Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech

Cassia Valentini-Botinhao[1], *Xin Wang*[2,3], *Shinji Takaki* [2], *Junichi Yamagishi*[1,2,3]

[1] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] National Institute of Informatics, Japan
[3] SOKENDAI University, Japan

cvbotinh@inf.ed.ac.uk, {wangxin,takaki,jyamagis}@nii.ac.jp

Abstract

The quality of text-to-speech (TTS) voices built from noisy speech is compromised. Enhancing the speech data before training has been shown to improve quality but voices built with clean speech are still preferred. In this paper we investigate two different approaches for speech enhancement to train TTS systems. In both approaches we train a recursive neural network (RNN) to map acoustic features extracted from noisy speech to features describing clean speech. The enhanced data is then used to train the TTS acoustic model. In one approach we use the features conventionally employed to train TTS acoustic models, i.e Mel cepstral (MCEP) coefficients, aperiodicity values and fundamental frequency (F_0). In the other approach, following conventional speech enhancement methods, we train an RNN using only the MCEP coefficients extracted from the magnitude spectrum. The enhanced MCEP features and the phase extracted from noisy speech are combined to reconstruct the waveform which is then used to extract acoustic features to train the TTS system. We show that the second approach results in larger MCEP distortion but smaller F_0 errors. Subjective evaluation shows that synthetic voices trained with data enhanced with this method were rated higher and with similar to scores to voices trained with clean speech.

Index Terms: speech enhancement, speech synthesis, RNN

1. Introduction

Statistical parametric speech synthesis (SPSS) systems [1] can produce voices of reasonable quality from small amounts of speech data. Although adaptation techniques have been shown to improve robustness to recording conditions [2] most studies on SPSS are based on carefully recorded databases. The use of less than ideal speech material is, however, of a great interest. The possibility of using found data to increase the amount of training material is quite attractive, particularly with the wealth of freely available speech data and increased processing power. In terms of applications, the creation of personalised voices [3] often relies on recordings that are not of studio quality. Quality of synthesised speech can be improved by discarding data that is considered to be too distorted but when data quantity is small or noise levels are too high discarding seems like a bad strategy. Alternatively speech enhancement can be used to pre-enhance the data.

Statistical model-based speech enhancement methods have been shown to generate higher quality speech in subjective evaluations over Wiener, spectral subtractive and subspace algorithms [4]. Recently there has been a strong interest towards methods using a deep neural network (DNN) [5, 6, 7, 8, 9] to generate enhanced acoustic parameters from acoustic parameters extracted from noisy speech. In [5] a deep feed-forward neural network was used to generate a frequency-domain binary mask using a cost function in the waveform domain. A more extensive work on speech enhancement using DNNs is presented in [6] where authors use more than 100 noise types to train a feed-forward network using noise-aware training and global variance [10]. Authors in [7] use text-derived features as an additional input of a feed-forward network that generates enhanced spectrum parameters and found that distortion is smaller when using text. In most of these studies around eleven frames (which represent a segment of at least 220 ms) are used as input to the network in order to provide the temporal evolution of the features. Alternatively authors in [8, 9] use a recursive neural network (RNN) for speech enhancement. It is difficult to compare results across studies as often authors evaluate their systems using different objective measures and no subjective evaluation. It seems however that neural network based methods outperform other statistical model-based methods and that the recursive structure is beneficial.

There have not been many studies on using speech enhancement for text-to-speech. In conventional SPSS, acoustic parameters that describe the excitation and the vocal tract are used to train an acoustic model. Authors in [11] found that excitation parameters are less prone to degradation by noise than cepstral coefficients. They found a significant preference for voices built using clean data for adaptation over voices built with noisy and speech that has been enhanced using a subspace-based speech enhancement method. In a work submitted on [12] we proposed the use of an RNN to generate enhanced vocoder parameters that are used to train an acoustic model of text-to-speech. We found that synthetic voices trained with features that have been enhanced using an RNN were rated of better quality than voices built with noisy data and data enhanced using a statistical model-based speech enhancement method. We found that using text-derived features as additional input of the network helps but not to a great extent and that fundamental frequency (F_0) errors are still quite large even after enhancement.

Most speech enhancement methods operate either on the magnitude spectrum or some sort of parametrisation of it, or on the binary mask domain that is used to generate an estimate of the clean magnitude spectrum [13]. To reconstruct the waveform, phase can be derived from the noisy signal or estimated. In such methods F_0 is not enhanced directly. We argued in [12] that enhancing the acoustic parameters that are used for TTS acoustic model training would generate better synthetic voices as it would not require waveform reconstruction. In this paper we investigate this hypothesis in more detail by comparing

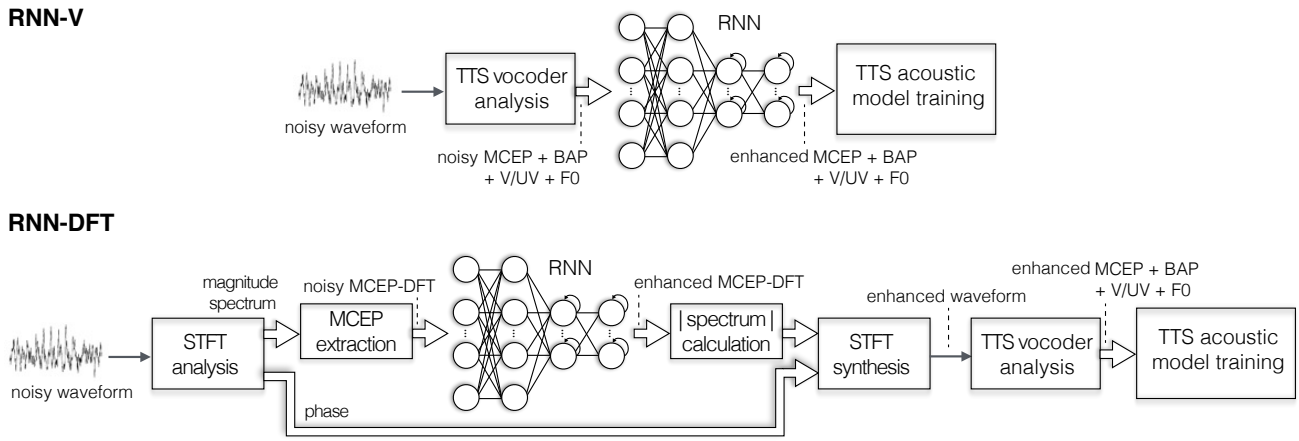

Figure 1: *Training a TTS acoustic model using an RNN-based speech enhancement method that enhances vocoded parameters directly (top) and a parametrisation of the magnitude spectrum (bottom).*

two RNN-based methods, one that operates in the TTS-style vocoder parameter domain as proposed in [12] and another that enhances a set of parameters that describe the magnitude spectrum. To simplify the comparison we do not use text-derived features in this work.

This paper is organised as follows: in Section 2 we present a brief summary of RNNs, followed by the proposed speech enhancement systems in Section 3 and the experiments in Section 4. Discussions and conclusions follow.

2. Deep recurrent neural networks

RNNs are networks that possess at least one feed-back connection, which could potentially allow them to model sequential data. Due to the vanishing gradient problem [14] they are difficult to train. Long short-term memory networks (LSTM) [15, 16] are recurrent networks composed of units with a particular structure and as such they do not suffer from the vanishing gradient and can therefore be easier to train. An LSTM unit is capable of remembering a value for an arbitrary length of time, controlling how the input affects it, as well as how that value is transmitted to the output and when to forget and remember previous values. LSTMs have been applied in a range of speech problems [17, 18], including regression problems such as text-to-speech [19, 20, 21, 22, 23] and as previously mentioned speech enhancement [8, 9]. LSTMs could be particularly interesting when training with real noisy data, i.e. recordings when speech is produced in noise and therefore changes accordingly.

3. Speech Enhancement using RNNs

Fig.1 shows the two RNN-based methods that we investigate in this paper. The diagram on the top represents the enhancement method proposed in [12]. We refer to this method as RNN-V. In this method we train an RNN with a parallel database of clean and noisy acoustic features extracted using the synthesis module of a vocoder that is typically used for SPSS. The acoustic features extracted using this vocoder are the Mel cepstral (MCEP) coefficients from a smoothed magnitude spectrum, band aperiodicity (BAP) values, the voiced/unvoiced (V/UV) decision and the F_0. These acoustic features are extracted at a frame level using overlapping F_0-adaptive windows. Once the RNN is trained it can be used to generate enhanced acoustic features from noisy ones, as displayed in the top diagram of Fig.1. These enhanced features are then used to train the TTS acoustic model.

The bottom of Fig.1 shows the alternative structure we propose in this paper, which we refer to as RNN-DFT. In this method we analyse the speech waveform using the short-term Fourier transform (STFT) to obtain the discrete Fourier transform (DFT) of each time frame. We calculate the magnitude value of this complex signal, which we refer to simply as the magnitude spectrum, as well as its phase. To decrease the dimensionality of the magnitude spectrum we extract M Mel cepstral coefficients from the N length magnitude spectrum, truncating the number of coefficients so that $M<N$. We refer to these coefficients as MCEP-DFT coefficients. We train an RNN with a parallel database of MCEP-DFT coefficients extracted from clean and noisy speech signals. Once the model is trained it can be used to generate enhanced MCEP-DFT from noisy ones. To reconstruct the speech signal these coefficients are converted to magnitude spectrum via a warped discrete cosine transform. The enhanced magnitude spectrum and the original phase obtained from the DFT extracted from the noisy waveform, as shown in the bottom of Fig.1, are combined and using the inverse discrete Fourier transform we obtain the waveform signal. This signal is once again analysed this time using the TTS-style vocoder and the extracted features are then used to train the TTS acoustic model.

4. Experiments

In this section we detail the database used to train and test these methods and the experiments conducted using vocoded and synthetic speech.

4.1. Database

We selected from the Voice Bank corpus [24] 28 speakers - 14 male and 14 female of the same accent region (England) and another 56 speakers - 28 male and 28 female - of different accent regions (Scotland and United States). There are around 400 sentences available from each speaker. All data is sampled at 48 kHz and orthographic transcription is also available.

To create the noisy database used for training we used ten different noise types: two artificially generated (speech-shaped noise and babble) and eight real noise recordings from the Demand database [25]. The speech-shaped noise was created by filtering white noise with a filter whose frequency response

Architecture	Training data	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)
NOISY	-	9.86 / 10.68	2.62 / 2.41	9.55 / 7.88	40.27 / 4.38
DNN	14 female + 14 male	5.69 / 6.10	1.96 / 1.82	4.00 / 4.25	27.09 / 10.90
RNN	14 female + 14 male	4.63 / 5.06	1.83 / 1.74	2.50 / 2.30	24.52 / 8.34
RNN	14 female	4.70 / 5.89	1.85 / 1.97	2.63 / 5.01	24.08 / 39.68
RNN	14 male	6.18 / 5.23	2.04 / 1.73	5.36 / 2.32	37.87 / 6.45
RNN	28 female + 28 male	4.59 / 5.05	1.86 / 1.72	2.46 / 2.15	24.90 / 8.43

Table 1: *Distortion measures calculated from the vocoded parameters of the female / male voice.*

matched that of the long term speech level of a male speaker. The babble noise was generated by adding speech from six speakers from the Voice Bank corpus that were not used either for either training or testing. The other eight noises were selected using the first channel of the 48 kHz versions of the noise recordings of the Demand database. The chosen noises were: a domestic noise (inside a kitchen), an office noise (in a meeting room), three public space noises (cafeteria, restaurant, subway station), two transportation noises (car and metro) and a street noise (busy traffic intersection). The signal-to-noise (SNR) values used for training were: 15 dB, 10 dB, 5 dB and 0 dB. We had therefore 40 different noisy conditions (ten noises x four SNRs), which meant that per speaker there were around ten different sentences in each condition. The noise was added to the clean waveforms using the ITU-T P.56 method [26] to calculate active speech levels using the code provided in [13]. The clean waveforms were added to noise after they had been normalised and silence segments longer than 200 ms had been trimmed off from the beginning and end of each sentence.

To create the test set we selected two other speakers from England of the same corpus, a male and a female, and five other noises from the Demand database. The chosen noises were: a domestic noise (living room), an office noise (office space), one transport (bus) and two street noises (open area cafeteria and a public square). We used four slightly higher SNR values: 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB. This created 20 different noisy conditions (five noises x four SNRs), which meant that per speaker there were around 20 different sentences in each condition. The noise was added following the same procedure described previously. The noisy speech database is permanently available at: http://dx.doi.org/10.7488/ds/1356

4.2. Acoustic features

Using STRAIGHT [27] we extracted 60 MCEP coefficients, 25 BAP components and using SPTK [28] we extracted F_0 and V/UV information with the RAPT F_0 extraction method [29].

	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)
NOISY	9.86 / 10.68	2.62 / 2.41	9.55 / 7.88	40.27 / 4.38
CLEAN*	1.84 / 1.61	1.24 / 1.10	0.58 / 0.62	17.14 / 1.84
NOISY*	9.41 / 10.13	2.75 / 2.50	10.39 / 8.49	41.17 / 4.70
OMLSA	8.19 / 8.36	3.15 / 2.77	8.73 / 8.28	34.03 / 6.31
RNN-V	4.59 / 5.05	1.86 / 1.72	2.46 / 2.15	24.90 / 8.43
RNN-DFT	4.90 / 5.22	2.44 / 2.32	2.06 / 2.44	22.59 / 3.31

Table 2: *Distortion measures calculated from the vocoded parameters of the female / male voice. CLEAN* and NOISY* refer to distortion calculated using parameters extracted from resynthesised clean and noisy signals.*

All these features were extracted using a sliding window of 5 ms shift. The resulting dimensionality of the vocoder features is 87.

Using a hamming window of 16 ms and a 4 ms shift we extracted the DFT of 1024 size. From its magnitude value we extracted 87 Mel cepstral coefficients. This number was chosen to match the number of parameters extracted using the STRAIGHT vocoder, making the comparison across methods fairer.

4.3. Speech enhancement methods

We trained different types of neural networks to map acoustic features extracted from noisy natural speech to features extracted from clean natural speech. The cost function used was the sum of square errors across all acoustic dimensions. Similar to [8] we set the learning rate to 2.0 e-5 and used the stochastic gradient descent to train the model with randomly initialised weights following a Gaussian distribution with zero mean and 0.1 variance. The momentum was set to zero. We used the CURRENNT tool [30] to train the models using a TESLA K40 GPU board.

As a conventional speech enhancement method we choose the statistical model-based method described in [31] that uses the optimally-modified log-spectral amplitude speech estimator (OMLSA) and an improved version of the minima controlled recursive averaging noise estimator as proposed in [32]. The code is available from the authors website and has been used as a comparison point for other DNN-based speech enhancement [6, 12].

4.4. Objective measures

In this section we present distortion measures calculated using the acoustic parameters extracted by the TTS vocoder. The distortion measures are the MCEP distortion in dB, the BAP distortion in dB, the F_0 distortion in Hz calculated over voiced frames and the VUV distortion calculated over the entire utterance. The measures are calculated at a frame level across all utterances of each test speaker (female/male) and averaged across frames. The distortion is always calculated using vocoded parameters extracted from clean speech as the reference. In the following sections we analyse the effect of network architecture, amount of training data, enhanced features and noisy type using these distortion measures as evaluation metric.

4.4.1. Network architecture and training data

Table 1 presents the distortion measures of the noisy test data (NOISY) and five neural network-based enhancement approaches that differ in terms of network architecture and amount of training data. All of these networks were trained using acoustic features derived from the TTS vocoder, following the RNN-V method.

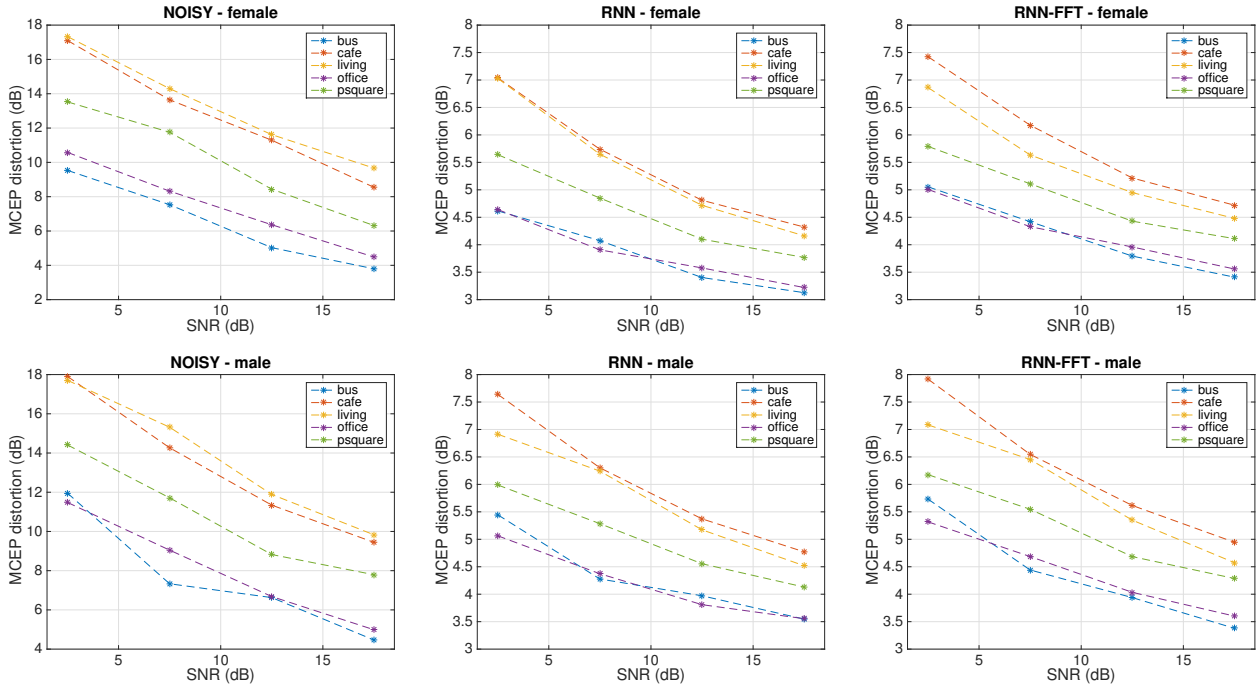

Figure 2: *Mel ceptral distortion per noise and SNR condition for female (top) and male (bottom).*

DNN refers to a deep neural network made of four feed-forward layers of 512 logistic units. RNN refers to a network with two feed-forward layers of 512 logistic units located closest to the input and two bidirectional LSTM (BLSTM) layers of 256 units closest to the output, as proposed in [12].

Most model-based speech enhancement methods train a model using data from both male and female speakers, but since the method proposed here is enhancing F_0 directly we also trained two separate models using only data from a single gender for comparison. The data used for training is noted by the column *Training data* in Table 1.

We can see from this table that RNN performance is better than DNN, particularly with respect to V/UV and MCEP distortion. The F_0 distortion of the male speaker data seem however to increase when using data from both genders for training. Results obtained using models trained with female and male data separately are only slightly better in terms of F_0 distortion but worst in terms of MCEP distortion, possibly due to the fact that the mixed gender model uses double the amount of data. Further increasing the amount of data from 28 to 56 speakers results in lower MCEP and V/UV distortion but does not improve BAP and F_0 distortions.

4.4.2. *Enhanced features and noise type*

In this section we focus on models trained with the most amount of data, i.e. 56 speakers of mixed gender. Table 2 shows the distortions of noisy speech (NOISY), resynthesised clean (CLEAN*) and noisy (NOISY*) speech, and the enhancement methods OMLSA, RNN-V and RNN-DFT. RNN-V is the same system listed in the last row of the Table 1.

The resynthesised data refers to the data that has been analysed and synthesised using the STFT settings described previously. The distortion observed in CLEAN* results are errors introduced by this process while the distortions observed in NOISY* are brought up by the resynthesis plus the presence of additive noise. As we can see in the table, BAP, VUV and F_0 distortion slightly increased when resynthesising the clean waveform (CLEAN*). Resynthesising noisy speech (NOISY*) does not seem to increase MCEP distortion (compare NOISY* and NOISY values) and only marginally increases other types of distortion. These results seem to indicate that the reconstruction process does not greatly affect the extraction of TTS acoustic features.

Regarding the enhancement methods, Table 2 shows that OMLSA results in more errors when compared to the RNN-based methods with respect to all acoustic features. RNN-V obtained lower MCEP and BAP distortion for both male and female speech, while RNN-DFT results in lower VU/V and F_0 errors. In fact only this method was able to decrease the F_0 errors of the male data.

For comparison we calculated the Mel cepstral distortion of the MCEP-DFT, i.e. the cepstral coefficients calculated from the magnitude spectrum obtained via STFT analysis. The coefficients extracted from clean speech were used as the reference. MCEP-DFT distortion of the female/male noisy speech data was found to be of $9.87/10.48$ dB. This value is similar to the one obtained for the MCEP distortion (NOISY row in Table 2). MCEP-DFT distortion decreases to $4.9359/5.3829$ dB when MCEP-DFT is enhanced using an RNN. Distortion decreased but is still larger than the MCEP distortion of RNN-V seen in Table 2.

In order to see how the performance of RNN-based methods depends on the noise type and SNR in Fig.2 we present the distortion broken down for each noise type and SNR. From these figures we can see that cafeteria (cafe) and living room (living) noises are the most challenging ones: MCEP distortion is quite high even after enhancement. This is most likely due to the fact that recordings of these noises often contained competing speaker, music and other non-stationary noises. Bus and office noises, often mostly stationary, seem to distort the signal less. The gap between the distortion brought by different noise

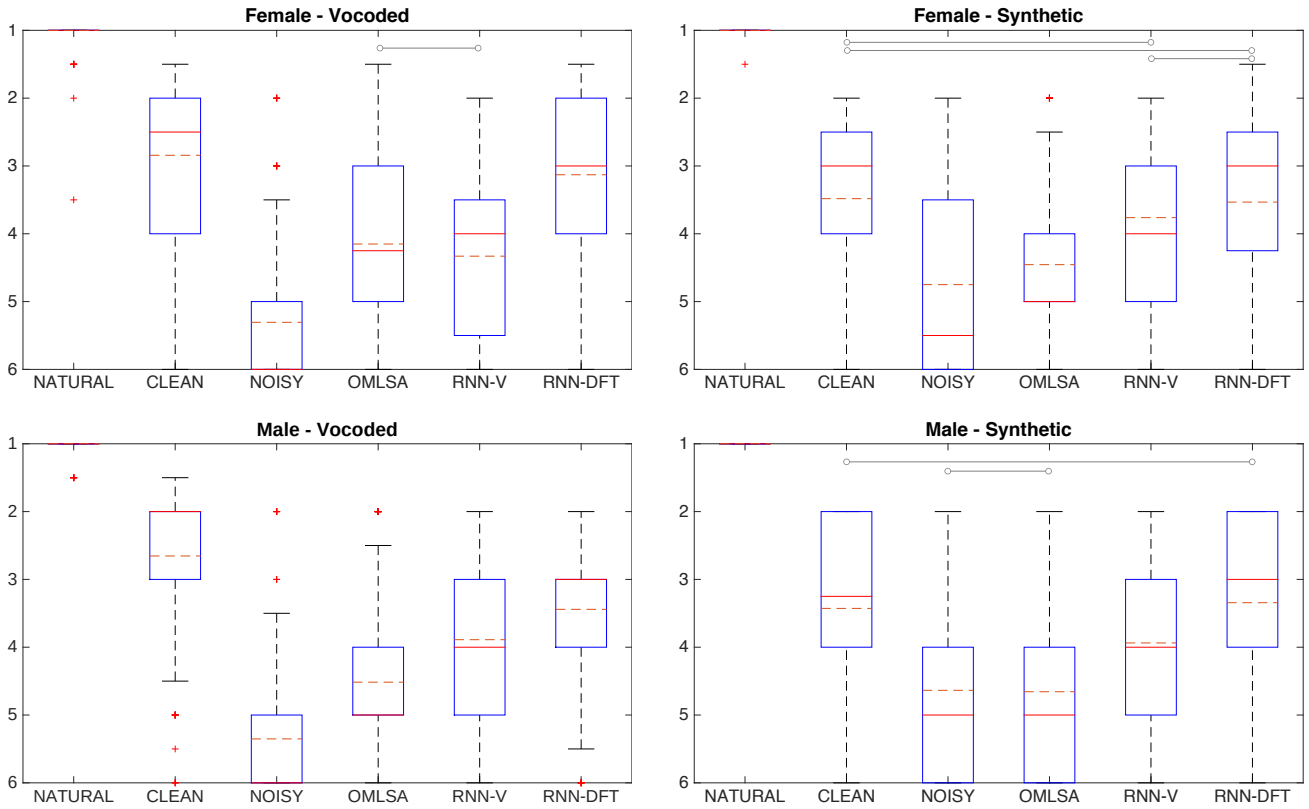

Figure 3: *Rank results of listening experiment with vocoded (left) and synthetic (right) speech of female (top) and male (bottom).*

types is made smaller with enhancement but still remains. The decrease in distortion after the enhancement seems to be higher for lower SNRs, in both RNN and RNN-DFT cases.

4.5. Text-to-speech

We built an hidden Markov model (HMM)-based synthetic voice for the female and the male test data by adapting a previously trained model of an English female speaker clean data [33, 34]. MCEP coefficients, BAP and Mel scale F_0 statics and delta and delta-deltas were used to train the model, forming five streams. To generate from these models we used the maximum likelihood parameter generation algorithm [35] considering global variance [10].

4.6. Subjective evaluation

We evaluated five different types of vocoded and synthetic speech: clean speech (CLEAN), noisy speech (NOISY) and speech enhanced by three methods (OMLSA, RNN-V, RNN-DFT). Vocoded speech is included in this evaluation to check whether the quality the synthetic voices is related to the quality of the enhanced vocoded samples. Notice that the OMLSA and RNN-DFT methods generate enhanced waveforms while RNN-DFT generate a sequence of enhanced vocoded parameters. To create vocoded speech of OMLSA and RNN-DFT we analysed and resynthesised the waveforms using the TTS vocoder. To generate vocoded speech of RNN-V we simply synthesised the enhaced parameters.

4.6.1. Listening experiment design

To evaluate the samples we created a MUSHRA-style [36] listening test. The test contained 30 screens organised in two bocks of 15 screens each: the first block with the male voice and the second with the female voice. The first half of each block is made of screens with vocoded speech samples while the second half contain screens of synthetic speech. The first screen of each block was used to train participants to do the task and familiarise them with the material. In each screen participants were asked to score the overall quality of a sample of the same sentence from each method on a scale from 0 to 100. We specifically asked listeners to rate overall quality considering both speech and background as some of the vocoded samples contained noise in the background. This is in accordance with the methodology proposed in [11]. A different sentence is used across different screens. 42 different sentences for each speech type (vocoded and synthetic) were used across six listeners. The sentences used for the vocoded speech were a subset of the ones recorded by the Voice bank corpus while the sentences used for synthesis were the Harvard sentences [37]. The training screen was constructed always with the same sentence and it was made of samples of vocoded speech. Natural clean speech was also included in the test so that participants would have a reference for good quality as well as checking if participants did go through the material and score it as 100 as instructed. We recruited 24 native English speakers to participate in this evaluation.

4.6.2. Results

Figure 3 shows the boxplot of listeners responses in terms of the rank order of systems for the female (top) and the male (bottom) voice of vocoded (left) and synthetic (right) speech. The rank order was obtained per screen and per listener according to the scores given to each voice. The solid and dashed lines show median and mean values. To test significant differences we used a Mann-Whitney U test at a p-value of 0.01 with a Homl Bonferroni correction due to the large number of pairs to compare. The pairs that were not found to be significantly different from each other are connected with straight horizontal lines that appear on the top of each boxplot.

As expected natural speech ranked highest and noise ranked lowest for all cases. RNN-DFT was rated higher among all enhancement strategies in all cases. The gap between clean and RNN-DFT enhanced speech is smaller for the synthetic speech style than for the vocoded speech. In fact for both genders the synthetic voice trained with RNN-DFT enhanced speech was not found to be significantly different than the voice built with clean speech. The increasing order of preference of the methods seem to be the same for vocoded and synthetic speech: OMLSA, followed by RNN-V and RNN-DFT. The benefit of RNN-based methods is seen in both vocoded and synthetic voices, while the OMLSA method improvements seems to decrease after TTS acoustic model training.

5. Discussion

We have found that the reconstruction process required in the RNN-DFT method does not seem to negatively impact the extraction of TTS acoustic features from noisy data. However we observed that the RNN-DFT method increases both MCEP and BAP distortion more than the RNN-V method. The assumption that phase can be reconstructed directly from the noisy speech data may have caused an increase in distortion. RNN-DFT seems however to decrease V/UV and F_0 errors when compared to RNN-V. This is somewhat unexpected as the RNN-V approach directly enhances the F_0 data. Both methods decreased MCEP distortion for all noises tested, making the gap between non-stationary and stationary noises smaller.

We argued in [12] that enhancing the acoustic parameters that are used for TTS model training should generate higher quality synthetic voices but subjective scores showed that RNN-DFT resulted in higher quality vocoded and synthetic speech for both genders. The RNN-DFT enhanced synthetic voice was in fact ranked as high as the voice built using clean data. We believe that RNN-V did not work as well because enhancing the F_0 trajectory directly is quite challenging, as F_0 extraction errors can be substantial in some frames (doubling and halving errors) while small in others.

6. Conclusion

We presented in this paper two different speech enhancement methods to improve the quality of TTS voices trained with noisy speech data. Both methods employ a recursive neural network to map noisy acoustic features to clean features. In one method we train an RNN with acoustic features that are used to train TTS models, including fundamental frequency and Mel cepstral coefficients. In the other method the RNN is trained with parameters extracted from the magnitude spectrum, as is usually done in conventional speech enhancement methods. For waveform reconstruction the phase information is directly obtained from the original noise signal while the magnitude spectrum is obtained using the output of the RNN. We have found that although Mel cepstral distortion is higher the second method was rated of a higher quality for both vocoded and synthetic speech and for the female and male data. The synthetic voices trained with data enhanced with this method were rated similar to voices trained with clean speech. In the future we would like to investigate whether similar improvements would apply to voices trained using DNNs and whether training an RNN directly with the magnitude spectrum could further improve results.

Acknowledgements This work was partially supported by EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF) and by CREST from the Japan Science and Technology Agency (uDialogue project). The full NST research data collection may be accessed at http://hdl.handle.net/10283/786.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based Speech Synthesis," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 581–584.

[3] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *J. of Acoust. Science and Tech.*, vol. 33, no. 1, pp. 1–5, 2012.

[4] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. ICASSP*, vol. 1, May 2006, pp. I–I.

[5] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. ICASSP*, April 2015, pp. 4390–4394.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[7] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc. Interspeech*, Sep. 2015, pp. 1760–1764.

[8] F. Weninger, J. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, Dec 2014, pp. 577–581.

[9] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, *Proc. Int. Conf. Latent Variable Analysis and Signal Separation*. Springer International Publishing, 2015, ch. Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR, pp. 91–99.

[10] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[11] R. Karhila, U. Remes, and M. Kurimo, "Noise in HMM-Based Speech Synthesis Adaptation: Analysis, Evaluation Methods and Experiments," *J. Sel. Topics in Sig. Proc.*, vol. 8, no. 2, pp. 285–295, April 2014.

[12] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. Interspeech*, (submitted) 2016.

[13] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 2007.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *J. Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *J. Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[17] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[18] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.

[19] S.-H. Chen, S.-H. Hwang, and Y.-R. Wang, "An RNN-based prosodic information synthesizer for mandarin text-to-speech," *Proc. ICASSP*, vol. 6, no. 3, pp. 226–239, 1998.

[20] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Proc. Interspeech*, 2014, pp. 1964–1968.

[21] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks." in *Proc. Interspeech*, 2014, pp. 2268–2272.

[22] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*. IEEE, 2015, pp. 4470–4474.

[23] S. Achanta, T. Godambe, and S. V. Gangashetty, "An investigation of recurrent neural network architectures for statistical parametric speech synthesis," in *Proc. Interspeech*, 2015.

[24] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA*, Nov 2013.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[26] *Objective measurement of active speech level ITU-T recommendation P.56*, ITU Recommendation ITU-T, Geneva, Switzerland, 1993.

[27] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, pp. 187–207, 1999.

[28] *Speech signal processing toolkit: SPTK 3.4*, Nagoya Institute of Technology, 2010.

[29] D. Talkin, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp. 495–518, 1995.

[30] F. Weninger, "Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *J. of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[31] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001.

[32] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[33] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66 –83, 2009.

[34] R. Dall, C. Veaux, J. Yamagishi, and S. King, "Analysis of speaker clustering strategies for HMM-based speech synthesis," in *Proc. Interspeech*, Portland, USA, Sep. 2012.

[35] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3. IEEE, 2000, pp. 1315–1318.

[36] *Method for the subjective assessment of intermediate quality level of coding systems*, ITU Recommendation ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, March 2003.

[37] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.