syllable-level network with a larger number of input features. This suggests that hierarchical architectures are capable of leveraging high-dimensional representations of suprasegmental contexts. Such is not the case for frame-level networks. In the following section, we report a listening evaluation aimed at validating these observations.

3.4. Subjective results

To assess the naturalness of speech samples produced by the trained systems, we conducted a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test [25]. This methodology allows the simultaneous comparison of multiple samples. Each sentence to be tested is assigned a set of stimuli. In our case, a single set of stimuli includes 7 samples: one from each system described in section 3.2 plus a final sample of matching vocoded speech.[3] This final sample is termed the *reference*. Within each set, samples are unlabeled and, for each participant, the order of the samples is randomized. Participants are then asked to judge the set of parallel samples on a scale from 0 (completely unnatural) to 100 (completely natural) with respect to the reference sample. The reference sample itself is included in the unlabeled samples. This ensures that participants provide accurate judgements and fixes the high end of the scale.

A total of 20 native English listeners participated in the listening test. Each participant rated 20 sets of stimuli produced from sentences taken from a held-out set. Sentence order was randomized for each participant. This allowed us to gather a total of 400 parallel comparisons. All tests were conducted in sound-insulated booths and all listeners were remunerated for their time.

Figure 3 shows the distribution of the stimuli for each condition in terms of the absolute values given by the test participants. Figure 4 shows the distribution in terms of their rank order, as derived from the absolute values. In these figures, feedforward networks are abbreviated as *ffwd*, cascaded networks as *casc*, and parallel networks as *par*. As before *hsel* indicates the hand-selected feature set and *std* the standard feature set.

4. Discussion

To better understand the results, we conduct a two-tailed paired t-test on the absolute values given by the listeners. To account for multiple comparisons, we perform a Holm-Bonferroni correction on all results. All system pairs are significantly different at the level of $p<.05$, except (ffwd-hsel, casc-hsel), (casc-hsel, casc-std), and (par-hsel, par-std). Furthermore, we conducted a double-sided Wilcoxon signed-rank test on the rank order results with a Holm-Bonferroni correction. The same pattern was observed, with the addition of two system pairs not showing statistically significant differences: (ffwd-hsel, par-hsel) and (ffwd-hsel, par-std).

In terms our initial hypotheses, we observe that adding a larger number of suprasegmental features to a frame-level network significantly damages performance. This might be problematic when exploring a better understanding of longer context for prosody modeling. How-

[3] Speech samples can be found in: `http://homepages.inf.ed.ac.uk/s1250520/samples/ssw9.html`

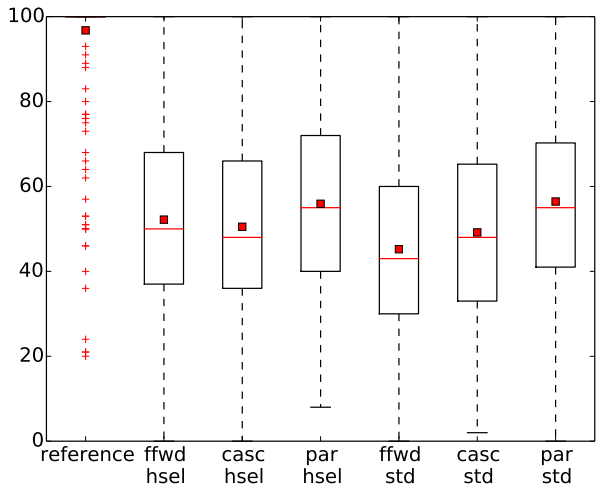

Figure 3: *Absolute results from the MUSHRA evaluation. Red horizontal line shows the median and the red square shows the mean.*

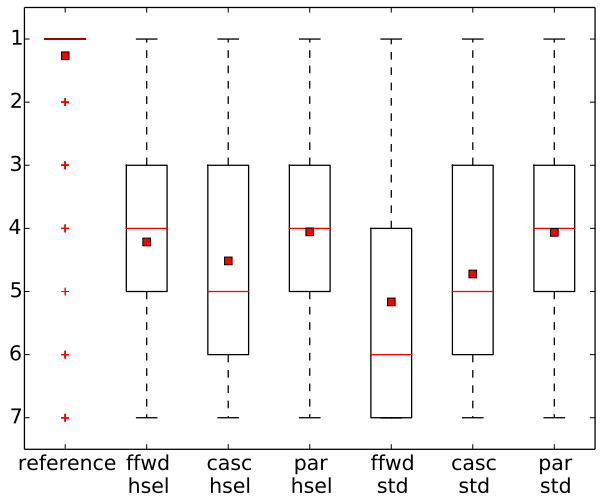

Figure 4: *Rank order results from the MUSHRA evaluation. Red horizontal line shows the median and the red square shows the mean.*

ever, we also observe that hierarchical systems are able to account for that difference if using high-dimensional noisy features. This difference does not occur for the hand-pruned feature set, where we failed to see significant improvements for the hierarchical systems in terms of rank order. This suggests that the hierarchical models may be operating as feature selectors for high-dimensional suprasegmental features. In our results, the hierarchical systems using the standard set are comparable to most systems using hand-selected features.

In terms of hierarchical strategies, we observe a preference for the parallel systems rather than the cascaded systems. This follows earlier conclusions, where this preference was also observed [14]. We could hypothesize that the frame-level part of the cascaded systems ends up depending too much on segmental features instead of balancing both sets. This is not the case for the parallel integration, as only one layer is used after concatenation. Further work could investigate this interpretation of the results by observing how the network weighs the various groups of features using techniques such as the ones described in [26].

As future work, the parallel neural network should be the focus of further research. It is unknown at this point whether decoupling the various linguistic levels could be useful. Similarly, it would be interesting to observe if these architectures have the capacity to leverage new features, such as text-derived word embeddings [21, 22, 27] or syllable bag-of-phones [27]. As suggested above, an attempt to visualize the impact of each linguistic level in the networks could be attempted [26]. Other lines of research could investigate how these hierarchical networks operate with recurrent systems, in a framework similar to that described in [15]. Finally, alternative acoustic features for the suprasegmental networks were not investigated. A strong possibility would be the use of selected components from wavelet-based decomposition of the *f0* signal [28, 29, 30].

5. Conclusion

Hierarchical systems structured as cascaded or parallel deep neural networks were investigated for decoupling segmental and suprasegmental features in statistical parametric speech synthesis. We have observed that, on expressive data, hierarchical systems are preferred over a standard feedforward network if using high-dimensional noisy features. This preference was not observed when using a hand-selected feature set. Hierarchical systems with a standard feature set are comparable to all systems using hand-selected features, which suggests they operates as a mostly as denoisers. We have also observed that parallel integration of segmental and suprasegmental features is preferred over cascaded integration. This preference was observed on both feature sets.

6. 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. We also thank Zhizheng Wu for contributions and insights into deep neural networks in this and earlier work.

7. 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] S. Nooteboom, "The prosody of speech: melody and rhythm," *The handbook of phonetic sciences*, no. 5, pp. 640–673, 1997.

[6] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," 2013.

[7] 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.

[8] 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.

[9] A. Stan and M. Giurgiu, "A superpositional model applied to f0 parameterization using dct for text-to-speech synthesis," in *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on*. IEEE, 2011, pp. 1–6.

[10] J. Latorre and M. Akamine, "Multilevel parametric-base f0 model for speech synthesis." in *INTERSPEECH*, 2008, pp. 2274–2277.

[11] N. Obin, A. Lacheret, X. Rodet *et al.*, "Stylization and trajectory modelling of short and long term speech prosody variations," in *Interspeech*, 2011.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," 2015.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] I. Recommendation, "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[26] K. C. Sim, "On constructing and analysing an interpretable brain model for the dnn based on hidden activity patterns," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 22–29.

[27] M. S. Ribeiro, O. Watts, and J. Yamagishi, "Syllable-level representations of suprasegmental features for dnn-based text-to-speech synthesis (under review)," in *submitted to Interspeech*, 2016.

[28] A. S. Suni, D. Aalto, T. Raitio, P. Alku, M. Vainio *et al.*, "Wavelets for intonation modeling in hmm speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[29] 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.

[30] M. S. Ribeiro, J. Yamagishi, and R. A. J. Clark, "A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis," in *Proc. Interspeech*, Dresden, Germany, September 2015.