# DEEP NEURAL NETWORK-GUIDED UNIT SELECTION SYNTHESIS

*Thomas Merritt*[1], *Robert A.J. Clark*[1*], *Zhizheng Wu*[1], *Junichi Yamagishi*[1,2], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]National Institute of Informatics, Japan

t.merritt@ed.ac.uk

## ABSTRACT

Vocoding of speech is a standard part of statistical parametric speech synthesis systems. It imposes an upper bound of the naturalness that can possibly be achieved. Hybrid systems using parametric models to guide the selection of natural speech units can combine the benefits of robust statistical models with the high level of naturalness of waveform concatenation. Existing hybrid systems use Hidden Markov Models (HMMs) as the statistical model. This paper demonstrates that the superiority of Deep Neural Network (DNN) acoustic models over HMMs in conventional statistical parametric speech synthesis also carries over to hybrid synthesis. We compare various DNN and HMM hybrid configurations, guiding the selection of waveform units in either the vocoder parameter domain, or in the domain of embeddings (bottleneck features).

***Index Terms***— speech synthesis, hybrid synthesis, deep neural networks, embedding, unit selection

## 1. INTRODUCTION

Statistical parametric speech synthesis (SPSS) systems were originally proposed in order to offer more flexibility (e.g., adaptability to target speech) than is possible with unit selection synthesis. However during the process of extracting and modelling speech parameters, followed by resynthesis, the naturalness of the speech is substantially reduced. As a consequence, these systems are consistently rated as less natural than unit selection, as we see in the results of many Blizzard Challenges [1, 2, 3, 4].

During previous investigations, some of the hypothesised explanations for the reduced quality of SPSS systems have been formally tested [5, 6, 7, 8]. The finding that across-linguistic-context averaging was harmful motivated us to build an HMM system which performed no such averaging [9]. In [6, 7, 8], we also identified the parametrisation step (i.e., vocoding) as introducing large degradations in quality, even before any modelling has taken place.

In order to increase the quality of speech above the ceiling imposed by vocoding, we conducted the current investigation. Our starting point is a prototypical unit selection system (Festival's Multisyn engine), in which very little processing of the speech waveforms is performed. We present an investigation into the effectiveness of hybrid synthesis (unit selection guided by SPSS) within the Multisyn framework.

## 2. PRIOR WORK

### 2.1. Unit selection

Unit selection synthesis is usually described as an optimisation problem: to find the sequence of units (diphones, in Multisyn) that minimises the sum of target costs and join costs [10], which involves trading off how well a candidate unit meets a required specification against how well it concatenates with neighbouring units. By defining the join cost to be zero for units that are contiguous in the database, unit selection effectively uses relatively large units of variable size.

Standard unit selection systems typically use mismatches between the linguistic specifications of the target and candidate units to compute a target cost. Distances between acoustic features are used to compute join cost [11, 12, 13].

Whilst speech within contiguous regions found in the database is effectively 'perfectly natural', unit selection speech generally suffers from concatenation artefacts. A variety of hybrid synthesis systems have been proposed to solve this problem by employing statistical models to predict the acoustic properties of speech, and then selecting units from the database that match [14, 15, 16].

### 2.2. Hybrid synthesis

Hybrid synthesis systems thus use statistical models (usually by generating speech parameter trajectories) as the basis of the target cost function [15, 16, 14]. An extension of this approach is 'multiform' synthesis in which some types of units are generated via vocoding, whilst others are retrieved from the speech database [17, 18, 19, 20, 21] although this is outside of the scope of our current investigation. Hybrid systems have performed very well in Blizzard Challenges [1, 2, 3, 4].

HMMs are the preferred statistical models in hybrid systems' target cost function, despite recent but compelling evidence that DNNs are superior to the regression tree employed

*Now at Google

in HMM systems [22, 23, 24]. One exception is the investigation in [21], where a bidirectional recurrent neural network (RNN) provides a prosodic target. However the authors in [21] did not use this system to synthesise exclusively using concatenation and instead used it in a multiform setup combined with modelled prosody. Previous investigations were made into the use of RNNs for extracting prosodic information from Mandarin which is then fed into a second RNN which outputs prosodic information [25]. The details of how this is used to inform selection of units is not made clear and was not formally tested.

In [9], context embeddings (which can also be called 'bottleneck features' when they are derived using a hidden layer of a feed-forward neural network [22]) were used to select rich-context HMM models (models which are trained only on samples where the linguistic contexts exactly match) in order to select better models for synthesis in the inevitable event that contexts seen at synthesis-time were not observed in the training data [26]. This outperformed conventional HMM synthesis, which provides a motivation to use these context embeddings to select units in a hybrid system. We use distance in embedding space (or distance in speech parameter space in some cases) to measure the mismatch between the target and a candidate unit.

## 3. MULTISYN

Multisyn is a general purpose unit selection framework enabling simple implementation of unit selection synthesis within the Festival toolkit [27, 28, 12]. Festival's Multisyn is used as one of the baselines for the Blizzard Challenge, and forms the basis of our hybrid unit selection systems. The unit size used in all systems reported here is the diphone. Although gains have been demonstrated using other sized units [14], this is outside the scope of the current investigation.

The Multisyn target cost function is a simple weighted sum of mismatches in selected linguistic features. The default weights were left unchanged for the baseline system used in this investigation, but the relative weight of the target cost compared to the join cost was manually tuned, for consistency with the hybrid systems to which it was compared.

The join cost for Multisyn is a sum of distances between 12 MFCCs, f0 and energy from the frames either side of the join [27, 28]. This default join cost was used in all systems.

Before performing the search to minimise the number of join and target costs required to be computed, it is necessary to pre-select a shortlist of candidates for each target position. The default pre-selection method in Multisyn returns candidates with matching diphone identity. In the event that this list is empty, a back-off scheme is invoked which uses manually-written phone substitution rules. Again, this default scheme was left in place, although it may be possible in future to use distance in embedding or speech parameter spaces in the backoff procedure.

## 4. PROPOSED HYBRID TARGET COST

The context embeddings derived from a neural network, or alternatively the actual speech parameters predicted at the output of the network, can be thought of as a non-linear projection of the input linguistics features. The projection is learned in a supervised manner, according to whatever optimisation criterion is used to train the network. It is this supervision from acoustic information that makes these DNN-derived features more powerful than the purely linguistic feature-based function used as standard in Multisyn. For example, linguistic features that are not predictive of acoustic properties will be discarded.

The motivation for using a DNN – that, crucially, has been trained to perform parametric speech synthesis – to provide the embeddings (rather than some other method), comes from the universally positive reports of DNN synthesis in recent literature.

Multisyn operates on diphone units, but the synthesis DNN we used operates on phone units. To map between these, we divided each phone into 4 sections. The features being used for the target cost (either context embeddings from a DNN [22], or output speech parameters from the neural network or an HMM) are gathered together across all frames within each of these 4 regions, from which we compute the mean and variance per section. The variance is floored at 1% of the global variance per feature (the floor value was chosen via informal listening). This is done in the same way for both candidate and target.

The Kullback Leibler divergence (KLD) [29] is computed for each of the 4 sub-phone regions individually. The use of KLD in embedding space follows on from our previous work on 'rich-context' modelling [9].

The KLD between distribution $f$ of the features computed for the frames corresponding to a given section in the test sentence, and distribution $g$, is:

$$D_{KL}(f||g) = \frac{1}{2}[log\frac{\mid \Sigma_g \mid}{\mid \Sigma_f \mid} + Tr[\Sigma_g^{-1}\Sigma_f] - d$$

$$+(\mu_f - \mu_g)^T\Sigma_g^{-1}(\mu_f - \mu_g)], \quad (1)$$

where $\mu$ and $\Sigma$ are mean and covariance and $d$ is the dimensionality of the feature vector. The KLD for each of the 4 sections comprising a diphone is summed together to give the final divergence score. The average of $D_{KL}(f||g)$ and $D_{KL}(g||f)$ was used in order to make the measure symmetrical.

The SPSS-derived target cost function used in this investigation is designed to be independent of phoneme duration, relying on the target cost to select candidates with suitable durations. However, work on explicit control of duration may be fruitful in the future.

**Table 1**. *Conditions included in listening test*

| ID | Description |
|---|---|
| N | Natural speech |
| M | Multisyn |
| LE | Multisyn with target cost derived from context embedding from 2nd layer of 6 layer DNN (as in [9]) |
| HE | Multisyn with target cost derived from context embedding from 5th layer of 6 layer DNN |
| NP | Multisyn with target cost derived from output from Stacked bottleneck DNN system [22] |
| HP | Multisyn with target cost derived from output from HTS demo with GV [30] |

## 5. EXPERIMENTS

### 5.1. Implementation

The systems shown in Table 1 were constructed in order to test the effectiveness of speech parameter trajectories (from HMMs or DNNs) and context embedding trajectories (from DNNs) for computing the target cost. As previously stated, the only component that differs between systems is the target cost, and the relative weight between target and join costs (tuned by informal listening).

Systems $LE$ and $HE$ use 32-dimensional context embedding features generated by a DNN similar to that described in [22, 31]. These come from the 2nd (lower layer of the DNN, closer to the linguistic input) or 5th (higher layer of the DNN, closer to the speech parameters output) layer of a 6 layer feed forward DNN, respectively. These systems are successors to the rich context system described in [9] but instead of generating the speech using a vocoder, they perform unit selection and concatenation.

System $NP$ uses the speech parameters output from the final layer of the stacked bottleneck DNN system presented in [22, 31]. The speech parameters form an 86-dimensional vector (60th order mel-generalised cepstrum, 25 band-aperiodicities, f0).

System $HP$ was included to represent a conventional HMM-guided hybrid system. This system uses the parameters generated by HMMs trained using the HTS demo recipe [30], including GV, to compute the target cost in much the same way as system $NP$ makes use of the generated speech parameters from the stacked bottleneck DNN system. The comparison between these systems $HP$ and $NP$ will tell us whether the gains offered by DNNs in SPSS carry over to the hybrid scenario.

Informal listening to the speech generated via vocoding from the speech parameters of systems $NP$ and $HP$ was conducted, in order to confirm that these systems generate speech of the quality expected. This vocoder-generated speech was not evaluated formally in the listening test reported below. The relative target cost vs join cost weight for all systems ($M$, $LE$, $HE$, $NP$ and $HP$) was tuned by informal listening using a few listeners.

2400 sentences from a male speaker of British English [32] were used as the training set for the HMMs and DNNs, and as the unit database in all systems. The text of 20 unseen Herald newspaper news sentences were used as a development set for tuning the target cost weight of each system. An additional 90 unseen Herald news sentences were then used for the listening test. For DNN synthesis (required to produce the embeddings or speech parameters of systems $LE$, $HE$ and $NP$), durations predicted by the HMM system were used; note that the durations of the final hybrid synthetic speech are determined by the unmodified natural durations of the candidate diphone units selected from the database. Before conducting the listening test, all utterances were volume normalised according to [33].

### 5.2. Experimental setup

The listening test followed the MUSHRA paradigm [34], comprising the systems shown in Table 1, with the same set up as in [7]. In a MUSHRA test, versions of single sentence generated under all conditions are presented side-by-side to the listener, allowing direct comparisons in naturalness to be made. The listener is required to rate the systems between 0 (completely unnatural) and 100 (completely natural). This paradigm was originally designed to evaluate audio codecs and we find that it is effective at prising apart relative differences between multiple systems because listeners have knowledge of the full range of those systems before making they judgements. System $N$ acts as a hidden (i.e., not labelled in the test) upper anchor. Listeners are instructed that there is one system that must be given a rating of 100. MUSHRA usually also includes a lower hidden anchor; it is unclear what should be used for this in the case of synthetic speech, so no lower anchor was included in this test (as was also the case in [7] and [9]).

This test was conducted with 30 listeners, with each listener rating 30 screens. Each screen presented 6 stimuli at once: a single sentence under all 6 conditions. The 30 listeners were split into 3 groups of 10 listeners, and each was presented with a disjoint set of 30 sentences; thus 90 different sentences were used. The stimuli played to listeners along with listener responses can be found at [35].

## 6. RESULTS

Figures 1 and 2 show the listeners responses from the MUSHRA test, in terms of the absolute values of their scores, and in terms of the rank order of systems derived from these scores, respectively. The dashed green lines added to the box plots show mean values. All tests for significant differences used Holm-Bonferroni correction due to the large number of condition pairs to compare.

All conditions are significantly different from each other in terms of absolute value, except between: $M$ and $HP$, $LE$ and $HE$, $LE$ and $NP$, $HE$ and $NP$. Significant differences are in agreement using a t-test and Wilcoxon signed-rank test at a p value of 0.01.

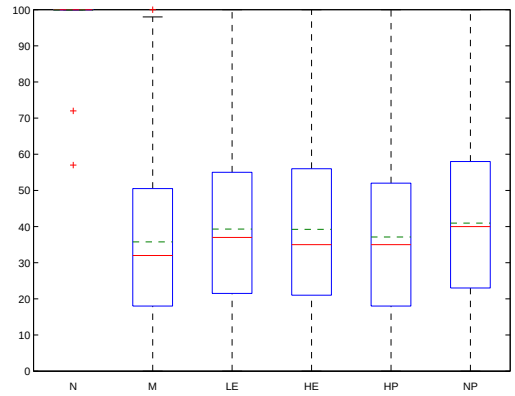


**Fig. 1**. *Boxplot of absolute values from MUSHRA test*

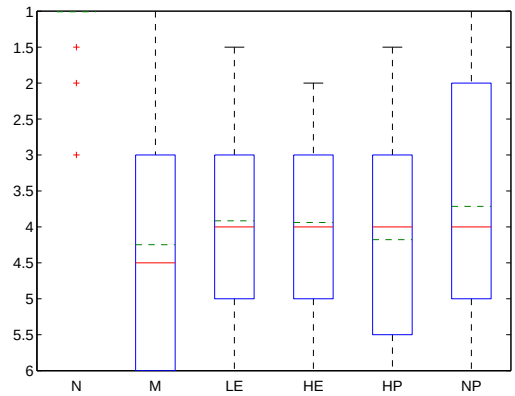


**Fig. 2**. *Boxplot of the rank order from MUSHRA test*

All conditions are significantly different from each other in terms of rank order, except between; $M$ and $HP$, $LE$ and $HE$. These significant differences are in agreement using a Mann-Whitney U test and a Wilcoxon signed-rank test at a p value of 0.01. There is a disagreement in statistical significance between conditions $LE$ and $NP$ with the Mann-Whitney U test finding this difference in ranking to be statistically significant whereas the Wilcoxon signed-rank test does not.

### 6.1. Comparison to baseline system $M$

We can see that the 'trajectory tiling' approach to unit selection described in [14], and implemented in our systems $LE, HE, HP, NP$ is generally effective, with all systems performing at least as well as the baseline, and significantly better in all cases where a DNN was used as the parametric model. We were not able to obtain significant improvements over baseline with HMM-generated speech parameter trajectories (system $HP$).

### 6.2. DNNs vs HMMs

The use of deep neural networks in systems $LE$, $HE$ and $NP$ provides significant improvements over both the baseline ($M$) and the HMM-driven hybrid system ($HP$). This demonstrates that the gains found in SPSS systems when moving from HMMs + regression trees to DNNs transfers over to the hybrid unit selections paradigm.

## 7. CONCLUSIONS & FUTURE WORK

We have proposed to use deep neural networks to guide unit selection systems, and have presented an experimental comparison of several different configurations of hybrid unit selection, all implemented within Festival's Multisyn framework. We found that the use of a DNN to generate features for use in the target cost was more effective than using an HMM, be that using the speech parameters generated at the output of the DNN or using context embeddings from a bottleneck layer.

In this investigation, only the target cost function was modified. However, further increases might be obtained in future work by improving the join cost function [36].

Although we found no significant differences between the use of speech parameters from a DNN compared to context embeddings, there is perhaps more consistency in listener judgements for the embedding-based systems ($LE, HE$) than the DNN speech parameter-based system ($NP$).

The context embedding features discussed here could be used elsewhere in the unit selection system, by using these features as a back-off function, replacing manual phone substitution rules, or to perform the initial pre-selection of units, instead of the current pre-selection of units by matching diphone identity.

Investigating different types of neural network for use in this hybrid framework is left as future work, but we expect that any improvement in parametric synthesis would carry over to the hybrid method for waveform generation. For example, mixture density networks (MDNs) might be used to directly produce a likelihood-based target cost instead of the KLD-based approach used here. Recurrent neural network (RNNs), which are more powerful sequence model, might also be used to generate the target trajectories.

Finally the systems investigated here made use of uniform subsections of diphones (the waveform concatenation unit) in conjunction with phoneme-level acoustic models. However investigations into using diphone-level acoustic models is of interest as this would allow state-sized representations of speech to be used instead and may further improve performance.

## 8. ACKNOWLEDGEMENTS

Thanks to Oliver Watts for his suggestions for tuning the target cost weight and to Gustav Eje Henter for help with the MUSHRA test implementation and analysis. This research was supported by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST). The NST research data collection may be accessed at
http://datashare.is.ed.ac.uk/handle/10283/786.

## 9. REFERENCES


[1] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge*, 2011.

[2] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2012," in *Proc. Blizzard Challenge*, 2012.

[3] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2013," in *Proc. Blizzard Challenge*, 2013.

[4] Simon King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014.

[5] Thomas Merritt and Simon King, "Investigating the shortcomings of HMM synthesis," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013, pp. 165–170.

[6] Thomas Merritt, Tuomo Raitio, and Simon King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," in *Proc. Interspeech*, 2014, pp. 1509–1513.

[7] Gustav Eje Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, 2014, pp. 1504–1508.

[8] Thomas Merritt, Javier Latorre, and Simon King, "Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech," in *Proc. ICASSP*, 2015.

[9] Thomas Merritt, Junichi Yamagishi, Zhizheng Wu, Oliver Watts, and Simon King, "Deep neural network context embeddings for model selection in rich-context HMM synthesis," in *Proc. Interspeech*, 2015.

[10] Andrew J Hunt and Alan W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[11] Alan W Black and Paul A Taylor, "Automatically clustering similar units for unit selection in speech synthesis.," in *Proc. Eurospeech*, 1997.

[12] Paul Taylor, Alan W Black, and Richard Caley, "The architecture of the festival speech synthesis system," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, 1998.

[13] Paul Taylor, "The target cost formulation in unit selection speech synthesis.," in *Proc. Interspeech*, 2006, pp. 2038–2041.

[14] Yao Qian, Frank K Soong, and Zhi-Jie Yan, "A unified trajectory tiling approach to high quality speech rendering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 280–290, 2013.

[15] Zhen-Hua Ling, Heng Lu, Guo-Ping Hu, Li-Rong Dai, and Ren-Hua Wang, "The USTC system for Blizzard Challenge 2008," in *Proc. Blizzard Challenge*, 2008.

[16] Zhi-Jie Yan, Yao Qian, and Frank K Soong, "Rich-context unit selection ( RUS ) approach to high quality TTS," in *Proc. ICASSP*, 2010, pp. 4798–4801.

[17] Vincent Pollet and Andrew Breen, "Synthesis by generation and concatenation of multiform segments.," in *Proc. Interspeech*, 2008, pp. 1825–1828.

[18] Alexander Sorin, Slava Shechtman, and Vincent Pollet, "Uniform Speech Parameterization for Multi-form Segment Synthesis," in *Proc. Interspeech*, 2011, pp. 337–340.

[19] Alexander Sorin, Slava Shechtman, and Vincent Pollet, "Psychoacoustic Segment Scoring for Multi-Form Speech Synthesis.," in *Proc. Interspeech*, 2012, pp. 2214–2217.

[20] Alexander Sorin, Slava Shechtman, and Vincent Pollet, "Refined Inter-segment Joining in Multi-Form Speech Synthesis," in *Proc. Interspeech*, 2014, pp. 790–794.

[21] Raul Fernandez, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory, "Using Deep Bidirectional Recurrent Neural Networks for Prosodic-Target Prediction in a Unit-Selection Text-to-Speech System," in *Proc. Interspeech*, 2015.

[22] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, 2015.

[23] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[24] Heiga Zen, "Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN," in *Proc. MLSLP*, 2015.

[25] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 226–239, 1998.

[26] Zhi-Jie Yan, Yao Qian, and Frank K Soong, "Rich context modeling for high quality HMM-based TTS," in *Proc. Interspeech*, 2009, pp. 1755–1758.

[27] Robert AJ Clark, Korin Richmond, and Simon King, "Festival 2–build your own general purpose unit selection speech synthesiser," in *Proc. SSW5*, 2004.

[28] Robert AJ Clark, Korin Richmond, and Simon King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[29] John R. Hershey and Peder A. Olsen, "Approximating the Kullback-Leibler divergence between Gaussian mixture models," in *Proc. ICASSP*, 2007.

[30] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW6*, 2007, pp. 294–299.

[31] Zhizheng Wu and Simon King, "Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features," in *Proc. Interspeech*, 2015.

[32] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao, "Hurricane natural speech corpus, [sound]," LISTA Consortium, doi:10.7488/ds/140, 2013.

[33] International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland, *Objective measurement of active speech level*, March 2011.

[34] International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, *Method for the subjective assessment of intermediate quality level of coding systems*, March 2003.

[35] Thomas Merritt, Robert A. J. Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King, "Listening test materials for "Deep neural network-guided unit selection synthesis", 2016 [dataset]," University of Edinburgh, The Centre for Speech Technology Research (CSTR), doi:10.7488/ds/1313.

[36] Alistair D. Conkie and Stephen Isard, "Optimal coupling of diphones," in *Progress in speech synthesis*, pp. 293–304. Springer, 1997.