Phone-Level Embeddings for Unit Selection Speech Synthesis

Antoine Perquin, Gwénolé Lecorvé, Damien Lolive, Laurent Amsaleg

Phone-Level Embeddings for Unit Selection Speech Synthesis

Antoine Perquin[1], Gwénolé Lecorvé[1], Damien Lolive[1], and Laurent Amsaleg[2]

[1] Univ Rennes, CNRS, IRISA
[2] Univ Rennes, CNRS, INRIA, IRISA
{antoine.perquin, gwenole.lecorve, damien.lolive, laurent.amsaleg}@irisa.fr

Abstract. Deep neural networks have become the state of the art in speech synthesis. They have been used to directly predict signal parameters or provide unsupervised speech segment descriptions through embeddings. In this paper, we present four models with two of them enabling us to extract phone-level embeddings for unit selection speech synthesis. Three of the models rely on a feed-forward DNN, the last one on an LSTM. The resulting embeddings enable replacing usual expert-based target costs by an euclidean distance in the embedding space. This work is conducted on a French corpus of an 11 hours audiobook. Perceptual tests show the produced speech is preferred over a unit selection method where the target cost is defined by an expert. They also show that the embeddings are general enough to be used for different speech styles without quality loss. Furthermore, objective measures and a perceptual test on statistical parametric speech synthesis show that our models perform comparably to state-of-the-art models for parametric signal generation, in spite of necessary simplifications, namely late time integration and information compression.

1 Introduction

Unit selection speech synthesis concatenates pre-existing segments of recorded speech, producing high-quality, natural sounding, oral renderings of sentences [2]. This process optimises a *target cost* function selecting the units that best match the linguistic descriptions of the phonemes to synthesize. Quality results from the involvement of linguistic experts who carefully design that function. These methods, however, suffer from concatenation errors and cannot generalize outside the pre-recorded units. Furthermore, they necessitate extremely costly human expertise, which might not exist when targeting or adapting to other domains or languages.

Instead, this paper proposes an embedding-based method allowing to rely on euclidean distances between units when optimizing their selection. Not only is this cheap compared to human expertise, but embeddings also facilitate domain adaptation. This paper presents four models, with two of them resulting in phone-level embeddings for unit selection. Three of them rely on a feed-forward Deep

Neural Network (DNN) whereas the last one uses a Long Short Term Memory layer (LSTM). Experiments using a large French speech corpus show that the use of the embeddings outperforms expert-based unit selection.

A few remarks are in order. One, temporal information can be integrated in different ways. We compare its early or later integration in DNNs, and study the use of temporal dependencies brought by an LSTM layer. Two, in an attempt to understand if mitigating the effects of the curse of dimensionality on the embeddings is beneficial, three models use varying layers sizes at their core. Three, as far as we know, no objective metrics for assessing the quality of such embeddings exist. As a proxy, we evaluate their ability to predict appropriate acoustic features. Four, in contrast to state of the art contributions that work for English, we work with French. This is the opportunity to assess how the embedding-based approach performs on French.

The remainder of this paper is as follows. Section 2 presents related work. Section 3 discusses the handling of time and dimensionality in the embeddings. Finally, Section 4 compares our approach to an expert unit selection system.

2 Related Work

The Statistical Parametric Speech Synthesis (SPSS) approach uses vocoders to synthesize speech from acoustic parameters predicted by a so-called *acoustic model* based on linguistic features [1]. Recent models are now based on DNNs [13]. Beside pros and cons of this approach (good flexibility but low quality), the resulting prediction models are interestingly independent of any linguistic expertise.

This approach gave rise to different extensions. On the one hand, the Wavenet model proposes to integrate vocoding in the prediction model, i.e., the raw waveforms are directly predicted [6]. End-to-end approaches such as [9] go further by predicting these waveforms directly from text, that is linguistic features are removed.

On the other hand, prediction models have also been used to transform linguistic features into other representations, based on which a new target cost for unit selection can be defined. In turn, the new representation space can be handled more easily than the symbolic space of the linguistic features. In [12], linguistic features are converted into acoustic ones based on hidden Markov models. The Kullback-Leibler divergence is then used as a target cost. Alternatively, [4, 8] have proposed intermediate representations extracted from hidden layers of DNNs, also called *embeddings*.

In [4], each phone is divided into 4 sections for which the mean and variance of the embedding of each frame is computed. The target cost function is the sum of the Kullback-Leibler divergence on the mean and variance of the 4 sections. This showed that there is no loss in the quality of speech when using embeddings instead of acoustic features. The subdivisions, however, still involve some human expertise.

In [8], a phone level embedding is obtained with a multi-modal model composed of an acoustic encoder and a linguistic encoder. The target cost is the euclidean

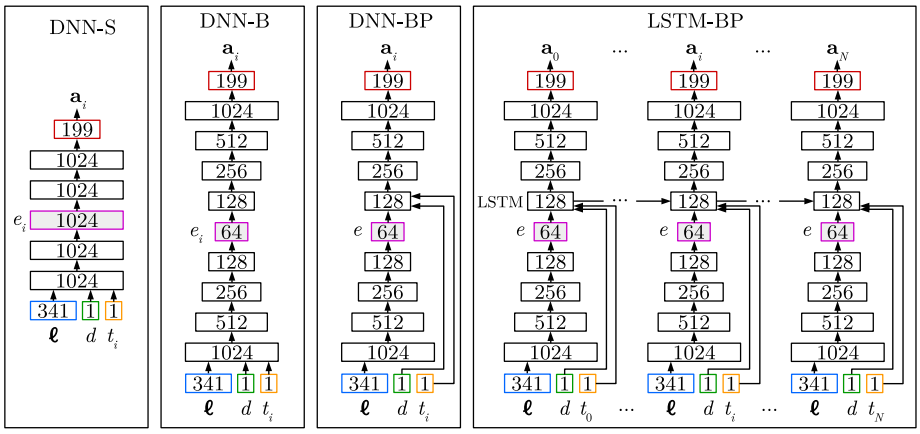

Fig. 1. Architecture of the 4 models. Layer sizes are reported. $\boldsymbol{\ell}$ denotes the vector of linguistic features, d the phone duration, t_i the time position in the phone for the i-th frame, $\mathbf{a}_i$ the corresponding acoustic features, and N is the total number of frames in the phone.

distance in the embedding space. This work is very similar to ours, and was actually conducted at the same time as [7]. Our work focuses on the effect of various time integration and information compression schemes, rather than those of multi-modality.

3 Handling of time and dimensionality

A key assumption in our work is that the quality of an embedding is correlated to the quality of the model from which it is derived, in our case acoustic models. In this section, different variants of acoustic models are studied, among which those proposed for unit selection. These models are objectively and perceptually evaluated to assess the behaviour of embeddings with respect to information compression and different strategies for time integration. This section first presents the models under study before describing the experimental dataset, then the objective and perceptual tests.

3.1 Models

Four models are studied, see Figure 1. The first model is a standard acoustic model (DNN-S), a simple feed-forward DNN, comparable to the one proposed in [10]. For a given phoneme, the model predicts acoustic features $\mathbf{a}_i$ of the i-th frame based on the linguistic feature vector $\boldsymbol{\ell}$. Those linguistic features provide information about the phoneme, e.g., its identity, the one of its close neighbours, its position in the syllable/word/utterance it belongs to, etc. The timing information is encoded as two numerical features: the phone duration d in

seconds and the relative position t_i of the frame i inside the phone. This timing information is useful to take into account the dynamics of acoustic features when realising a phone. For the frame i, the output of the middle layer can be seen as the frame embedding e_i.

The second model, DNN-B, has a similar architecture as DNN-S but introduces a bottleneck layer. This bottleneck is obtained by gradually decreasing then increasing the size of the hidden layers. e_i is then a compression of the linguistic features $\boldsymbol{\ell}$. As a side effect, compressing the embedding space avoids the curse of dimensionality and allows for tractable similarity measures.

The limitation of both models is that the resulting embeddings would only represent the frame currently predicted, whereas phone-level embeddings are needed for unit selection. To solve this problem, we propose to postpone making use of the timing information (d and t_i) until after the embedding layer, as in [8]. Applying this principle on DNN-B gives the model DNN-BP, P for Phone-level. For a given linguistic feature vector $\boldsymbol{\ell}$, we obtain a phone-level embedding e.

Finally, to attempt to model the timing dependency across frames, we propose to replace the layer after the embedding layer with an LSTM in DNN-BP to obtain the model LSTM-BP. While the previous models are trained on independent frames, LSTM-BP is trained on the full frame sequence of the considered phone. This method decreases the shuffling possibilities over the training set but could lead to better predictions, and thus better embeddings.

The implementation details of the different models are as follows:

- DNN-S: 5 hidden layers of size 1024. The total number of parameters is 4,75 millions.
- DNN-B: The bottleneck scheme is symmetrically designed: 9 hidden layers of size 1024, 512, 256, 128, 64, 128, 256, 512, 1024. The total number of parameters is 1.95 millions.
- DNN-BP: Same as DNN-B, except timing is postponed. The total number of parameters is 1.95 millions.
- LSTM-BP: Same as DNN-BP, except the second 128 dimensional layer is replaced with an LSTM layer of size 128 too. The total number of parameters is 2.04 millions.

All hidden layers rely on the hyperbolic tangent (*tanh*) activation function. For all models, the output layer is with a linear activation.

3.2 Dataset and experimental setup

Our models were trained on a corpus corresponding to the reading of a French audio-book by a professional French speaker resulting in approximately 11 hours of speech for a total of 3300 utterances (approximately 390 000 phonemes). Speech is expressive (narration, acted dialogues), and sentences are complex (long sentences, formal register) due to the style of the audiobook's author (Marcel Proust). 105 utterances were held out for the listening test, while the rest was shuffled at the frame level (or phone level for model LSTM-BP) and then split

into three sets : 90% for the training set and 5% each for the validation set and test set.

About 110 linguistic features are considered for each phone. Categorical attributes represent information about quinphones, syllables, articulatory features, and part of speech for the current, previous and following words. They are encoded in one-hot. 34 other features are numerical, such as the position of the phone inside the word or the utterance. After encoding, the overall linguistic vector is of size 341. The linguistic features and the timing information were normalised to the range [0.01, 0.99]. Each linguistic feature was manually extracted, without automatic annotation.

The acoustic features, extracted using the WORLD vocoder [5], consist of a 60 dimension Mel-Generalized Cepstral coefficients (MGC) vector, a 5 dimension band-aperiodicity (BAP) vector and the fundamental frequency F_0. Those features were extracted every 5 ms. The F_0 coefficient was linearly interpolated on unvoiced parts, a boolean attribute keeps track whether the frame was voiced or not and the logarithm was applied to F_0. Finally, the deltas and delta-deltas were computed for MGC, BAP and F_0. In total, the acoustic vector is of size 199. The acoustic features were centered and normalized to unit variance.

The implementation was done using Keras with TensorFlow. Training was done on a GTX 1080 Ti, over 100 epochs using RMSPROP with the mean square error as a loss function. The model weights with the best performance on the validation set were saved. Those models were trained using the true duration values.

3.3 Objective Evaluation

The 4 models were evaluated in an acoustic modelling perspective. Table 1 reports the quality of the predicted acoustic features according to the following measures:

- MCD : Mel-Cepstral Distortion on MGC coefficients.
- BAP : a distortion measure on BAPs.
- V/UV : Voiced/unvoiced error rate.
- RMSE(F0) : Root mean squared error on F_0.

Those measures are computed between the acoustic features predicted by the DNNs and the reference acoustic features. The results from a state of the art acoustic model are also directly reported from [11]. They were not trained on the same data, nor even the same language. They are presented for the sake of a sanity check.

First, our measures are higher than those from [11], especially regarding F_0 and voicing error. We believe the high error on those two aspects is due to the high expressiveness of our speech corpus. Taking this into account, these results can be considered as acceptable. Second, by comparing model DNN-B and DNN-BP, we can see that the displacement of the timing information raises the error for all measures by only a small margin. Then, while one would expect DNN-S to lead to the best results, it appears that its performance is a bit worse

Table 1. Objective evaluation of the predictions of our models.

	MCD (dB)	BAP (dB)	V/UV (%)	RMSE(F_0) (Hz)
DNN-S	5.22	0.48	17.2	18.3
DNN-B	5.06	0.35	12.6	17.9
DNN-BP	5.09	0.36	13.7	18.2
LSTM-BP	5.80	0.49	19.7	19.5
DNN (reported from [11])	4.54	0.36	11.38	9.57

than for DNN-B and DNN-BP. Possible explanations are the larger number of parameters for model DNN-S (more difficult to reach the global optimum for weights) or the larger number of layers in the DNN-B and DNN-BP. At least, the results show that the bottleneck does not hurt. On the contrary, LSTM-BP leads to the worst results. This is surprising since adding extra information about the previous frames should not degrade the prediction. In our opinion, these results mainly come from the fact that training data are organised as sequences of frames, thus reducing the diversity of observations during training, whereas the other models are trained on shuffled frames. Despite those bad results, we chose to keep LSTM-BP to evaluate the effect of an acoustic model quality on its embeddings and on the resulting synthesized speech.

3.4 Perceptual Evaluation

The previous objective measures are not perfect estimators of the overall speech quality. For example, a significant improvement in MCD does not necessarily translate to a perceptual improvement of the synthesized speech. Thus we want to observe the perceptual impact of embeddings on SPSS.

For each model, synthetic speech has been generated based on the predicted acoustic features using an SPSS approach. In practice, the WORLD vocoder is used. Long utterances were split into breath groups of 4-5 seconds. The timing information (d and t_i) was derived from a duration model (DNN with 6 hidden layers and *tanh* activation) predicting the number of frames N of a phone based on its linguistic feature vector $\boldsymbol{\ell}$. The mean absolute error of this model is 3.7 frames. d is derived by multiplying N by the sampling rate, t_i is simply $\frac{i}{N}$. A listening test was conducted with 21 French native speakers. They were asked to rate between 0 and 10 the overall quality of speech utterances. Synthetic speech coming from the 4 models was presented along with natural speech (NAT) and utterances vocoded based on the reference acoustic features (VOC). Each listener was given 10 in-domain utterances and 10 out-of-domain utterances. For in-domain evaluation, we used the 105 sentences left-out from the dataset, while for out of domain evaluation we used 100 phonetically balanced French sentences. Thus, every utterance was rated by at least two different listeners.

In accordance with the objective results, the listening test shows that our models do not perform well in SPSS mode, as can be seen on Figure 2. While our

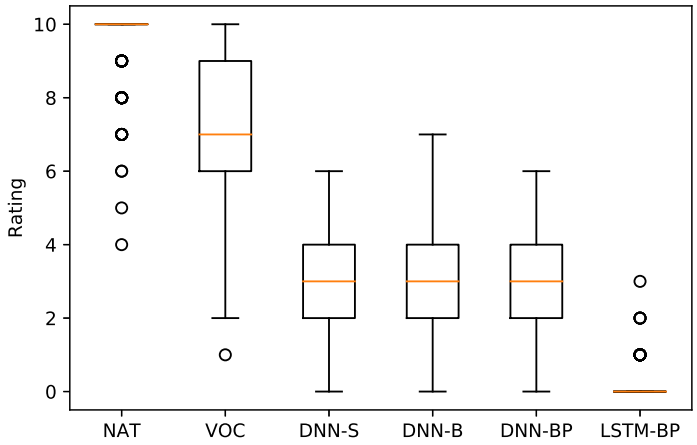

Fig. 2. Results of the listening test for the models in statistic parametric mode for in-domain utterances.

listeners reacted well to perfect SPSS (analysis-synthesis), giving a 7.2 mean score to the system VOC, they gave a mean score of around 3 to our first 3 models, and agreed that system LSTM-BP's productions were incomprehensible with a mean score of almost 0. This is not surprising since this system had a higher MCD than other systems, which is measured on a log-scale. There is no real statistically significant (p-value greater than 0.05) difference between systems DNN-S, DNN-B and DNN-BP which proves that the displacement of the timing information does not cause any perceptual loss in quality. Surprisingly, the natural speech received a couple of marks below 7. The corresponding utterances were perceived as having unnatural prosody because they were cut to be shortened.

4 Comparison with Expert Unit Selection

In this section, the proposed unit selection method is compared to a system where the target cost is defined based on expert knowledge.

4.1 Unit Selection Engine

Once phone-level embeddings have been extracted, we use them to guide the unit selection process. Before synthesis, for each phone in the database, the corresponding embedding is computed and stored. At synthesis time, a pre-selection reduces the set of candidate units to those corresponding to the same phoneme as the target, in order to reduce the computation time. The number of candidate units is reduced further by searching the 25 nearest neighbours of the

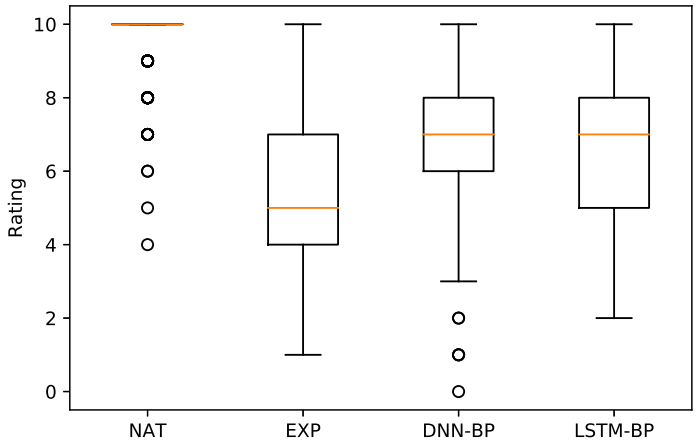

Fig. 3. Results of the listening test for the models working in unit selection mode for in-domain utterances.

target phone in the embedding space among the pre-selected units. Finally, the lattice of these nearest neighbours for each phone is decoded to find the sequence of units minimizing the sum of the target and concatenation cost.

The target cost is defined as the euclidean distance in the embedding space between the candidate and target phone. For the expert system, the target cost was originally defined as a weighted sum of linguistic features and has since been improved over the years. The join cost for all systems is the same as in [3], defined as a sum of euclidean distances on acoustic features between following candidate units.

4.2 Perceptive Evaluation

Since there is no proposed measure in the literature to evaluate the quality of an embedding in relation to speech synthesis, we directly address the subjective evaluation of the embeddings with listening tests. The results for in-domain and out-of-domain utterances can be found on Figure 3 and Figure 4 respectively.

For in-domain utterances, the 3 proposed models were all awarded really high grades (multiple times rated with a 10/10 score for each system) but also really low notes (up to 0 for system DNN-BP). However, on average, the expert system received a mean score of 5.4, system DNN-BP received 6.8 and system LSTM-BP received 6.6. While we cannot statistically distinguish the two systems with embeddings (p-value greater than 0.05), both are a statistically significant improvement over the expert system. Interestingly, even if the system LSTM-BP performed really poorly for SPSS, it receives good results for the unit selection paradigm.

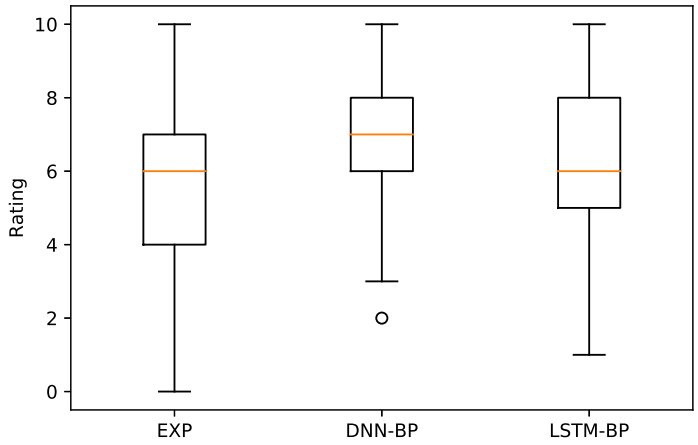

Fig. 4. Results of the listening test for the models working in unit selection mode for out-of domain utterances.

For out-of-context utterances, natural speech from the same voice as the one used in the database was unavailable. The expert system received a mean score of 5.9, system DNN-BP received 7.0 and system LSTM-BP received 6.4. This time, the difference between the two systems using embeddings is significant. However, they both perform without significant loss over in-domain synthesis. Surprisingly, the expert system appears to have been better rated on out-of-domain utterances. A probable explanation is the absence of natural speech as a higher baseline during the MUSHRA test. Still, the embedding systems remain significantly better than the expert one proving that the embeddings are general enough to be used in other domains.

5 Conclusion

In this paper, we proposed two models to extract phone-level embeddings in the context of DNN-driven unit selection. The models were compared to other DNNs on the task of acoustic modeling. The experiments highlighted that late integration of time and information compression (bottleneck) do not impact the quality of feature prediction, even though the use of LSTM did not seem conclusive yet. Then, experiments on unit selection showed that quality of synthesized speech based on embeddings resulting from our models is perceptually preferred over an approach with an expertly defined target cost. They also demonstrated that the proposed embeddings are general enough to be used in multiple domains.

Besides these results, our study also highlights the fact that the link between the acoustic feature prediction and the quality of unit embeddings is not clear.

For instance, the LSTM-BP model led to good unit selection results whereas it was very bad for SPSS. It would thus be interesting to further study how embeddings could be evaluated with objective measures, e.g. by analyzing the topological properties of the embedding spaces. Then, extensions of the proposed models should be tested to produce better embeddings, and to better understand dependencies across different types of information. Multi-modality, as in [8], is a first direction. The integration of duration into the embeddings is another. Finally, our objective in the long term is to deal with large heterogeneous speech databases (different speakers, different languages, etc.). Apart from the previously raised questions, this would require to study the compliance of the embeddings with fast database searches, especially in regard to approximate nearest neighbours techniques.

Acknowledgments

This study has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015.

References

1. Black, A.W., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 4, pp. 1229–1232 (2007)
2. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 1, pp. 373–376 (1996)
3. Lolive, D., Alain, P., Barbot, N., Chevelu, J., Lecorvé, G., Simon, C., Tahon, M.: The irisa text-to-speech system for the Blizzard challenge 2017. In: Proceedings of the Blizzard Challenge Workshop (2017)
4. Merritt, T., Clark, R.A., Wu, Z., Yamagishi, J., King, S.: Deep neural network-guided unit selection synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5145–5149 (2016)
5. Morise, M., Yokomori, F., Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems 99(7), 1877–1884 (2016)
6. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. In: Proceedings of the ISCA Speech Synthesis Workshop (SSW). pp. 125–125 (2016)
7. Perquin, A.: Big deep voice: indexation de données massives de parole grâce à des réseaux de neurones profonds. Master's thesis, University of Rennes 1 (2017)
8. Wan, V., Agiomyrgiannakis, Y., Silen, H., Vit, J.: Googles next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). pp. 1143–1147 (2017)

9. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al.: Tacotron: Towards end-to-end speech synthesis. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). pp. 4006–4010 (2017)
10. Wu, Z., King, S.: Improving trajectory modelling for dnn-based speech synthesis by using stacked bottleneck features and minimum generation error training. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24(7), 1255–1265 (2016)
11. Wu, Z., Watts, O., King, S.: Merlin: An open source neural network speech synthesis system. In: Proceedings of the ISCA Speech Synthesis Workshop (SSW). pp. 218–223 (2016)
12. Yan, Z.J., Qian, Y., Soong, F.K.: Rich-context unit selection (RUS) approach to high quality TTS. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). pp. 4798–4801 (2010)
13. Ze, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7962–7966 (2013)