The CSTR entry to the Blizzard Challenge 2016

Thomas Merritt, Srikanth Ronanki, Zhizheng Wu, Oliver Watts

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

T.Merritt@ed.ac.uk, srikanth.ronanki@ed.ac.uk

Abstract

This paper describes the text-to-speech system entered by The Centre for Speech Technology Research into the 2016 Blizzard Challenge. This system is a hybrid synthesis system which uses output from a recurrent neural network to drive a unit selection synthesiser. The annual Blizzard Challenge conducts side-by-side testing of a number of speech synthesis systems trained on a common set of speech data. The task of the 2016 Blizzard Challenge is to train on expressively-read children's storybooks, and to synthesise speech in the same domain. The Challenge therefore presents an opportunity to test the effectiveness of several techniques we have developed when applied to expressive speech data.

Index Terms: hybrid synthesis, statistical parametric speech synthesis, deep neural network, recurrent neural network, unit selection

1. Introduction

The CSTR entry to this year's Blizzard Challenge builds on the hybrid Multisyn [1, 2] system introduced in [3]. Hybrid synthesis brings the benefits of extremely natural-sounding unit selection (which is unaffected by the degradations introduced by vocoding [4, 5, 6]), whilst also exploiting the flexibility of statistical parametric speech synthesis (SPSS). The data used for this year's Challenge was obtained from professionally-read child-directed audio books and is therefore much more prosodically rich than the more standard prompt-based speech data used in [3]. The amount of data (5 hours) is also greater than used in [3] (2 hours).

The experiment presented in [3] established that improving the underlying SPSS of a hybrid synthesiser results in improvements to the concatenated output speech. The current system therefore incorporated two major improvements to the underlying SPSS model compared to the system prensented in [3]: the decision tree duration model is replaced with a bi-directional long short-term memory (LSTM) recurrent neural network, and the feed-forward DNN acoustic model is replaced with an LSTM network. The neural networks used in this entry were trained using our open-source neural network TTS toolkit, Merlin [7].

2. System Description

2.1. Data

The database – provided to the Challenge by Usborne Publishing Ltd. – consists of the speech and text of 50 children's audiobooks spoken by a British female speaker. We made use of a segmentation of the audiobooks carried out by another Challenge participant[1] and kindly made available to other participants. The total duration of the audio is approximately 4.33 hours after segmentation. Three audiobooks from the given corpus were held out to act as an internal development set to gauge system performance before generating the final test data. The held-out data consists of three whole short stories: *Goldilocks and the Three Bears*, *The Boy Who Cried Wolf* and *The Enormous Turnip*, having a total combined duration of approximately 10 minutes.

2.1.1. Sentence selection

Harnessing the variety of speaking styles present in expressively-read audiobooks might enable us to produce less robotic-sounding TTS systems. However, initial experiments showed that the extreme variation in parts of the training data for the Challenge resulting in poor unit selection. We therefore filtered the data using the active learning approach described in [8]: 198 utterance-level acoustic features are extracted, and 15 sentences initially labelled as *keep* or *too expressive* by an expert listener. Uncertainty sampling [9] using an ensemble of decision trees was then used to select a further informative sample to be hand-labelled; this process continued for 20 minutes (real time). A classifier built on the entire set of hand-labelled data was then used to determine the subset of available sentences to be used for training. 11.5% of the training sentences were discarded in this way; informal comparison suggested this resulted in more stable synthesis with fewer unwarranted prosodic excursions.

2.2. Text processing

We used the Festival English front-end with the British Received Pronunciation version of the Combilex lexicon [10]. 127 items were added to cover words appearing in the training data but otherwise absent from the dictionary. There were slight differences in the lexicon-lookup procedures used in preparing the annotation for training the SPSS model and those employed by the Festival front-end used for Multisyn. The resulting inconsistencies were dealt with by aligning the DNN's phone sequences to those expected by Multisyn in an ad hoc fashion. Given sufficient time to retrain the system from scratch, we expect making label creation consistent across the SPSS and unit selection modules of our system to lead to improved synthesis quality.

2.3. Parametric system

The parametric system was implemented using LSTMs in a conventional two-stage approach. In the first stage, a duration model is used to predict phone durations to form frame-level linguistic features. In the second stage, an acoustic model is used to generate parameters from those linguistic features.

[1] Innoetics: https://www.innoetics.com

2.3.1. Duration model

The duration model trained for our entry to the Challenge made use of a modified version of the multi-level modelling approach with LSTM mixture density networks (MDNs) proposed in [11] for robust duration modelling. We exploit the benefits of including long-range dependencies in duration prediction by using recurrent neural networks and by simultaneously predicting durations at multiple levels (state, phone, syllable and word). In [11], phone-level duration was used as a multi-task side-objective, alongside the main task of predicting the durations of the states within the phone. The phone-level prediction is discarded at run-time, but requiring the network also to make this prediction results in improvements on the main task. We here extend this approach to include also the syllable- and word-level. Furthermore, the duration model used is *statistically robust*: by training a multi-component MDN, some components can be used to account for bad data (garbage components). By then synthesising from a single mixture component (e.g. the one with the largest mass), datapoints that trained the other components – and the behaviour that led to those datapoints – are ignored in synthesis.

The approach described was used only to predict durations for forming frame-level linguistic features as input for the prediction of acoustic parameters. The hybrid Multisyn unit-selection system, however, doesn't make use of any duration-derived features in its target cost function. Including features based on our robust multi-level duration model in the unit selection process is left for future work.

2.3.2. Acoustic model

The durations predicted by the bi-directional LSTM described above are combined with linguistic features derived from a set of questions about linguistic context to create the frame-level linguistic features which is the input to the uni-directional LSTM RNN acoustic model. This LSTM RNN is then trained at the frame-level to map from the linguistic context to vocoder parameters (static, delta and delta-delta features of 60 Mel-cepstra, 25 BAPs and log-f_0) and a binary feature denoting whether the frame is voiced or unvoiced. Following the prediction of the frame-level vocoder parameter distributions, maximum likelihood parameter generation (MLPG) and postfiltering are performed to arrive at the final generated parameter trajectories.

In SPSS these parameter trajectories would then be passed through the vocoder to produce a speech waveform. Instead, we use them as targets for selecting waveform units as follows. First, the synthesised parameters for each phoneme are split uniformly across time into 4 section. In each of the 4 sections, a Gaussian distribution is fitted to each of the vocoder parameters. The variances of these Gaussian distributions are floored at 1% of the global variance per parameter, following [3]. These 4 uniform sections per phone allow diphone representations to be created from the phone predictions produced by the SPSS system: 2 sections from each of the phones associated with a diphone are used to create a representation for that diphone.

Comparable distributions were generated for the candidates in the unit database, based on vocoder parameters derived from the training data and natural durations obtained by forced alignment.

2.3.3. Feature extraction for acoustic model training

We obtained a state-level forced alignment of the sentence-segmented data described above using context independent HMMs, similar to [12]. Festvox's `ehmm` [13] was used to insert pauses into the annotated phone sequences based on the acoustics. Each phone was then characterised by a vector of 481 text-derived binary and numerical features: these features are a subset of the features used in decision-tree clustering questions from the HTS public demo [14]; numerical features queried by those questions were used directly where possible.

For duration modelling, all these features were used as input and normalised to the range of $[0.01,\ 0.99]$. The output for training is an eight-dimensional vector of durations for every phone, comprising five sub-state durations, the overall phone duration, syllable duration and whole word duration. We use this form of multi-task learning to improve the model; the three additional features (phone, syllable, and word durations) act as a secondary task to help the network learn more about suprasegmental variations in duration at word level.

For acoustic modelling, the input uses the same features as duration prediction, to which 9 numerical features were appended. These capture frame position in the HMM state and phoneme, state position in phoneme, and state and phoneme duration, similar to [12]. For output features, STRAIGHT [15] was used to extract 60 mel-cepstrum coefficients, 25 band aperiodicities, logarithmic fundamental frequency ($\log F_0$) along with delta and delta-delta features every 5ms. Unvoiced regions of $\log F_0$ were linearly interpolated before computing delta and delta-delta features. To which, a binary feature denoting whether the frame is voiced or unvoiced was added. For both the duration and acoustic data, a per-component mean and variance normalisation was applied prior to model training, with the transformation reversed as part of synthesis.

2.3.4. Duration and acoustic model training

The duration model used phone-level linguistic features as input and are optimised to predict the (mean and variance normalised) duration of the phones in the training data. For training, the model was configured with five feed-forward layers of 1024 nodes each and a final bi-directional LSTM hidden layer consisting of 512 nodes. The output layer was configured with a single-component, maximum-likelihood Gaussian MDN. Whereas, the acoustic model used frame-level linguistic features as input to predict the vocoder parameters. For training, the model was configured with five feed-forward layers of 1024 nodes each and a final uni-directional simplified LSTM hidden layer consisting of 512 nodes [16].

Both the networks were initialised using small random weights, with no pre-training. Each prediction system was trained with a fixed learning rate, manually tuned to yield close-to-optimal results on the development set in 30 epochs or less. Early stopping was used to avoid overfitting, by aborting training once the objective function on the development set had failed to improve for five epochs. The neural network training was performed broadly as for the basic system described in [12, 16] using Merlin [7].

2.4. Unit selection waveform renderer

A modified form of Festival's Multisyn engine [2] was used for the unit selection stage of our system. To compare the suitability of a given candidate diphone in the unit database with the 4 distributions representing a synthesised diphone, the symmetrised

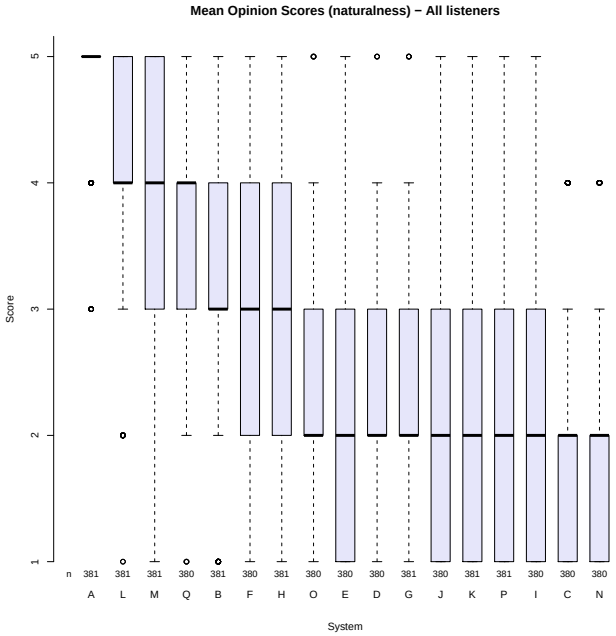

Figure 1: Our system(Q): Mean opinion score for naturalness of the synthesized speech with ratings from all listeners.

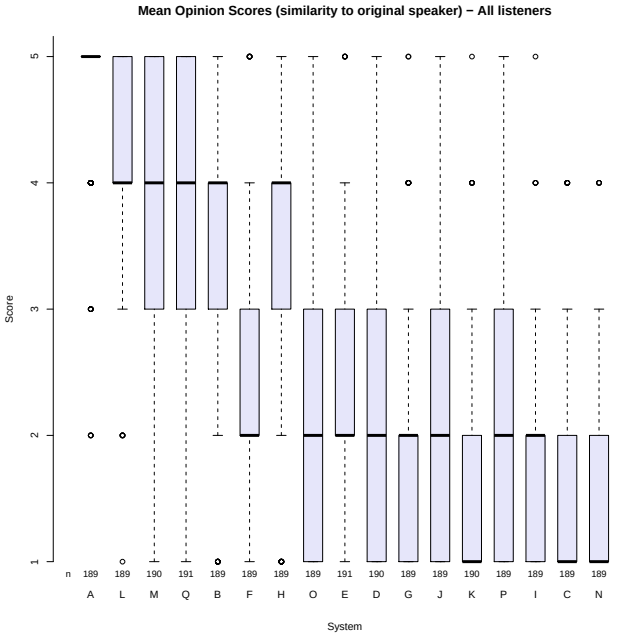

Figure 2: Our system(Q): Mean opinion score for speaker similarity with ratings from all listeners.

Kullback Leibler divergence (KLD) [17] is used. The KLD is computed between each of the 4 candidate unit's distributions and the corresponding target unit distributions individually. The resulting 4 scores are then summed to produce the final target score.

The standard Multisyn join cost (sum of distances between 12 MFCCs, f_0 and energy from the frame either side of the join) is retained, as well as the standard pre-selection criterion of candidate units (by matching diphone identity). The standard Multisyn Viterbi search (with pruning to reduce the search time) is performed in order to optimise target cost and join cost. Also the standard Multisyn back-off rules are used where the target diphone to be synthesised is not present in the training data.

2.5. Speech synthesis

At synthesis time, duration is predicted first, and is used as an input to the acoustic model to predict the speech parameters. Maximum likelihood parameter generation (MLPG) using pre-computed variances from the training data is applied to the output features for synthesis, and global-variance (GV statistics computed from training material) is applied to the resulting MCC trajectories. These parameter trajectories are then used to produce diphone coefficients. The Festival Multisyn engine was used to compute the target and joint cost between target unit and pre-selected candidate units to select the final candidate, as explained above. The final waveform synthesis was done by joining the selected units. No additional smoothing or post-modification of prosody was performed after joining the units: this is left for future work.

From the sentences synthesised in this way, files were made containing whole paragraphs, chapters and books as required by the Challenge by simply concatenating the waveforms. While proper exploitation of long-distance contexts ought to improve synthesis quality, no contexts outside the current sentence were used for the present submission.

3. Results

The identifier for our system in the published results is Q.

3.1. Naturalness

We first consider the results for naturalness (making use of the published statistical analysis of significant differences between naturalness of systems at the 1% level with Bonferroni corrected alpha) [18]. Mean opinion scores for naturalness from all listeners on book sentences are shown in Figure 1. Our system outperformed all three baseline systems(B–D). Among the 13 challenge participants, our system is outperformed by only a single system (L). The same trend can be seen across the scores made by paid listeners, speech experts and on-line volunteers.

3.2. Speaker similarity

We now consider mean opinion scores for speaker similarity. The mean opinion scores for speaker similarity from all listeners on book sentences are shown in Figure 2. Considering ratings from all listeners (or any other listener group), no other system was significantly better than ours and our system was in turn significantly better than 11 other systems. These results show the effectiveness of waveform concatenation systems for speaker similarity.

3.3. Evaluation of audiobook paragraphs

We now consider the results for evaluation of audiobook paragraphs – that have been evaluated on several other factors like stress, intonation, emotion, pleasantness, listening effort, speech pauses and overall impression. Considering ratings from all listeners on overall impression, our system showed similar performance as in the case of the isolated sentence evaluation of naturalness and speaker similarity. Only two systems (L and M) outperformed us and our system was significantly better in turn than the remaining systems (cf. Figure 3). Considering ratings from paid listeners on overall impression, only system L outperformed ours. Considering ratings for other individual fac-

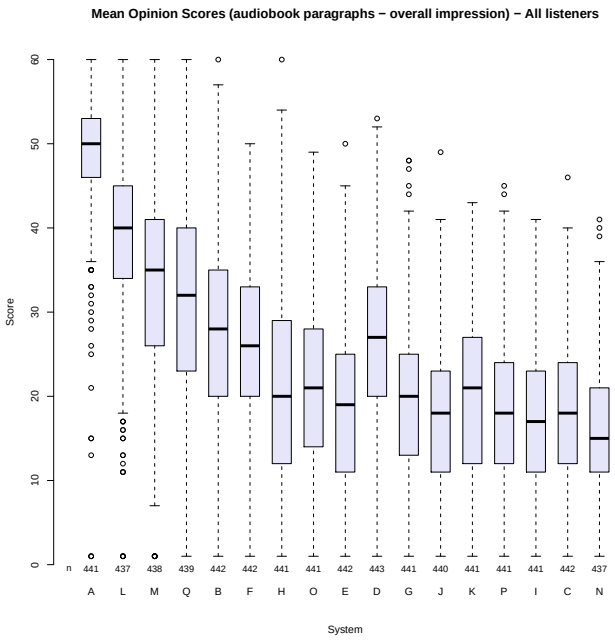

Figure 3: Our system(Q): Mean opinion score for overall impression with ratings from all listeners.

tors (e.g., stress, intonation and pleasantness) from all listeners, only the aforementioned systems L and M consistently outperformed ours. Overall, our system outperforms between 9 and 13 other systems in evaluation of each of these factors, performing best in emotion and pleasantness.

3.4. Intelligibility (SUS)

We now consider the results for intelligibility of semantically unpredictable sentences (making use of the published statistical analysis of significant difference between word error rates of the systems). Taking into account ratings from all listeners, there are only three other systems out of 16 (D, F, L) significantly better than ours. Considering only paid listeners, there are only two other systems (F and L) significantly better than ours. Out of 16 systems evaluated by paid listeners, 10 were not significantly more or less intelligible than ours, 3 were significantly less intelligible, and only 2 significantly more intelligible. The results show that our system is quite effective on intelligibility as well. Overall, our system has shown consistent performance (standing in top four) in all the factors evaluated for the Challenge.

4. Conclusions & future work

For this year's CSTR Blizzard Challenge entry the hybrid system introduced in [3] was improved (both its duration model and acoustic model) and applied for the first time to expressive speech data.

The results of the evaluation are on the whole very positive, but there are still a number of potential future improvements which could be made to the hybrid synthesis system described here. These include adopting consistent lexicon-lookup for both the SPSS and unit selection systems, performing modifications to smooth the joins between units, and the explicit inclusion of predicted duration in the unit selection synthesis target cost.

5. Acknowledgements

We thank Robert A. J. Clark for useful discussion and advice. 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`.

6. References

[1] R. A. Clark, K. Richmond, and S. King, "Festival 2–build your own general purpose unit selection speech synthesiser," in *Proc. SSW*, 2004.

[2] R. A. Clark, k. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[3] T. Merritt, R. A. J. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Proc. ICASSP*, 2016.

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

[5] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. 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.

[6] T. Merritt, J. Latorre, and S. King, "Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech," in *Proc. ICASSP*, 2015.

[7] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proc. SSW*, Sunnyvale, USA, 2016.

[8] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, Aug. 2013, pp. 121–126.

[9] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*.

[10] S. Fitt and K. Richmond, "Redundancy and productivity in the speech technology lexicon - can we do better?" in *Proc. Interspeech 2006*, Sep. 2006.

[11] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust TTS duration modelling using DNNs," in *Proc. ICASSP*, vol. 41, Shanghai, China, March 2016, pp. 5130–5134.

[12] 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 *Proc. ICASSP*, 2015, pp. 4460–4464.

[13] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. ICASSP*, 2006, pp. I–853–I–856.

[14] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW*, vol. 6, 2007, pp. 294–299.

[15] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[16] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. ICASSP*, 2016, pp. 5140–5144.

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

[18] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," *Proc. Blizzard Challenge Workshop*, 2007.