Light Supervised Data Selection, Voice Quality Normalized Training and Log Domain Pulse Synthesis

Gilles Degottex, Pierre Lanchantin, Mark Gales

University of Cambridge, United Kingdom

gad27@cam.ac.uk, pkl27@cam.ac.uk, mjfg100@cam.ac.uk

Abstract

Training acoustic models with, and synthesising, expressive speech is a challenge for Text-to-Speech (TTS) systems. The 2017 Blizzard Challenge offers an opportunity to tackle this problem by releasing data from "lively" recordings of children books. This paper describes the System J submission to the Blizzard Challenge 2017 - Task EH1. Three potential approaches to handling expressive speech within a DNN-based system are discussed. First, mistranscribed and outlier content can be removed from the training data by using lightly-supervised training approaches. Second, the impact of paralinguistic information that cannot be predicted by the contextual labels is handled by marginalising out these aspects when training the acoustic model. This should reduce the implicit averaging effect that normally occurs. Finally, the system makes use of a new vocoder that has the potential to be more flexible than other state-of-the-art solutions. Results of the Challenge show that, even though the intelligibility and pauses are of reasonable quality and an internal test shows improvements using the new vocoder, the marginalisation over the voice quality removed most of the intonation and expressivity, leading to more degradation of the overall impression than expected.

Index Terms: parametric speech synthesis, pulse model

1. Introduction

Text To Speech (TTS) systems usually need many components that try to reproduce every element that mimic a speaker. From phonetics to signal processing in a statistical modelling framework, these systems are quite complex even though recent results [1, 2] promise simpler pipelines in the future. Specific comparisons are necessary to study some specific components. Participation to a challenge that assesses many systems on overall criteria are also necessary in order to provide an overview of TTS techniques. TTS systems are often based on pre-recorded voices dedicated to this task (e.g. [3]). However, existing recordings can be used to alleviate the burden of making dedicated recordings. This type of recordings often provide a higher degree of variations in terms of expressiveness, acting, voice qualities, etc. namely paralinguistic information. The recording of a voice dedicated to a TTS system is often tailored and directed in a way that simplifies its processing and modelling at the cost of over-simplifying the voice. Modelling the voice of a speaker reading children books is, therefore, a challenging and necessary task for developing more flexible and expressive TTS systems.

For our submission to this Blizzard Challenge, we chose to use a DNN-based Speech Parametric Synthesis System (SPSS) [4], because we believe this approach should provide a flexibility that concatenative-based synthesis cannot offer in many applications. However, the paralinguistic information carried in a lively reading is quite difficult to model for current SPSS systems. While concatenative systems can blindly reproduce the paralinguistic information by copying the speech content and limit discontinuities, SPSS systems have no choice but to predict it based on the contextual input information. The input being usually contextual phonetic content, it is not correlated to the produced paralinguistic information and the SPSS systems have little chances to model it appropriately. Because there is usually no input that help discriminating the use of paralinguistic content, the statistical model end up averaging the spectral representation of the waveform. This results in muffling effects and increases noise in transients that post-processing techniques usually attempt to alleviate.

In this paper, we tried to deal with this extra variability at three different levels. The first suggested idea is to remove erratic data from the training data. Even though that might seems to be a drastic choice, it is safe to believe that some expressions, onomatopoeia and exaggerations made by the speaker can be put aside in order to avoid outliers in the training data while preserving most of the original voice variability. Our second idea is to normalise the voice during training with respect to the voice quality. We can first assume that it exists a component of the paralinguistic information that is correlated to the voice quality. Our hypothesis is that if we marginalise the acoustic model over the voice quality, the DNN model can focus on the phonetic content and the paralinguistic component that is predictable from the contextual phonetic labels. We expect the resulting voice to be more consistent, less muffled, overall less averaged. We assume that the component of the paralinguistic information which is correlated to the contextual phonetic labels will preserve enough variability of the original voice (intonations, expressivity, etc.). Therefore, this approach is sort of a bet since we do not know a priori the balance between the phonetic-correlated paralinguistic information and that correlated to voice quality. We hope that enough of the paralinguistic information can be preserved through the contextual phonetic labels while the excess can be marginalise by the voice quality. The third and last idea is to use a novel vocoder for parameterizing the waveform. We suggest to use the Pulse Model in Log-domain (PML) that has been recently presented [5]. This vocoder makes use of a noise model that is convolutive instead of being additive with the deterministic content, conversely to the traditional source-filter model. This noise model makes use of a binary mask to activate noise in the time-frequency plan. For the initial presentation of PML, we suggested to model this mask through an intermediate Phase Distortion Deviation (PDD) feature [5]. In this submission, we modelled the noise mask directly by the DNN using an adapted output layer.

The next section describes the system used for synthesizing the sentences of system J. Namely, the overall structure is first described, then the three innovative elements are detailed that correspond to the main differences one can find between known systems [4] and our submission. The last section presents the results of the listening tests carried out for this challenge.

2. System J

The overall process follows the Merlin SPSS pipeline [4], which uses a sentence by sentence architecture. For both training and testing stage, text is first converted to phonetic labels and linguistic contexts are append to these labels. At training stage, an HMM-GMM model was first trained in order to align the context labels on their corresponding waveforms [6]. Acoustic features were then extracted from the waveforms and a DNN-based acoustic model was trained in order to predict the acoustic features from the context labels. For this acoustic model, the parameters of a 3-stacked Bidirectional LSTM (BLSTM) of 1024 units were optimised by gradient descent. At testing stage, durations of context phonetics labels are predicted using the HMM-GMM model, and the acoustic features predicted by the BLSTM are used to resynthesize a waveform (see Fig. 1). In addition to this relatively standard pipeline, the three novelties discussed in the introduction are added in our submission and described below.

2.1. Alignment and duration prediction

To align context labels on the recordings, an HMM-GMM HTS system [6] was first trained using five-state, left-to-right, no-skip hidden semi-Markov models (HSMMs [6]). STRAIGHT's features were used for these alignments. The rest of the topology of the HMM models and systems was similar to the one used for the Nitech-HTS system ([6]). Multiple iterations of training and re-alignment provided state-aligned phonetic labels used for training the acoustic model. In order to produce inputs to the acoustic model, one-hot encoding was used to represent the state-aligned context labels providing 592 binary input features. 9 linear numerical input features were also added representing the position of the label within the sentence, word, phoneme, etc.

Similarly to previous submissions to the Blizzard Challenge 2016 [7], we added one extra binary flag to the input features that is representing the neutral/expressive state of the text. This context flag was simply obtained by locating the segments of text between quotes.

During testing, the duration of the context labels was predicted using the HMM-GMM system.

2.2. Light supervised training for data selection

The first idea to deal with the variability of the voice is to discard the data that seem to be outliers and might degrade the modelling uselessly.

A lightly supervised approach [8] was first used for the alignment and the selection of the training data. The output from a speech recogniser, using a language model biased towards the original transcripts, was compared to the original transcripts and a Phone Matched Error Rate (PMER) computed between the two for each recognised segment. The maximum PMER allows segments to be selected for training while ensuring that the word/phone supervision information is reasonably accurate. Segments corresponding to text between quotes were tagged in an attempt to identify expressive speech. Pauses longer than 60ms were also marked in the transcriptions. A total of 2h50mn of speech from 40 of the provided audiobooks was aligned and selected with PMER=0% including 38mn of marked expressive speech.

2.3. Voice quality normalized training

The variability of the speaker's voice adds a layer of difficulty compared to voices that are recorded for TTS purposes (e.g. Arctic databases [3]). In technical terms, the standard input labels are too poor for discriminating all the possible instances of a single phoneme and, as a consequence, the voice parameters are averaged. The usual perceptual result is an effect of muffling on the vowels and increased noise in the transients.

The only way to improve the discriminative capabilities of the network is to enrich its input features. In previous Blizzard Challenges [7], some participants added TOBI features for this purpose. In this work, knowing that the variability of the output is partly due to paralinguistic information which is correlated to the the voice quality, we chose to contextualise the training using a voice quality features vector. Namely, in addition to the text-related inputs (phonetic labels, sentence structure, expressivity flag, etc.), we concatenated an extra vector of 11 voice quality features that are computed from the target waveform (See Fig. 1, left side). Those voice quality features were computed using the COVAREP repository [9] (Normalised Amplitude Quotient, Quasi-Open Quotient, H1-H2, Harmonic Richness Factor, Parabolic Spectral Parameter, Cepstral Peak Prominence, Maxima Dispersion Quotient, Peak Slope, Maximum Voiced Frequency, Rd glottal parameter and its confidence value). During training time, we do not want the DNN to rely on these extra features for predicting the phonetic content of the waveform. We only want these features to marginalise the training over the voice quality variance. Thus, in order to remove any phonetic content from these features, we averaged them across voiced segments and interpolated the values in the unvoiced segments (See Fig. 1, left side). During testing time, since we do not know the voice quality features of the target sentence, the voice quality feature vector is replaced by an averaged vector (See Fig. 1, right side) computed from the voice quality features extracted at the middle state of all possible vowels of the training data. Knowing that using an average voice quality feature vector might select an average voice, the training is still able to better discriminate the outputs with respect to the given inputs. Thus, we assume that selecting an average input vector ourself after a training that could to map predictable data should be better than letting the neural network face unpredictable data and select an average output on its own.

Future work might want to predict this voice quality feature vector based on textual inputs of the full paragraph in order to recover part of the voice variability which is lost during this marginalisation.

2.4. Noise mask modelling for a pulse-based vocoder

This section presents the new vocoder used in this submission as well as the special output layer used to model the noise mask, conversely to the initial presentation [5].

2.4.1. Pulse Model in Log-domain (PML): Analysis/Synthesis

The PML synthesis process needs the following features that are illustrated in Fig. 2: A fundamental frequency curve $f_0(t)$, which does not exhibit voicing decisions; The REAPER f_0 estimator was used in this work [10] and the zero values were filled by linear interpolations between voiced segments and extrapolated at the beginning and end of the signal. The VTF response $V(t, \omega)$, which is assumed to be minimum phase. The spectral envelope estimate provided by STRAIGHT vocoder [11] was used in this work and compressed on a mel scale of 60 coef-

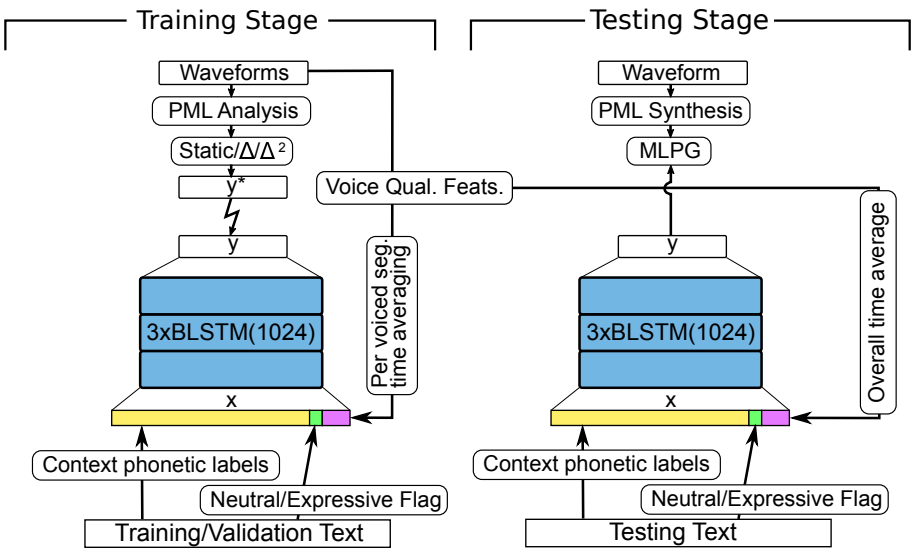

Figure 1: *Architecture of our BLSTM-based pipeline in training ans testing stages.*

ficients; A binary mask $M(t,\omega)$ in the time-frequency space. Here 0 is for deterministic regions and 1 for noisy regions. In this work, this mask is derived from the Phase Distortion Deviation (PDD) [12] PDD(t,ω) as described below. For statistical modelling, this mask is compressed on 24 frequency bands whose bandwidths follow a Bark scale.

Since $f_0(t)$ and $V(t,\omega)$ are extracted using state-of-the-art methods previously published (REAPER and STRAIGHT, respectively), we describe here only the computation of the noise mask. The Phase Distortion Deviation (PDD) [12, 13, 14, 13] is used for this purpose and the mask is obtained by thresholding the PDD values.

In order to compute PDD, the Phase Distortion (PD) at each harmonic frequency is first computed [12]:

$$\text{PD}_{i,h} = \phi_{i,h+1} - \phi_{i,h} - \phi_{i,1} \tag{1}$$

where $\phi_{i,h}$ is the phase value at frame i and harmonic h, as measured by a sinusoidal model [15, 16, 17]. A step size of one forth of a fundamental period was used in this work to split the analysed signal into frames as in [12]. PDD is then computed as the short-term standard-deviation of PD:

$$\begin{aligned} \text{PDD}_i(\omega) &= \operatorname*{std}_i\left(\text{PD}_i(\omega)\right) \\ &= \sqrt{-2\log\left|\frac{1}{N}\sum_{n\in C} e^{j(\text{PD}_n(\omega))}\right|} \end{aligned} \tag{2}$$

where $C = \{i - \frac{N-1}{2}, \cdots, i + \frac{N-1}{2}\}$, $N = 9$ in this work and $\text{PD}_i(\omega)$ is the continuous counterpart of $\text{PD}_{i,h}$ obtained by linear interpolation across frequency.

In [12], it is shown that the measurement of phase variance saturates as the variance increases. Consequently, a threshold of 0.75 was used to force the variance to a fixed and higher value in order to ensure the proper randomization of the noise segments. Therefore, in this work the same threshold was used for building the mask: $M(t,\omega) = 1$ if PDD$(t,\omega) > 0.75$ and zero otherwise.

The generation of the waveform follows a pulse-based procedure, similarly to the synthesis process of the STRAIGHT vocoder. Short segments of speech signals, called *pulses* (roughly the size of a glottal pulse) are generated sequentially. In both voiced and unvoiced segments, the voice source of each pulse, is made of a morphing between a deterministic impulse and Gaussian noise. This source is then convolved by the Vocal Tract Filter (VTF) response and then overlapped-add with the other pulses. The paragraphs below describe the details of this procedure.

A sequence of pulse positions t_i are first generated all along the speech signal according to the given $f_0(t)$ feature:

$$t_{i+1} = t_i + 1/f_0(t_i) \tag{3}$$

with $t_0 = 0$. Then, to model the speech signal around each instant t_i, the following simple formula is applied:

$$S_i(\omega) = e^{-j\omega t_i} \cdot V(t_i,\omega) \cdot N_i(\omega)^{M(t_i,\omega)} \tag{4}$$

where $N_i(\omega)$ is the Fourier transform of a segment of Gaussian noise starting at $\frac{t_{i-1}+t_i}{2}$ and finishing at $\frac{t_i+t_{i+1}}{2}$, whose central instant t_i is re-centered around 0 (to avoid doubling the delay $e^{-j\omega t_i}$ for the noise in $S_i(\omega)$). Additionally, the noise $N_i(\omega)$ is normalized by its energy to avoid altering the amplitude envelope that has to be controlled by $V(t,\omega)$ only.

The first complex exponential defines the overall position of the voice source (e.g. the position of the Dirac impulse of the deterministic source). $V(t_i,\omega)$ defines the amplitude spectral envelope and its minimum phase. $M(t_i,\omega)$ provides the means to switch between deterministic or noisy voice source at any time-frequency point.

In order to build the complete speech signal from the pulses generated by (4), overlap and add is applied:

$$\check{s}(t) = \sum_{i=0}^{I-1} \mathcal{F}^{-1}\Big(S_i(\omega)\Big) \tag{5}$$

where I is the number of pulses in the synthesized signal.

2.4.2. *PML: Noise Mask (NM) modelling*

In the first presentation of PML [5] for TTS, PDD was predicted by the acoustic model and then thresholded in order to produce

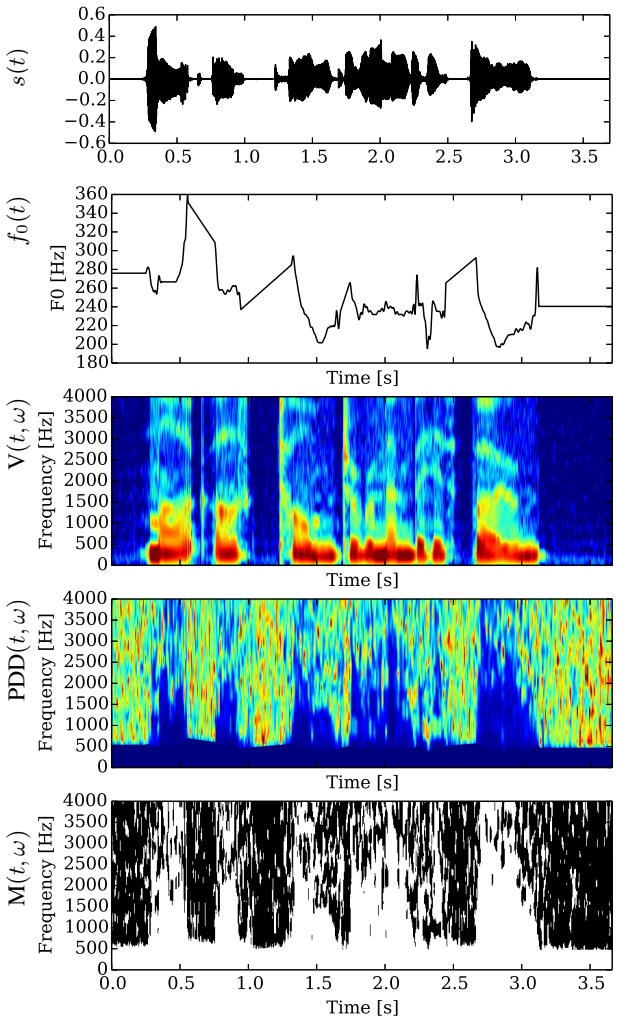

Figure 2: *From top to bottom: a recorded waveform used to extract the following elements; The continuous fundamental frequency curve $f_0(t)$; the amplitude spectral envelope $V(t,\omega)$; the Phase Distortion Deviation PDD(t,ω) (a measure of phase randomness. The warmer the colour, the bigger the PDD value and the noisier the corresponding time-frequency region); the binary mask $M(t,\omega)$ derived from PDD, which allows to switch the time-frequency content from deterministic (white) to random (black). The features that are necessary for PML synthesis in this submission are only: $f_0(t)$, $V(t,\omega)$ and $M(t,\omega)$.*

the binary mask given to the PML synthesis process. For this submission, the noise mask is directly modelled by the acoustic model.

When modelling PDD, its first and second approximate derivatives are normalized by their mean and variance. However, when modelling NM directly, its values are already bounded in $[0, 1]$. Thus, it does not seem necessary to normalise NM. Moreover, using a linear output for these values is not advised as the DNN would have to model the boundaries at 0 and 1 whereas they are known a priori. For this reason, we modelled the static NM values using a sigmoid output function. For the 1st and 2nd approximate derivatives, we used hyperbolic tangent normalized in amplitude to 0.5 and 2, respectively, to match the values' intervals given by the windows used for the derivatives' approximation. Note that this leads to a mix output layer in the acoustic model where the first 183 $(3 \cdot 60 + 3 \cdot 1)$ values are linear outputs, as in STRAIGHT-based systems, and the remaining 72 values $(3 \cdot 24)$ are non-linear outputs.

In the following, results of an experiment are presented in order to evaluate the impact of the noise mask model on the overall quality, by modelling either PDD, or the noise mask directly as described above. The two noise model were also compared to a STRAIGHT-based synthesis. For the STRAIGHT synthesizer, the output features were the same as the ones used for the HTS systems used for the alignment (see above). Input features were normalised to [0.01, 0.99] and output features were normalised to zero mean and unit variance. The same 60 Mel-cepstral coefficients and log f_0 values were used for the 3 systems, only the noise features were different (aperiodicity, mel-PDD and NM for STRAIGHT, PML-PDD and PML-NM, respectively).

A Comparative Mean Opinion Score (CMOS) listening test was carried out to assess the difference of quality. 50 test sentences were synthesised by each system. Since duration models are out of the scope of this experiment, the durations used here were extracted from the original recordings. Similarly, common f_0 curves and amplitude spectral envelopes were used among all synthesis methods in order to focus on the difference of PDD vs NM modelling. The systems trained for STRAIGHT were used to build the common features (for PML syntheses, $f_0(t)$ was then linearly interpolated in unvoiced segment to obtain a continuous $f_0(t)$ curve). Each listener taking the test assessed the 3 pairs of each system combinations for 8 random sentences among the 50x6=300 synthesized sentences [18]. Using crowdsourcing, 47 listeners took the test properly and the results are shown in Fig. 3. The "Preference test" results are deduced from the CMOS test by counting the number of assessments bigger than 1 favouring each system and those equal to zero for the no-preference choice.

Results in Fig. 3 show that the NM modelling yielded on average better scores than both STRAIGHT and PDD-based modelling. Solid brackets on the right show significant differences for p-values$<$0.001. The improvement from PDD to NM modelling shows that the noise can be successfully modelled by a simple binary mask.

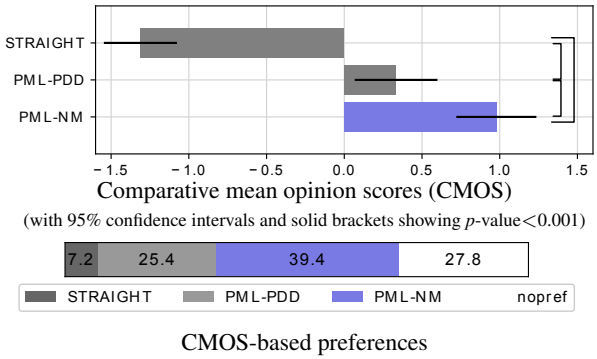

Figure 3: *Results of listening test: Baseline STRAIGHT; PML synthesis using Phase Distortion Deviation (PDD) modelling ; PML synthesis using Noise Mask (NM) modelling*

3. Comparisons in Blizzard Challenge

The Blizzard Challenge carried out listening tests for assessing various characteristics of the synthesis provided by the submit-

ted systems. Only part of the results are shown below, in order to focus on the most interesting elements. Paid listeners, volunteers and speech experts took the listening tests. The plots below show aggregated results for the three types of listeners.

In most comparisons below, four references are available: **A** Original recording; **B** Benchmark Unit selection synthesis [19]; **C** Benchmark Hidden Markov Model using GMM (HTS) (similar to [6]); **D** Benchmark DNN [4]; And our system is **J**.

3.1. Overall impression on paragraphs

The first listening tests consisted in rating synthesized paragraphs. The results for the "overall impression" rating is shown in Fig. 4. The quality provided by System J is clearly not as convincing as most other systems. Detailed results in Fig. 5,6,7 give some insight on the potential reasons behind this bad overall impression. The speech pauses seem to be as good as the average systems, thus, it cannot be the reason for a major degradation compared to the other systems. Given the results shown in Fig. **??**, and assuming comparable systems use STRAIGHT (or similar) as a vocoder, we can consider that the PML vocoder is neither the main reason of the overall degradation. However, The intonation and emotion characteristics have been clearly rated as among the worst. The voice quality being highly correlated to these two characteristics, it seems the voice quality normalisation might be among the source of the degradation. Even though the initial motivation was to simplify the voice variations in order to improve its consistency, it seems that anything which was correlated to the voice quality has been carried off, including the intonations and emotions.

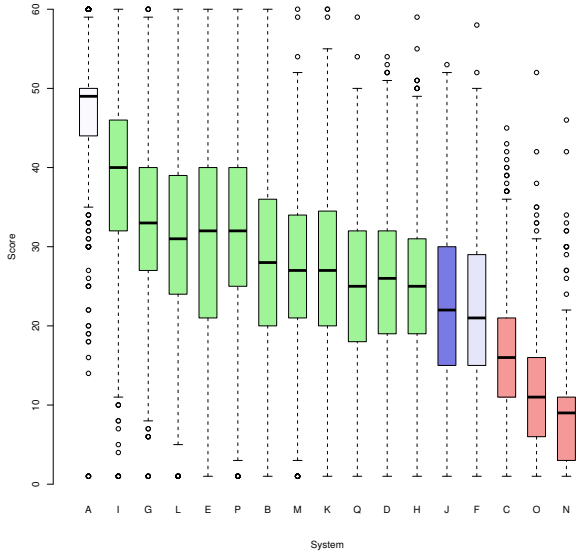

Figure 4: *MOS: Overall impression. Recording is shown with a white box, System J with a blue box, System with significantly better overall impression in green and significantly worse overall impression in red.*

3.2. Similarity and naturalness on isolated sentences

Similarity, the identity of the speaker was also assessed using a scale from 1 to 5 on isolated sentences, as well as an overall naturalness. Results are shown in Fig. 8 and 9 (with the same coloring as in previous figures). Compared to the other evaluations, the similarity to the original speaker provided by system J is comparable to other systems K, Q, D and F. In terms of naturalness, system J is not significantly different from 3 other systems. Those results seem coherent with the overall impression of the previous listening test.

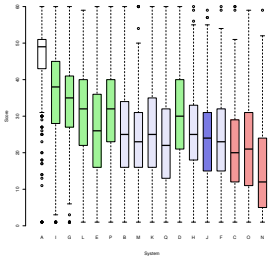

Figure 5: *MOS: Pauses (colors as in Fig.4)*

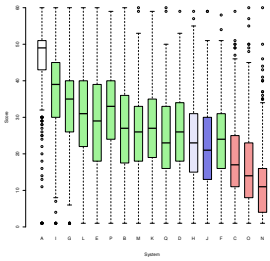

Figure 6: *MOS: Intonation*

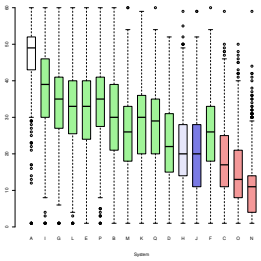

Figure 7: *MOS: Emotion*

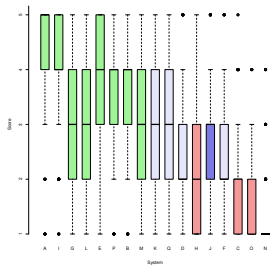

Figure 8: *MOS: Similarity*

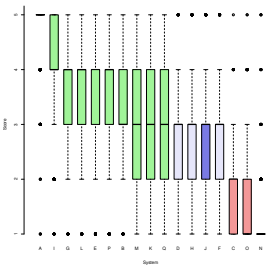

Figure 9: *MOS: Naturalness*

3.3. Intelligibility

Semantically Unpredictable Sentences (SUS) have been used to test the intelligibility of the synthetic speech. Listeners heard one utterance and typed in what they heard (only once). The word error rate (WER) is then computed by comparing the number of recognised words over the total number of words. Fig. 10 shows the results of this listening test. Despite the bad overall impression for System J reported in Fig. 4, the WER reported here is comparable to the best systems present in this Challenge. According to a Wilcoxon's signed rank test, the only system which has a significantly lower WER than System J is System D. Even though the voice quality normalisation degraded the intonation and emotion characteristics, it might actually have helped to simplify the training for the elements that are essential for intelligibility. During training, since the statistical model can rely on the voice quality features to predict most of the paralinguistic information, it has more flexibility (or "learning capacity") for modelling the phonetic content. During synthesis, by using an average voice quality feature vector, the voice quality variations is discarded, as well as the intonations,

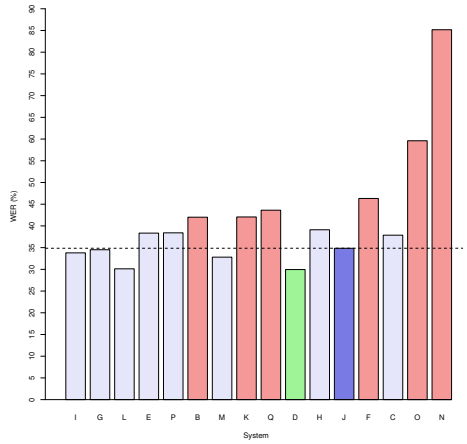

Figure 10: *Intelligibility: Measure of Word Error Rate (WER) using semantically unpredictable sentences. The dashed line is aligned on the result of System J. Only System D (the DNN benchmark, in green) has a significantly lower WER than System J. Systems with significantly higher WER than System J are shown in red.*

emotions and paralinguistic information correlated to the voice quality that might interfere with linguistic information. It remains only the linguistic information, i.e. mainly the phonetic content, that might then appear more prominent and clearer to the listener. A separate listening test for testing only the voice quality marginalisation would help to prove this point.

4. Conclusions

In this paper, we presented our submission to the Blizzard Challenge 2017 - Task EH1. We tried to deal with the high variability of the voice by three different means, i.e. data selection for training, voice quality normalisation and a new vocoder. The results of the Challenge have shown that the intelligibility provided by our system is good. Internal results have also shown that the vocoder used in our system should provide a better quality compared to similar systems. However, the normalisation of the voice quality in the acoustic model seems to have more degraded the overall impression than brought the consistency that we hoped for. Indeed, intonation and emotions (that are correlated to the voice quality) have been assessed as very low compared to most of the other systems. Since the intelligibility has been assessed as good despite the overall bad impression, it seems that conditioning the training over the voice quality might be a way to alleviate the work of the acoustic model with respect to expressivity. However, the results show that an average voice quality feature vector should not be used and should be predicted by an auxiliary model. This will be the subject of future works.

5. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 655764. The research for this paper was also partly supported by EPSRC grant EP/I031022/1 (Natural Speech Technology).

6. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[2] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proc. ICLR*, 2017.

[3] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. ISCA Speech Synthesis Workshop*, 2003, pp. 223–224, http://www.festvox.org/cmu_arctic.

[4] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. 9th Speech Synthesis Workshop (SSW9)*, 2016, pp. 218–223.

[5] G. Degottex, P. Lanchantin, and M. Gales, "A pulse model in log-domain for a uniform synthesizer," in *Proc. 9th Speech Synthesis Workshop (SSW9)*, Sunnyvale, CA, USA, September 2016, pp. 230–236. [Online]. Available: http://gillesdegottex.eu/wp-content/papercite-data/pdf/DegottexG2016pml.pdf

[6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[7] T. S. S. S. I. Group, "The Blizzard Challenge 2016 [Online]," http://www.synsig.org/index.php/Blizzard_Challenge_2016/, 2016.

[8] P. Lanchantin, M. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P. Woodland, and C. Zhang, "The development of the cambridge university alignment systems for the multi-genre broadcast challenge," in *Proc. IEEE ASRU*, 2015.

[9] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - a collaborative voice analysis repository for speech technologies," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, http://covarep.github.io/covarep/, 2014.

[10] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR [Online]," by Google on Github: https://github.com/google/REAPER, 2015.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[12] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 38, 2014. [Online]. Available: http://asmp.eurasipjournals.com/content/2014/1/38

[13] G. Degottex and N. Obin, "Phase distortion statistics as a representation of the glottal source: Application to the classification of voice qualities," in *Proc. Interspeech*, 2014, pp. 1633–1637.

[14] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *Proc. Interspeech*, 2014, pp. 1653–1657.

[15] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[16] Y. Stylianou, "Harmonic plus noise models for speech combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, TelecomParis, France, 1996.

[17] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 21, no. 10, pp. 2085–2095, 2013.

[18] T. I. R. Assembly, "ITU-R BS.1284-1: En-general methods for the subjective assessment of sound quality," ITU, Tech. Rep., 2003.

[19] K. Richmond, V. Strom, R. Clark, J. Yamagishi, and S. Fitt, "Festival multisyn voices for the 2007 blizzard challenge," in *Proc. Blizzard Challenge*, 2007.