Speech Enhancement of Noisy and Reverberant Speech for Text-to-Speech

Speech Enhancement of Noisy and Reverberant Speech for Text-to-Speech

Cassia Valentini-Botinhao, Junichi Yamagishi, *Senior Member, IEEE,*

***Abstract*—Text-to-speech voices created from noisy and reverberant recordings are of lower quality. A simple way to improve this is to increase the quality of the recordings prior to text-to-speech training with speech enhancement methods such as noise suppression and dereverberation. In this paper we opted for this approach and to perform the enhancement we used a recurrent neural network. The network is trained with parallel data of clean and lower quality recordings of speech. The lower quality data was artificially created by adding recordings of environmental noise to studio quality recordings of speech and by convolving room impulse responses with these clean recordings. We trained separate networks with noise only, reverberation only and both reverberation and additive noise data. The quality of voices trained with lower quality data that has been enhanced using these networks was significantly higher in all cases. For the noise only case, the enhanced synthetic voice ranked as high as the voice trained with clean data. For the most realistic and challenging scenario, when both noise and reverberation were present, the improvements were more modest, but still significant.**

I. Introduction

Although considerable progress has been made in the text-to-speech area, particularly in statistical parametric speech synthesis (SPSS) , there is still little effort being put towards improving synthetic voices trained with lower quality recordings of speech. Most research projects and commercial systems are based on carefully recorded databases that contain very low levels of noise and reverberation. Although this is the case in many applications, there are some applications where other kinds of speech material is of a great interest. For instance, the generation of personalised voices [1] tend to rely on recordings from the end user over which we have limited control. Beyond the application driven scenario, improving quality of voices trained with lower quality data can potentially increase the amount of training material that can be used to create synthetic voices, particularly given the wealth of freely available speech data. The significant quality drop observed when the training data is noisy [2] or reverberant [3] can be compensated in a few different ways. Adaptation techniques have been shown to help but only to a certain extent [2]. Another way to improve quality is to discard data that is considered to be too distorted [2]. That becomes a bad strategy when there is not enough data, when distortion levels are too high or both. Alternatively, speech enhancement can be used to 'clean' the training data. In this paper we refer to speech enhancement as the process of removing additive noise, often called noise suppression, as well as removing the effects of the room acoustics, i.e. dereverberation.

There are a great variety of noise suppression methods in the speech enhancement literature. Methods that are based on statistical models have been shown to produce higher quality speech than methods such as spectral subtraction, Wiener filter and subspace-based ones [4]. An alternative methodology whose popularity is growing is to use neural networks to map acoustic parameters extracted from noisy speech to parameters describing the underlying clean data [5]–[9]. It is hard to compare results across studies as the choice of evaluation metrics is inconsistent, and highly application specific. Often no subjective evaluation is performed or results are shown in terms of automatic speech recognition (ASR) performance. We mention here a selection of neural network based noise suppression studies that illustrate some of the challenges and techniques used in this area. In the work described in [6] authors train a feed-forward neural network with noise-aware training and global variance estimation using more than 100 different noise conditions [10], both techniques seem to improve results. Authors in [7] investigated the use of additional input features derived from the underlying spoken text and found that spectral distortion decreases when text-based features are included. In these studies around eleven frames (at least 220 ms) of acoustic features are used as the network input. Alternatively, authors in [8], [9], use only one acoustic frame as the input to a recurrent neural network (RNN) composed of long short-term memory (LSTM) units. They reported improvements with regards to ASR performance.

Dereverberation algorithms, i.e. methods that aim to remove reverberation from recordings, are a separate category of speech enhancement. Reverberation, unlike additive noise, is a non-linear distortion, and potentially harder to remove. Dereverberation can be achieved via inverse filtering in the frequency domain, room impulse response (RIR) modelling in the time domain and non-linear mapping using artificial neural networks [11]. In the subjective evaluation of the REVERB Challenge 2016 [11] it was found that many systems significantly decreased the amount of reverberation perceived by listeners but only one system [12] improved subjective quality when compared to the unprocessed reverberated signal. This system used a RIR-modeling dereverberation method [13] based on the hypothesis that a RIR can be represented as a Gaussian stationary noise signal multiplied by an exponential decay. Non-linear mapping methods using neural networks achieved reasonably good performance in subjective quality and subjective reverberation scores [14] as well as in ASR

C. Valentini-Botinhao and J. Yamagishi are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, United Kingdom (e-mail: cvbotinh@inf.ed.ac.uk, jyamagis@inf.ed.ac.uk). J. Yamagishi is also with the National Institute of Informatics, Japan.

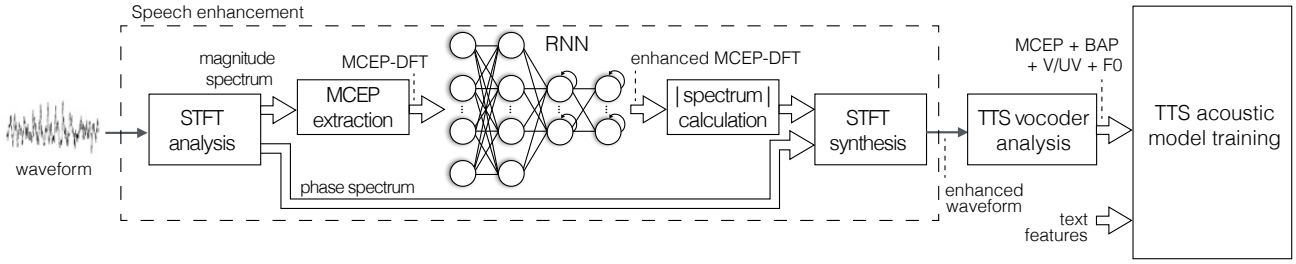

Fig. 1. Proposed framework for training TTS acoustic models using an RNN-based speech enhancement method to pre process the speech waveform prior to acoustic model training.

performance [15]. While the authors in [14] used a feed-forward network with 15 frames of log spectrum as the input, the authors in [15] used a network containing bidirectional LSTM (BLSTM) layers to map one frame of reverberant to one frame of clean log Mel spectrum, which the authors reported improved the results.

In a real scenario, however, speech recordings are contaminated with background noise and reverberation at the same time. There have been some studies that have attempted to remove both noise and reverberation from the speech signal, either by performing noise suppression and dereverberation sequentially or proposing an integrated approach [16]–[21]. More recently, authors in [21] made use of a feed-forward neural network with 11 frames input and one frame output to perform both denoising and reverberation. The authors reported their method obtained higher objective quality scores than an approach based on the ideal binary mask, but for unseen noises no baseline was used.

Much of the applied literature on speech enhancement is concerned about the effect that enhancement has on ASR performance. There have thus been comparatively few studies on the effect that noise and speech enhancement have on text-to-speech. Unlike in ASR and speech enhancement, the acoustic model used for SPSS is trained with acoustic parameters that describe not only the vocal tract but also the excitation signal. Authors in [2] found that extraction of band aperiodicities and fundamental frequency from noisy speech signals using STRAIGHT [22] generated less errors than that of cepstral coefficients. They asked listeners to choose between an HMM-based synthetic voice trained using clean speech and one trained with noisy speech (noise is added to the clean data) and they found a significant preference for the voice that used clean data. An interesting finding was that when the noise signal is continuous (babble noise), having more noisy adaptation data does not substantially increase objective quality. The best alternative seems to either use less but clean data or pre-enhance the signal before adaptation. The authors have evaluated using a non-negative matrix factorization noise suppression method but found no significant difference on listeners naturalness scores compared to voices trained with speech corrupted by babble noise. Recently we proposed the use of an RNN to directly enhance vocoder parameters extracted from noisy speech for the purpose of text-to-speech training [23]. We found that synthetic voices trained with enhanced acoustic features were rated significantly better, in terms of subjective quality, than voices trained with speech corrupted by a range of noise conditions. We also found that using text-derived features as additional input of the enhancement network was not always beneficial most likely due to alignment errors using noisy data. In that experiment we enhanced the vocoder parameters directly, including the F_0 stream. Traditionally, speech enhancement methods, however, tend to operate either on the magnitude spectrum or a parametrisation of it [24]. To reconstruct the waveform, phase is derived directly from the Fourier transform of the noisy signal or estimated from it. The F_0 information is therefore not enhanced directly. In a follow-up experiment [25] we compared using an RNN to enhance the TTS-style vocoder domain as in [23] and on the magnitude spectrum instead. To simplify the comparison no additional text-derived features were used. We found that the method operating on the magnitude spectrum obtained higher quality scores and comparable to those obtained by voices trained with clean speech.

In this article we revise parts of the work reported in [25]: the construction of a noisy version of an existing database that was purposely created for training TTS voices and the evaluation of RNN-based enhancement models on this data. Here we extend the analysis of these results showing the performance of these models with different test conditions, when noise type and level match or do not match the training data. Beyond the study presented in [25], in this article we introduce another two freely available datasets: one corrupted by reverberation and one with speech corrupted by both additive noise and reverberation. To apply similar enhancement techniques to this more challenging data we adopt an existing technique for phase estimation and present this new framework as well as objective and subjective results on this new data. In contrast with the work presented in [2] we adopt a DNN-based enhancement technique, more specifically a recursive neural network. We trained and tested this model with higher quality speech data (48kHz sampling rate) corrupted by a wider variety of noise types and levels as well as reverberation.

In Section II we present the proposed speech enhancement TTS framework. In Section III we describe the data that we created for training the speech enhancement and the TTS models. Sections IV and V present details of how we trained these models respectively, followed by Sections V-B and VII where we present objective results for the enhanced natural speech, and subjective results for both the enhanced natural speech and the synthetic speech (text-to-speech). Discussions and conclusions follow.

II. Proposed framework

Figure 1 shows a schematic of the TTS training framework adopted in this work. In this framework speech enhancement takes place prior to acoustic model training, acting like a pre-processing stage. The speech enhancement is done at a frame level and on parameters extracted from the magnitude spectrum. The magnitude spectrum is derived from the complex short-term Fourier transform (STFT). From the N length magnitude spectrum we extract M Mel cepstral coefficients, where $M<N$ via truncation. We refer to these coefficients as MCEP-DFT. An RNN is used to generate enhanced MCEP-DFT from the distorted ones. The generated coefficients are then converted to magnitude spectrum via a warped discrete cosine transform. The enhanced magnitude spectrum and the phase spectrum obtained from the input waveform are combined and using the inverse discrete Fourier transform we obtain the enhanced waveform signal. For the purpose of acoustic model training, this signal is once again analysed, this time using a TTS-style vocoder, which in this work is the STRAIGHT vocoder [22]. The extracted features are then used to train the acoustic model together with the linguistic features that are extracted from the underlying text in the signal. The linguistic features are initially aligned using acoustic features derived from the enhanced waveform.

In this framework the RNN used for speech enhancement is previously trained with a parallel database of MCEP-DFT extracted from clean and distorted speech in order to minimize the error between generated features and features extracted from clean speech data. In this work the distortion can be either additive noise, reverberation (via convolution with a room impulse response) or both.

III. Database

A. Clean speech

We selected 56 voices - 28 male and 28 female - from the VCTK corpus[1]. All voices are of native English speakers but of different accents (Scotland and United States mostly). The database provides around 400 sentences of each speaker (speech recordings and orthographic transcription is available). The recordings are sampled at 48 kHz. For testing purposes, we selected two other speakers, identified as p232 (male) and p257 (female), both from England. The sentence level waveforms were trimmed at the beginning and the end in order to remove silence segments longer than 200 ms.

B. Additive noise

The noise recordings used for either training or testing are derived from the multichannel Demand database [26], more specifically from the first channel of the 48 kHz versions of the noise recordings. We added noise to the clean speech sentences using the ITU-T P.56 standard [27] for calculating the active speech level, using the code provided in [24]. The noise segments were chosen randomly from the longer noise signals.

[1]available here: http://dx.doi.org/10.7488/ds/1994

TABLE I
THE ROOM IMPULSE RESPONSES (RIR) THAT WERE USED TO CREATE THE REVERBERANT SPEECH DATASET LISTED ACCORDING TO INCREASING ORDER OF $\overline{\text{T60}}$. THE $\overline{\text{T60}}$ VALUE IS THE AVERAGE BETWEEN THE T60 MEASURED IN MICROPHONE POSITION 1 (T60_1) AND 2 (T60_2). THE CONDITIONS SELECTED FOR THE TEST SET ARE INDICATED WITH X. THE REMAINING CONDITIONS WERE USED FOR TRAINING.

RIR Database	Room setting	T60_1 (ms.)	T60_2 (ms.)	$\overline{\text{T60}}$ (ms.)	Test
Artificial	Small room	97	107	102	
MIRD [28]	Small room	137	168	152	X
Artificial	Medium room	230	247	239	
MIRD [28]	Medium room	292	326	309	
ACE [29]	Office 1	370	345	357	
ACE [29]	Meeting room 2	382	362	372	X
ACE [29]	Office 2	402	429	416	
ACE [29]	Meeting room 1	462	460	461	
MARDY [30]	Reflective walls	477	560	518	
MIRD [28]	Large room	540	603	572	X
ACE [29]	Lecture room 1	525	676	600	
ACE [29]	Building lobby	727	821	774	

For training, we used ten different noise types: eight noise recordings from the Demand database and two artificially generated noises (speech-shaped noise and babble noise). We created the speech-shaped noise from white noise by filtering it with a filter whose frequency response matched that of the long term average spectrum of a male speaker. We created the babble noise by adding recordings from six speakers of the VCTK corpus (not used either for either training or testing). The eight noise recordings chosen for training were: kitchen, meeting room, office cafeteria, restaurant, subway, car, metro and traffic. Each of these noises were added to the speech signal at four different signal-to-noise (SNR) values: 15 dB, 10 dB, 5 dB and 0 dB. In total there were 40 different noisy conditions (ten noises x four SNRs). This meant that per speaker we had around ten different sentences in each condition.

For testing, we used five other noise recordings from the Demand database: living room, office, bus, street cafe and a public square. We used four slightly higher SNR values than the ones used for training: 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB. This resulted in 20 different noisy conditions (five noises x four SNRs), which meant that per speaker we had around 20 different sentences in each condition.

To create these mismatched test set conditions, we used slightly higher SNR values (or conversely slightly lower for training) because we wanted to train using the worst case scenario but evaluate in practical SNR values. We do not expect these models to enhance material recorded at negative SNRs (note that our application is creating voices for TTS systems) and we believe it is more beneficial to train using more challenging data (lower SNRs), in case recording quality is worse than expected.

The noisy speech database as well as the silence trimmed clean speech set created for this work is permanently available at: http://dx.doi.org/10.7488/ds/2117

C. Reverberation

In order to recreate the reverberation effect we convolved this material with a variety of single channel room impulse

responses (RIR) derived from publicly available databases. In order to cover a variety of reverberation levels and room settings we used RIR from three different databases: the MARDY database [30], the MIRD database [28] and the ACE Challenge database [29]. Each of these databases provide RIRs recorded with microphones placed in at least two different locations given a certain room configuration. The documentation of the MIRD and MYRC database describe that microphone position is either one or two meters away from the source (loudspeaker), while the ACE database documentation does not report what the two different positions reflect. All RIR signals were downsampled to 48 kHz to match the sampling frequency of the speech signal.

To choose the different conditions we computed the T60 values for each room and microphone position using the tool provided in [28]. T60 is a measure for the reverberation time. It quantifies how long it takes for the impulse sound level to decrease 60 dB. Table I shows the T60 for each microphone position and the average per room. The table also shows the conditions we choose for training and testing.

To compose the training set we selected seven different rooms: the medium size room of MIRD, the Office 1, Office 2, Meeting Room 1, Lecture Room 1 and Building Lobby rooms of ACE, and the reflective room settings of MARDY. To account for environments with less reverberation we generated two artificial rooms using the tool from [31]. The room settings used to create the artificial RIRs were: a three by three meters room with a T60 of 130 ms (small room) and a five by four meters room with a T60 of 250 ms (medium size room). Both rooms were of six meters high.

To compose the test set we selected the following three room environments: the small and large rooms from the MIRD database and the Meeting Room 2 from the ACE database. In total there were 18 conditions used for training (two microphone positions x seven real rooms + two artificial rooms) and six for testing (two positions x three real rooms). Note that the test conditions were chosen to reflect small (152 ms), medium (372 ms) and large (572 ms) levels of reverberation. The reverberant speech data created for this work is permanently available at: http://dx.doi.org/10.7488/ds/1425

D. Additive noise and reverberation

To create the noisy reverberant speech training set we selected five out of the ten rooms that were used to create the reverberant database and all the ten noise types and four SNRs that were used to create the noisy database. The rooms selected were: small room from the artificial database, medium room from the MIRD database, and office 1, meeting room 1, lecture room 1 from the ACE database. This meant that there were in total 200 (five rooms x ten noises x four SNRs) conditions covered, which in practice means that there were around two sentences per speaker and condition. To create the test set the same two speakers used previously were selected and the same test set noise types, SNRs and rooms used. This resulted in 60 (three rooms x five noises x four SNRs) conditions, which means that around six sentences per speaker matched one of this conditions.

Following a similar procedure as in [21], the noisy and reverberant speech was created as follows:

$$y = x * h_1 + \alpha(n * h_2) \tag{1}$$

where x and n refer to the clean speech and the noise waveform respectively, h_1 and h_2 refer to the room impulse response recorded using microphone position 1 and 2, $*$ stands for the convolution operator and α is calculated according to the desired SNR of the condition. Microphone position 1 is the one closer to the loudspeaker as we would expect that the speech source is closer to the microphone than the background noise.

The noisy and reverberant speech database is permanently available at: http://dx.doi.org/10.7488/ds/2139

IV. Speech enhancement

A. Baselines

As the baseline noise suppression method we adopted the method described in [32]. This method uses the optimally-modified log-spectral amplitude speech estimator (OMLSA) and the minima controlled recursive averaging noise estimator as proposed in [33]. This method can be classified as a statistical model-based method. It might be consider a weaker baseline but it has been used as a comparison point for other neural network based speech enhancement studies [6] and it is freely available from the authors website.

As a baseline dereverberation method we used the freely available dereverberation algorithm of the open source tool Postfish [34]. The Postfish dereverberation tool compresses decaying segments of the speech signal to attenuate the effect of reverberation while keeping the attack segments intact. Despite its simplicity it has been successfully applied in the past to pre-process data for the purpose of TTS training [3]. In this work we used a smoothing setting of 40 ms and a release of 400 ms as used in [3].

B. RNN-based speech enhancement

To train the speech enhancement neural network we extracted MCEP-DFT features using a hamming window of 16 ms and a 4 ms shift. From each windowed speech frame we extracted a DFT of 1024 size and from its magnitude value we extracted 87 Mel cepstral coefficients, which we refer here as MCEP-DFT features. We chose this value as it matches the overall number of parameters extracted using the STRAIGHT vocoder, a comparison point for our feature experiments with additive noise. The input of the network is the frame level MCEP-DFT extracted from the lower quality speech signal and the target output is the MCEP-DFT extracted from the underlying clean speech signal of that particular frame. A different network was trained with only the noisy data (RNN-N), the reverberant data (RNN-R) and the noisy and reverberant data (RNN-NR).

The network architecture in this work is fixed to two feed-forward layers of 512 logistic units (located closest to the input) and two layers containing 256 bidirectional LSTM (BLSTM) units (closest to the output). To train the networks we used as cost function the sum of square errors taken

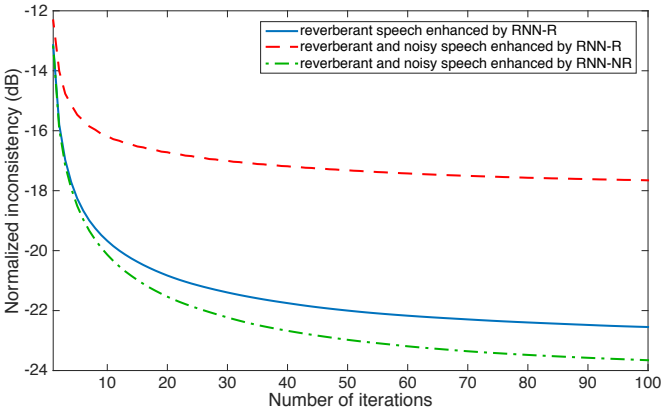

Fig. 2. Normalized inconsistency measure averaged across all test sentences versus the number of iterations taken by the modified update signal reconstruction method described in [36].

across all acoustic dimensions. We trained the models using the stochastic gradient descent method and following the hyperparameter choices in [8]: learning rate of 2.0 e-5 and momentum of zero. We used the CURRENNT tool [35] to train the models with a TESLA K40 GPU board.

C. Phase estimation for RNN-based speech enhancement

We observed that reconstructing speech using the phase spectrum from reverberant speech generated considerable audible artifacts. The reverberation effect affects the phase spectrum more than additive noise because of the convolution operation. This means that the inconsistency between the enhanced magnitude spectrum and the reverberant phase spectrum can be large. To alleviate this effect, rather than using the phase spectrum from the reverberant speech we estimated a more appropriate phase spectrum using the method referred in [36] as the modified update, based on the Griffin-Lim algortihm [37]. In this method the phase spectrum is updated iteratively with the phase of the Fourier transform of the inverse transform of the complex signal.

Figure 2 shows the evolution of the normalized inconsistency measure (as defined in [36]) when estimating the clean phase spectrum from enhanced reverberant speech (blue curve) and from enhanced noisy reverberant speech (red and green curves). The red curve refers to the case when the magnitude spectrum has been enhanced by the neural network trained only with reverberant speech (RNN-R) while the green curve is the case when it has been enhanced by the RNN trained only with noisy and reverberant data (RNN-NR). We can see that, for these three cases, inconsistency drops steadily. The iterative procedure was not used for data enhanced using the RNN-N network as it was not as effective. We believe that happened because the inconsistency between enhanced magnitude and noisy reverberant phase spectrum is smaller when the enhancement does not remove the reverberation effect from the magnitude spectrum.

V. Text-to-speech

A. Acoustic feature extraction

To train the text-to-speech acoustic model we extracted vocoder-type features using STRAIGHT [22], [38]: 60 Mel cepstral coefficients (MCEP) coefficients, 25 band aperiodicities (BAP) components. We extracted fundamental frequency (F_0) and voiced/unvoiced (V/UV) information using SPTK [39] F_0 extraction routine with the RAPT's extraction method option [40].

All acoustic features were extracted using a sliding window of 5 ms shift. MCEP features were calculated from STRAIGHT's smooth spectrum that was extracted using two pitch-adaptive window functions. The shape of these two windows are designed in order to smooth peaks and valleys respectively [38]. The BAP features were extracted using one of these windows, following the procedure described in [38]. F_0 feature was extracted using a Hanning window [40].

B. Acoustic model training

It is possible to improve the quality of HMM-based TTS voices trained with noisy speech using adaptation techniques [2], [41], [42]. For this purpose the clean speech of other speakers is used to train a model and this model is adapted to the target speaker's noisy speech data.

There have been various studies on the noise robustness of the different HMM-based adaptation techniques for TTS, e.g. the CSMAPLR (constrained structural maximum a posteriori linear regression) method [2], [41], the EMLLR (eigenspace-based maximum likelihood linear regression) method [2] and the CAT (cluster adaptive training) method [42]. CAT and EMLLR are similar techniques as they project the adaptation data into a linear space trained on clean data. While the authors in [42] found that CAT is better than CSMAPLR, in terms of perceived quality, authors in [2] found that CSMAPLR is superior to EMLLR.

We trained two types of HMM-based models. A speaker dependent one, trained only with data of the target speaker (around 400 sentences), and a speaker adapted one. As there is no publicly available implementation of CAT, we decided to use CSMAPLR for our experiments. To train the adapted model we adopted a two stage adaptation, a CSMAPLR stage followed by a maximum a posteriori (MAP) update (CSMAPLR-MAP) [41], [43]. The model used for adaptation was trained with data of an English female speaker [41], [44], chosen as it produced higher likelihood when adapting with the clean test data. For both adapted and speaker dependent models, static frame level MCEP, BAP and Mel scale F_0 as well as delta and delta-deltas were used as observation vectors. We used the maximum likelihood parameter generation algorithm [45] considering global variance [10] to generate acoustic trajectories using these models from text.

In order to check if adaptation is beneficial for our own data (higher sampling frequency, different types of additive noise and speakers) we performed a preference test (8 native English speakers evaluated 48 pairs of sentences each). We found that listeners significantly preferred the speaker adapted to the speaker dependent voice, for the case that the voice is trained with noisy data and speech enhanced using the RNN-N model. The only exception was the enhanced adapted male voice that was preferred over the speaker dependent alternative but not significantly. These results were obtained using 95%

TABLE II
NOISY DATA EVALUATION: PERCENTAGE IMPROVEMENT RELATIVE TO UNPROCESSED NOISY SPEECH IN DIFFERENT TEST CONDITIONS FOR THE FEMALE / MALE VOICE.

RNN-N (V)					
Test conditions	MCEP-DFT (%)	MCEP (%)	BAP (%)	V/UV (%)	F_0 (%)
matched	-	60.67 / 59.42	35.99 / 35.07	78.07 / 74.97	51.81 / 17.60
partially mismatched	-	60.28 / 58.77	34.46 / 33.09	79.34 / 77.57	46.45 / -11.00
mismatched	-	53.74 / 52.81	29.28 / 28.63	74.51 / 72.59	38.88 / -87.90
RNN-N					
Test conditions	MCEP-DFT (%)	MCEP (%)	BAP (%)	V/UV (%)	F_0 (%)
matched	58.79 / 57.22	58.45 / 58.23	14.01 / 7.29	79.33 / 73.80	51.99 / 47.72
partially mismatched	57.75 / 55.99	57.94 / 57.56	12.16 / 6.32	80.40 / 76.31	49.34 / 30.69
mismatched	49.95 / 48.66	50.51 / 51.12	7.22 / 3.73	78.74 / 69.04	44.41 / 24.43

TABLE III
NOISY DATA EVALUATION: OBJECTIVE MEASURES CALCULATED FOR THE FEMALE / MALE VOICE.

	MCEP-DFT (dB)	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)	STOI	PESQ
NOISY	9.87 / 10.48	9.86 / 10.68	2.62 / 2.41	9.55 / 7.88	40.27 / 4.38	0.95 / 0.95	2.79 / 2.88
CLEAN*	-	1.84 / 1.61	1.24 / 1.10	0.58 / 0.62	17.14 / 1.84	0.98 / 0.98	3.63 / 3.82
NOISY*	-	9.41 / 10.13	2.75 / 2.50	10.39 / 8.49	41.17 / 4.70	0.94 / 0.95	2.78 / 2.87
OMLSA	-	8.19 / 8.36	3.15 / 2.77	8.73 / 8.28	34.03 / 6.31	0.93 / 0.94	2.87 / 2.97
RNN-N (V)	-	4.59 / 5.05	1.86 / 1.72	2.46 / 2.15	24.90 / 8.43	0.89 / 0.91	1.82 / 2.26
RNN-N	4.94 / 5.38	4.90 / 5.22	2.44 / 2.32	2.06 / 2.44	22.59 / 3.31	0.95 / 0.95	3.11 / 3.25

TABLE IV
REVERBERANT DATA EVALUATION: OBJECTIVE MEASURES CALCULATED FOR THE FEMALE / MALE VOICE.

	MCEP-DFT (dB)	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)	STOI	PESQ
REVERB	12.65 / 12.04	11.19 / 11.03	3.54 / 3.24	18.81 / 15.62	39.53 / 16.45	0.65 / 0.65	2.17 / 2.33
Postfish	-	10.17 / 9.83	3.40 / 3.14	14.96 / 13.17	38.34 / 15.42	0.75 / 0.74	2.21 / 2.41
RNN-R-r	5.97 / 5.97	6.51 / 6.42	2.97 / 3.04	7.07 / 7.94	44.31 / 13.01	0.83 / 0.84	2.37 / 2.50
RNN-R	5.97 / 5.97	5.89 / 5.73	3.03 / 3.07	6.19 / 6.47	47.28 / 9.65	0.86 / 0.87	2.52 / 2.66

TABLE V
NOISY AND REVERBERANT DATA EVALUATION: OBJECTIVE MEASURES CALCULATED FOR THE FEMALE / MALE VOICE.

	MCEP-DFT (dB)	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)	STOI	PESQ
NOISYREVERB	18.94 / 18.10	17.78 / 18.13	3.76 / 3.66	23.33 / 23.59	66.56 / 25.19	0.62 / 0.63	1.74 / 1.88
OMLSA	-	14.50 / 14.52	4.08 / 3.77	24.89 / 21.54	54.15 / 41.18	0.61 / 0.61	1.89 / 2.03
Postfish	-	17.09 / 17.32	3.76 / 3.65	23.08 / 23.22	66.74 / 26.13	0.64 / 0.65	1.83 / 1.99
RNN-N	11.11 / 10.61	11.77 / 10.92	3.37 / 3.38	20.67 / 16.98	58.47 / 25.52	0.63 / 0.65	1.87 / 2.17
RNN-R	12.88 / 13.03	14.58 / 15.27	3.55 / 3.48	20.53 / 21.49	85.49 / 21.34	0.73 / 0.75	1.58 / 1.72
RNN-NR	9.29 / 8.74	9.53 / 9.32	3.19 / 3.31	8.75 / 10.66	44.60 / 18.66	0.83 / 0.84	2.18 / 2.28

confidence intervals retrieved with a two-tailed binomial test. As listeners preferred the adapted voices, adaptation was always used for the remainder of this work.

VI. OBJECTIVE EVALUATION OF SPEECH ENHANCEMENT TECHNIQUES

In this section we present objective results for the enhanced natural speech data.

We do not report objective metrics for the synthetic speech as there is no reliable and agreed upon metric to evaluate TTS voices. The lack of a natural speech reference with the same duration structure and the great variety of distortions that are potentially introduced by TTS systems are a few of the challenges that researchers face when trying to create such metrics. The work presented in [46] shows that it is possible to learn a reliable metric when an extensive amount of subjective ratings of a specific TTS engine are available, which it is not a realistic scenario.

A. Objective measures

We present two types of distortion measures: vocoder parameter distortions and intelligibility/quality measures. All measures are calculated using the clean speech as the reference.

The first type of measures serves as an indication of the errors that the vocoder makes when extracting parameters from natural speech that is not 'clean'. The distortion measures we calculated are the MCEP distortion (in dB), the BAP distortion (in dB), the F_0 distortion (in Hz) calculated over voiced frames and the V/UV distortion (in %) calculated over the entire utterance. For comparison we also calculated the distortion of the MCEP-DFT parameters. Each measure is calculated at a frame level across all utterances of each test speaker (female/male) and then averaged across frames. For all these measures a lower score indicates better results.

Besides the feature level distortion measures we also present two measures that are commonly reported in speech enhancement studies: the STOI and the PESQ. The Short-Time Objective Intelligibility (STOI) [47] is an objective intelligibility measure claimed to correlate especially well with listeners scores for conditions where noisy speech is processed by a noise suppression algorithm [47]. The measure is the linear correlation coefficient between a time-frequency (T-F) representation of clean and a normalized T-F representation of noisy

speech averaged over time frames. The T-F representation is obtained by: one-third octave band analysis of windowed time frames of 25.6 ms with $50\,\%$ overlap. STOI values are always between zero and one. The Perceptual Evaluation of Speech Quality (PESQ) [48] was designed as a measure for predicting the quality of speech signals transmitted over telephone lines and it became an ITU standard for evaluating telecommunication networks in 2000. The measure is the difference between the loudness spectra of clean and noisy speech averaged over time and frequency and then mapped into a zero to five scale to match the mean opinion score scale. For both STOI and PESQ, higher scores represent better results.

B. Additive noise results

In this subsection we compare the unmodified noisy speech against the OMLSA baseline and two types of RNN-based enhancement methods. The first type, the RNN-N method, uses a model trained with magnitude spectrum derived features (MCEP-DFT). The model used for the second method, the RNN-N (V), is trained with the features that are used to train the TTS acoustic model (MCEP, BAP and F_0). The output of this method can therefore be transferred directly to the TTS acoustic module, while the output of the first method needs to be synthesized to a waveform and then further analysed, as shown in Figure 1. RNN-N (V) offers a clear advantage over RNN-N, however the task of learning how to enhance multiple acoustic streams might be more challenging.

1) Matched conditions: Table II shows values of relative improvement (in percent) with regards to unprocessed noisy speech in different test conditions. The test conditions were: matched (same noise and SNRs as used in training), partially mismatched (same noises but different SNRs) and mismatched (different noises and different SNRs). The final condition is the one we described as our normal test set in the Section III). The SNR values used for the partially mismatched conditions are the same used for the mismatched one.

We can observe, as expected, that the best performance occurs when both noise type and SNR values used for testing are the same as the ones used for training (around 60% MCEP improvement for both RNN-N (V) and RNN-N). There is a decrease in performance when the test and training conditions do not match, as can be observed by the drop in MCEP and particularly F_0 improvements for the male voice. The mismatched SNR and particularly mismatched noise type had the effect of increasing F_0 extraction errors for the RNN-N (V) model. For the model trained using MCEP-DFT features (RNN-N) the mismatched condition did not affect vocoder extraction performance as much.

2) Mismatched condition: Table III shows the distortion values calculated from noisy speech (NOISY), resynthesised clean speech (CLEAN*), resynthesised noisy speech (NOISY*), and speech enhanced by the enhancement methods OMLSA, RNN-N (V) and RNN-N. The resynthesised entries refers to speech data that has been analysed and reconstructed using the STFT settings used to extract MCEP-DFT. The motivation to include these baselines is to observe the errors introduced by this process.

We can see that vocoder distortion is relatively small (less than 2 dB of MCEP distortion) when the STFT analysis and synthesis process is applied to the clean waveform (CLEAN*). Distortion values are in general the same for noisy and resynthesized noisy speech (compare NOISY* and NOISY values). These seem to indicate that the MCEP-DFT extraction process, that is necessary for the speech enhancement framework we propose here, does not greatly affect feature extraction for acoustic model training.

In terms of enhancement methods, the results in Table III show that the RNN-based methods decrease errors substantially more when compared to OMLSA. RNN-N (V) obtained lower MCEP and BAP distortion but higher VU/V and F_0 errors. In fact only RNN-N was the only enhancement method able to decrease the F_0 errors of the male data, even though F_0 is not directly enhanced in this method. MCEP-DFT distortion decreases from $9.87/10.48$ dB to $4.94/5.38$ dB (female/male) when MCEP-DFT is enhanced using an RNN (RNN-N). A slightly bigger drop is observed when enhancing MCEP features directly in terms of MCEP distortion: $9.86/10.68$ dB (NOISY) to $4.59/5.05$ dB (RNN-N (V)). STOI and PESQ scores seem to indicate that the best enhancement method is the RNN-N followed by OMLSA and RNN-N (V). Note that RNN-N (V) obtained STOI and PESQ scores that are lower than the unmodified NOISY data even though the distortion based measures are smaller. We believe this could be a limitation of PESQ and STOI as these metrics have not been proposed with these kind of data in mind (speech enhanced with a model that modifies acoustic features like pitch and aperiodicity).

C. Reverberation results

Table IV shows the distortion measures calculated for the female / male reverberant speech (REVERB) and for speech that has been enhanced using different enhancement methods (Postfish, RNN-R-r and RNN-R). Both RNN-R and RNN-R-r methods are based on RNN trained using the reverberant database. The RNN-R-r and RNN-R methods differ on the fact that for the RNN-R case we applied the phase estimation iterative method described in Section IV-C, while for RNN-R-r phase is directly derived from the reverberant data.

Although the noisy and the reverberant test sets are not necessarily directly comparable, we can observe that reverberation seems to affect V/UV extraction and the F_0 extraction to a greater extend, while MCEP distortion is only slightly higher.

Enhancement using the reference dereverberation method Postfish is relatively poor, with MCEP and V/UV distortion above still 10 dB and 13%. The RNN-based enhancement methods RNN-R-r and RNN-R, obtained the lowest distortions across all acoustic features except F_0 error for the female speaker. Estimating the phase spectrum rather then directly using the reverberated spectrum seems to improve MCEP and V/UV results as well as F_0 errors for the male speaker. The MCEP-DFT relative improvement observed for RNN-R is comparable to the improvement obtained by RNN-N with the noisy database, i.e. almost halving the error. Both STOI

and PESQ seem to indicate that the best enhancement method is the RNN-R followed by RNN-R-r and Postfish.

D. Additive noise and reverberation results

Table V presents the distortion measures for noisy reverberant speech (NOISYREVERB) and noisy reverberant speech that has been enhanced: OMLSA, POSTFISH, RNN-N, RNN-R and RNN-NR. RNN-NR method is based on RNN trained using the noisy reverberant database.

We can see that overall distortion measures increase highly when both reverberation and additive noise are present, affecting both MCEP, voicing decision errors and F_0 extraction greatly.

Postfish and OMLSA enhancement had little positive impact on these measures: voicing decision errors still remain high above $21\,\%$ and MCEP distortion does not decrease more than $14\,$dB. We believe these results are expected because these methods were specially designed to deal with data that is corrupted by only one specific type of distortion (either additive or convolutional). Postfish is based on the idea that the main distortion effect is temporal smearing and OMLSA operates with a noise estimator that expects noise to be additive only. The RNN-N and RNN-R methods however are data-driven and do not assume any particular distortion model. The RNN trained with additive noise was able to reduce this number to about $11\,$dB. while RNN-R performed relatively worse with MCEP distortions above $14\,$dB. Enhancement using the network trained with matched conditions (RNN-NR) caused fewer extraction errors overall. MCEP-DFT and MCEP distortion is around $9\,$dB (again we see that RNN-based enhancement halved MCEP-DFT errors). More interestingly we can see that voiced/unvoiced errors heavily decreased to less than $8.75\,\%$ and $10.66\,\%$ for the female and male voice and that F_0 extraction errors also decreased greatly for both voices. STOI and PESQ scores indicate that the best enhancement method is the RNN-NR. The noise suppression method RNN-N obtained lower STOI scores (intelligibility indicator) than the dereverberation method RNN-R but higher PESQ scores (quality indicator). A similar pattern is seen when comparing the noise suppression method OMLSA to the dereverberation method Postfish.

VII. Subjective evaluation

In this section we present results of three separate listening tests performed to evaluate with noisy, reverberant and noisy reverberant data.

A. Listening test design

We performed three MUSHRA-style [49] listening tests. We asked listeners to rate overall quality considering both speech and the background, as some of samples contained noise. Listeners were presented a sequence of screens. In each screen they could play the audio produced from different systems for the same sentence material. Listeners were asked to rate these systems against each other on a scale from 0-100 using the sliders on the screen. The test had 30 screens, the first half contained the variants of the male voice, and the second, the female speaker. In each half the first screen was used to train the participant with the task and the remaining 14 contained samples of vocoded (first seven screens) and synthetic speech (last seven screens). Clean natural speech was included in each screen so that participants would have a reference for good quality and to check if participants did go through the material and score it as 100 as instructed. Across six participants, for the same speech type (vocoded versus synthetic), 42 different sentences were used. For the vocoded speech we used a a subset of the sentences available from the test set while the sentences used for text-to-speech synthesis were the Harvard sentences [50]. We recruited 24 native English speakers for each test.

B. Subjective measures

We present results in boxplots of the rank order and in tables in terms of raw scores. The rank results are obtained per screen and per listener. The boxplots solid and dashed lines refer to the median and mean values of each distribution. To check if differences in rank order were significantly different we used a Mann-Whitney U test, at a p-value of 0.01, and with a Holm Bonferroni correction due to the large number of pairs to compare. In each boxplot pairs we draw horizontal straight lines connecting pairs of conditions that were not found to be significantly different from each other. We present tables with the raw scores given to each voice to give an illustration of the absolute score given to each system. The raw scores are calculated per screen and per listener. We believe that the rank values are a better indication of the ordering of the systems because the rank values are less corrupted by participants preferences on how to utilize the 0-100 range.

C. Additive noise results

We compared five conditions of vocoded and synthetic speech: clean speech (CLEAN), noisy speech (NOISY) and speech enhanced by three methods: OMLSA, RNN-N (V), RNN-N. As mentioned in the previous section, the RNN-N (V) method uses a model trained with TTS-vocoder parameters that include magnitude spectrum, aperiodicity measures and F_0, while the RNN-N is trained with magnitude spectrum only features. The advantage of using the RNN-N (V) method is that the enhanced features can be directly used by the TTS module, without the need for waveform reconstruction and further acoustic feature analysis.

To create vocoded speech of OMLSA and RNN-N conditions, we analysed and resynthesised with STRAIGHT the speech waveform enhanced by each of these methods. To create vocoded speech for the RNN-N (V) condition, we synthesised the enhanced parameters using STRAIGHT.

The boxplot of rank order for the female (top) and the male (bottom) voice of vocoded (left) and synthetic (right) speech is presented in Figure 3. Table VI presents the raw scores given to each voice (from the scale of 0-100) averaged across listeners and sentences.

As expected, natural clean speech (NATURAL) ranked highest and NOISY lowest for all voices. The order of preference (in terms of median rank values) for speech enhancement

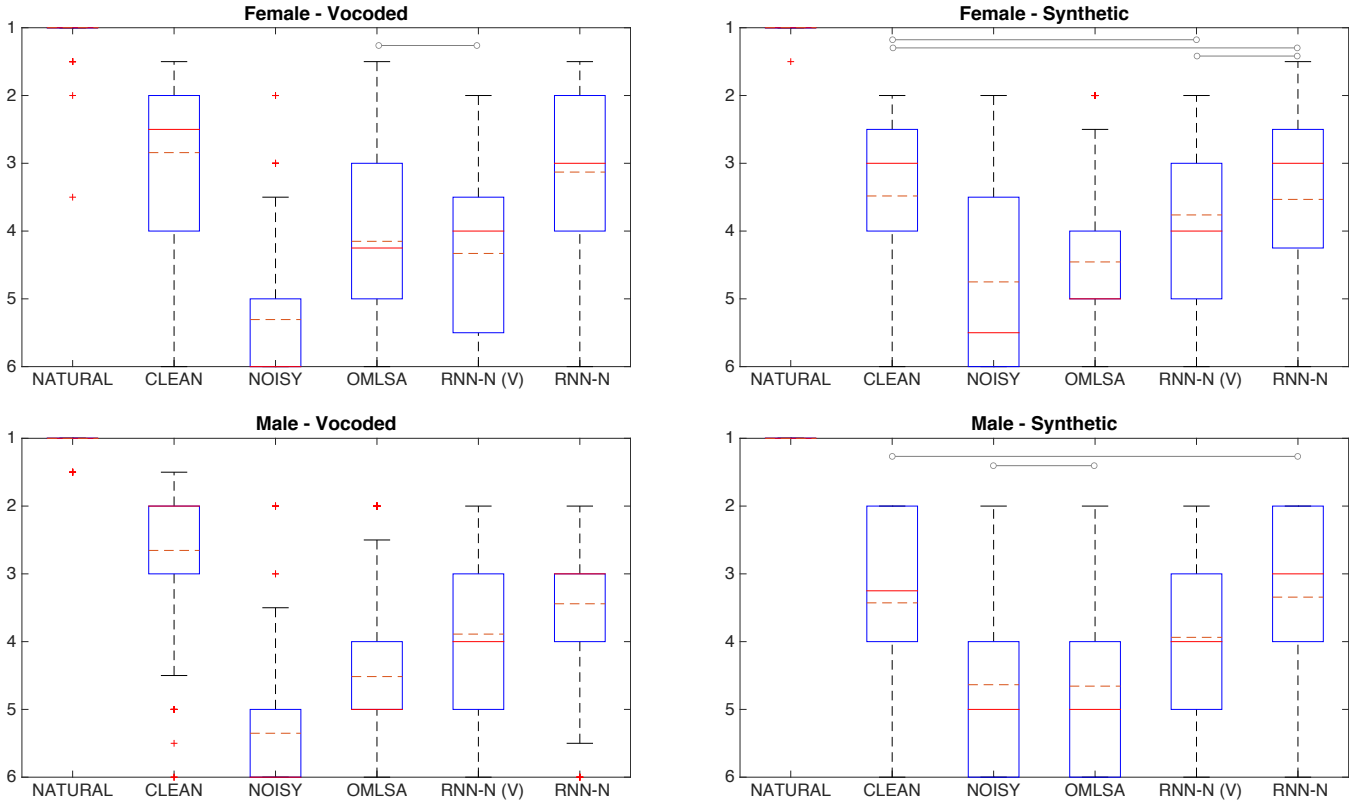

Fig. 3. Noisy data evaluation: rank order values for vocoded (left) and synthetic (right) speech of female (top) and male (bottom).

TABLE VI
NOISY DATA EVALUATION: MEAN SCORES OF LISTENING EXPERIMENT WITH VOCODED AND SYNTHETIC VOICES. NATURAL SPEECH IS OMITTED AS IT WAS ALWAYS RATED 100.

	CLEAN	NOISY	OMLSA	RNN-N (V)	RNN-N
Female - Vocoded	66.59	35.46	47.33	47.17	59.17
Male - Vocoded	68.02	33.75	43.61	51.50	54.53
Female - Synthetic	22.10	17.10	17.89	21.15	22.42
Male - Synthetic	27.54	20.10	20.69	23.66	26.84

methods is the same in all cases: OMLSA, followed by RNN-N (V) and RNN-N. The gap between RNN-N enhanced speech and CLEAN speech is smaller for the synthetic speech style than for the vocoded speech. The synthetic voice trained with RNN-N enhanced speech data was not found to be significantly different than the voice trained with clean speech for both voices. In contrast to what PESQ and STOI predicted (see Table III), the samples from RNN-N (V) were rated superior to noisy speech.

The gap between clean and noisy conditions is smaller for the synthetic voices, which can indicate that adaptation removed some of the noise of the data. The benefit of RNN-based methods is seen for all cases. The OMLSA method improvements seems to decrease after TTS acoustic model training. The absolute scores given to the noisy condition are at least 1.3 times lower than the ones given to clean voices.

D. Reverberation results

Participants rated five conditions: clean speech (CLEAN), reverberant speech (REVERB) and speech enhanced by Postfish (POSTFISH), and the RNN-based methods: RNN-R-r and RNN-R, as described in Section V-B. Vocoded and synthetic speech of each condition was evaluated.

Figure 4 shows the rank order boxplot for the female (top) and the male (bottom) voice of vocoded (left) and synthetic (right) speech. Table VII presents the raw scores given to each voice averaged across listeners and sentences.

As expected, natural clean speech (NATURAL) ranked highest and reverberated speech lowest for all cases. The order of increasing preference for all cases is: POSTFISH, followed by RNN-R-r and RNN-R. The rank difference between RNN-R enhanced speech and clean speech is smaller for the synthetic speech style than for the vocoded speech, but still remains a significant difference. The RNN-based methods improves listeners preferences of both vocoded and synthetic speech, while the POSTFISH method improvements seems smaller after TTS acoustic model training, which was also observed for the OMLSA method with additive noise. The raw scores given to the reverberant conditions are at least two times lower than the scores of the clean voices. The difference between clean and reverberant speech seems for vocoded and synthetic speech is very similar, which could indicate that acoustic model training via adaptation does not alter the reverberation effect substantially.

E. Additive noise and reverberation results

Listeners rated five conditions: clean speech (CLEAN), noisy reverberant speech (NOISYREVERB) and speech enhanced by the three RNNs: RNN-N, RNN-R, RNN-NR. We did not include any other baseline in this evaluation because

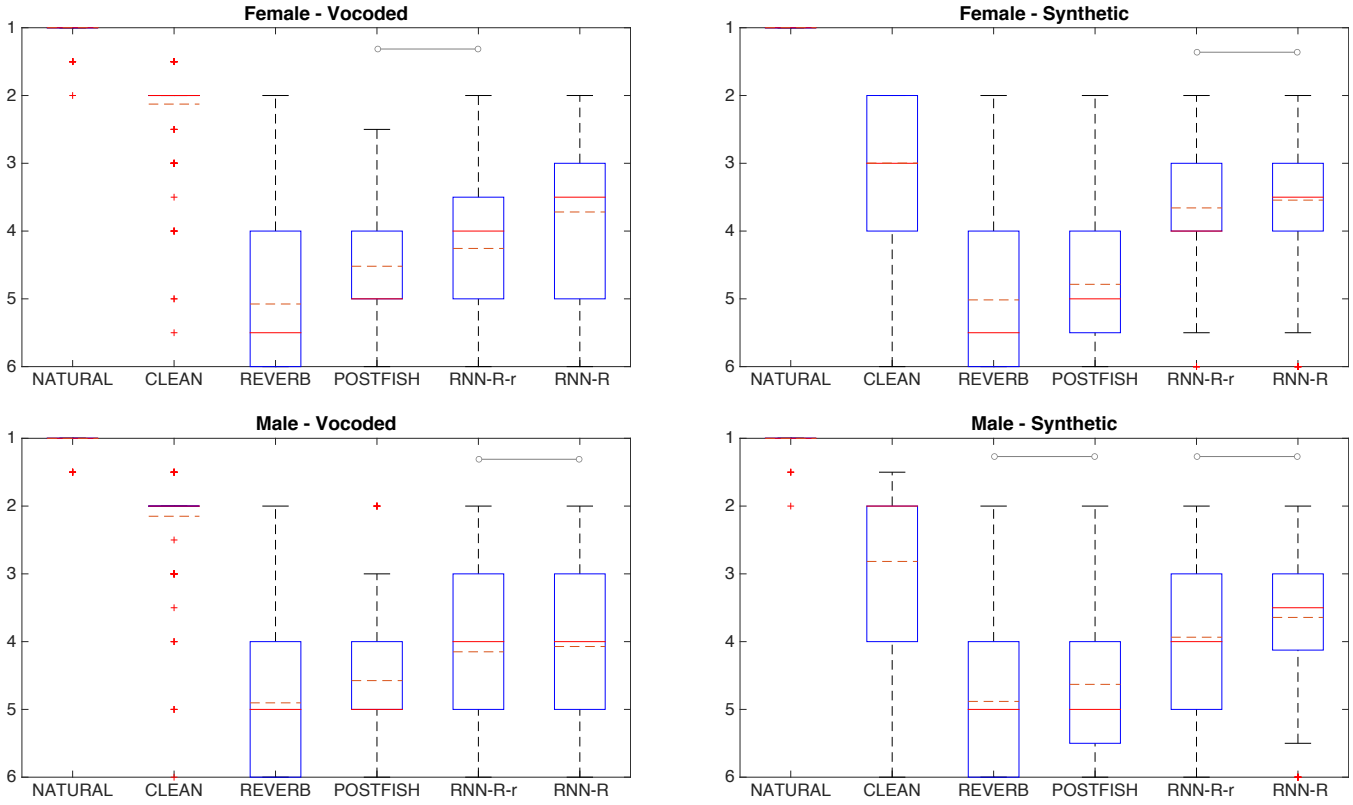

Fig. 4. Reverberant data evaluation: rank order values for vocoded (left) and synthetic (right) speech of female (top) and male (bottom).

TABLE VII
REVERBERANT DATA EVALUATION: MEAN SCORES OF LISTENING EXPERIMENT WITH VOCODED AND SYNTHETIC VOICES. NATURAL SPEECH IS OMITTED AS IT WAS ALWAYS RATED 100.

	CLEAN	REVERB	POSTFISH	RNN-R-r	RNN-R
Female - Vocoded	75.81	35.49	40.06	45.88	51.32
Male - Vocoded	71.15	30.83	33.68	39.87	40.76
Female - Synthetic	23.50	9.68	11.66	19.25	21.10
Male - Synthetic	30.59	11.57	14.97	20.33	24.06

their results were too poor in informal listening tests and objective measures.

Figure 5 shows the rank order boxplot for the female (top) and the male (bottom) voices of vocoded (left) and synthetic (right) speech. Table VIII presents the raw scores obtained by each voice averaged across listeners and sentences.

Clean speech ranked highest and noisy reverberant speech lowest in all conditions. The best results were obtained by the network trained with both noisy and reverberant data (RNN-NR) and the improvements were higher for the synthetic and the female voices. Interestingly, dereverberated noisy synthetic speech (RNN-R) ranked higher than noise suppressed reverberant synthetic speech (RNN-N) even though the preference on the vocoded speech level was the opposite. This could indicate that the HMM adaptation process is more robust to noise than to reverberation. The absolute scores given to the noisy reverberant condition are at least four times lower than the ones given to clean voices.

VIII. DISCUSSIONS

In this section we revisit the more interesting results presented here with regards to the speech enhancement framework (acoustic feature space and phase estimation) and to distortion conditions (additive noise and reverberation). Additionally, we contextualize our work in view of current trends of speech enhancement and text-to-speech.

A. Speech enhancement framework

From the noise only experiments we found that the reconstruction process required by the proposed speech enhancement framework does not negatively impact the extraction of TTS acoustic features using STRAIGHT. In terms of Mel cepstral distortion measures, enhancing the TTS acoustic features directly, without the need for waveform reconstruction, seems to be the better option. However, in terms of objective intelligibility and quality measures, as well as subjective listeners scores, the proposed framework achieved better results. We also observed that voiced/unvoiced and fundamental frequency related errors are smaller when the enhancement is done over the magnitude spectrum even though these features are not directly enhanced. In an unified framework where enhancement and acoustic model training are optimized together, operating on the same acoustic feature might not be ideal.

We observed from the reverberation experiments that phase estimation improves most objective measures although results with synthetic speech were not significantly better. This seems to indicate that the artifacts caused by incorrect phase information were not captured by the TTS acoustic model.

B. Effect of different distortions

For additive noise the synthetic voice trained with speech data enhanced by the proposed method was ranked as high as

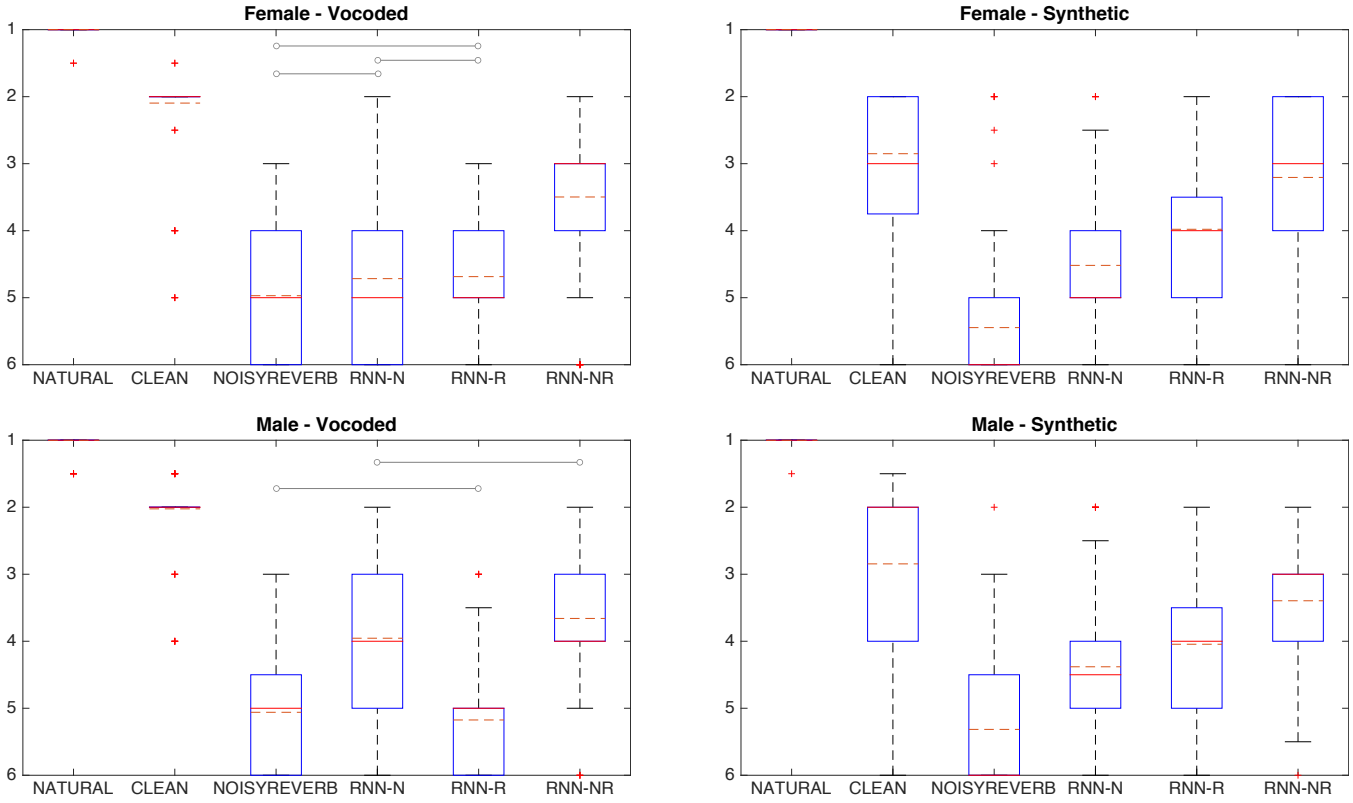

Fig. 5. Noisy and reverberant data evaluation: rank order values for vocoded (left) and synthetic (right) speech of female (top) and male (bottom).

TABLE VIII
NOISY AND REVERBERANT DATA EVALUATION: MEAN SCORES OF LISTENING EXPERIMENT WITH VOCODED AND SYNTHETIC VOICES. NATURAL SPEECH IS OMITTED AS IT WAS ALWAYS RATED 100.

	CLEAN	NOISYREVERB	RNN-N	RNN-R	RNN-NR
Female - Vocoded	74.67	17.16	22.75	20.71	38.26
Male - Vocoded	75.16	17.14	31.19	15.01	34.40
Female - Synthetic	19.12	3.20	8.31	11.57	16.49
Male - Synthetic	29.53	5.77	12.79	16.07	20.70

the voice trained with clean data. In most realistic situations noise and reverberation coexist. When both noise and reverberation are present, the RNN-based enhancement significantly improved the quality of both vocoded and synthetic voices, but the clean voice was still significantly better than the enhanced version. This happened even thought the network was also trained with relatively matching conditions (data corrupted by noise and reverberation) and with similar amount of data.

We observed, for the additive noise case, that enhancement results can be improved if the RNN model is trained with the speech data corrupted with the exact noise type and level that is present in the data that needs enhancing. This motivates the use, at run time, of modules that estimate SNR and noise characteristics so that a model trained with matched conditions can be applied. Another way to improve performance to unseen and more challenging scenarios (distortion type and speakers) would be to append speaker and environment information as additional input of the network, in order to facilitate learning. At enhancement time this information would be estimated from the distorted signal. Noise aware training has been shown to improve neural network based enhancement for larger and challenging datasets [6].

When comparing listeners results across the different distortion types (additive noise, reverberation) we can also make a few interesting observations. While additive noise decreases the segmental quality of vocoded and synthetic speech, reverberation tends to decrease intelligibility as well. When both distortions are present the resulting voice is not only of a lower quality but also often unintelligible. Listeners rated the noisy reverberant condition with scores at least four times lower than the clean condition, while individually the reverberant and noisy condition were rated two and 1.2 times worse than the clean condition respectively. Although we did not perform intelligibility tests, informal listening indicates that noisy reverberant speech enhanced with the RNN-R, i.e. reverberated noisy speech, sounds clearer and more intelligible, while the speech enhanced with the network trained with noisy data (RNN-N), i.e. noise suppressed reverberant speech, sounds less noisy but at times unintelligible. Samples of clean, distorted and enhanced vocoded and synthetic speech are available at: http://homepages.inf.ed.ac.uk/cvbotinh/se/

In all our experiments the distortion was artificially created by either adding noise that has been previously recorded or convolving speech with an impulse response recorded in a different room. This artificial scenario, although far from realistic, is useful because it allows us to create larger datasets for training the enhancement module. However, the performance of a system trained with artificially created noises can decrease substantially due to mismatched conditions if tested with real recordings of speech. In a more realistic situation we would not have access to the clean speech required for such parallel

sets. In such cases, other training frameworks where parallel data is not necessary have to be exploited.

C. Our work in context

Since the conclusion of our experiments other more advanced neural-network based denoising techniques have been proposed, such as generative adversarial networks [51], Wavenet-style based systems [52], [53] and convolutional neural networks [54], [55]. The latter showing improvements upon RNN based methods [55] in terms of PESQ and STOI scores. For noise supression and dereverberation, authors in [56] proposed to use an RNN to estimate the power spectral density (PSD) prior to prediction filter estimation and inverse filtering, showing ASR improvements compared to the baseline method that iteratively calculated PSD on the enhanced speech signal. None of these techniques have been evaluated in terms of how well they can improve TTS quality. Directly enhancing and synthesizing the raw waveform could be a way to avoid vocoder errors that appear when extracting features from distorted data. Using more complex models should also improve results for the more realistic scenario where speech and noise are not strictly uncorrelated and where the distortion have a longer temporal effect.

In this work we have not tried to enhance phase derived features and up to the conclusion of this work we were not aware of any published study on this subject. Noise suppression studies tend to focus on modifying the magnitude spectrum alone as it is mostly accepted that additive noise does not corrupt the phase enough to create audible artifacts. Speech enhancement is often also only seen as a front end for an automatic speech recognition engine, which tends to not require perfect phase information. Phase distortions caused by additive noise are perceptually relevant to humans [57]. Modelling the phase spectrum is however not an easy task. As opposed to the magnitude spectrum, the phase spectrum, if not properly extracted and unwrapped, does not present a clear continuous structure, even during voiced segments, as illustrated in [58]. One relevant work in this direction is the one recently published by [54], where the real and imaginary parts of the noisy spectrogram are mapped to that of the spectrogram of the clean signal using convolutional neural networks. Although authors obtained better objective results when compared to enhancing the magnitude spectrum in isolation (and using the noisy phase spectrum for reconstruction), they do not compare this with other phase reconstruction techniques such as the Griffin-Lim algorithm.

Finally, all experiments described in this work are on HMM-based speech synthesis and the adaptation techniques developed for that. For HMM-based speech synthesis model adaptation using a model trained with clean speech can be a way to improve noise robustness, as shown in this paper and in others cited here. Up to the submission of this work we are not aware of any work regarding the effect of noise and or reverberation on DNN-based speech synthesis and DNN adaptation techniques. Our own preliminary experiments indicate that both noise and reverberation have a significant impact on the quality of DNN-based TTS voices, but further experiments are still on going.

IX. Conclusions

In this paper we proposed the use of recurrent neural network to remove additive noise and reverberation of speech material used for training a text-to-speech system. We presented a series of objective and subjective evaluations on the quality of vocoded and synthetic speech created from clean (studio recordings), distorted (noisy and/or reverberated recordings) and speech that has been enhanced using recurrent neural networks. To train the network we extracted the magnitude Fourier transform of clean and distorted speech and use it as target and input of the network respectively. The network is trained with a variety of noise types and levels as well as reverberation effects caused by a range of impulse responses derived from different rooms and microphone positions. We found that synthetic speech quality can be significantly improved by simply improving the quality of the recordings used for training the voices. The most challenging scenario, where both additive noise and reverberation are present, was judged worst by listeners. However, enhancement was still significantly beneficial.

Acknowledgements This work was partially supported by EPSRC 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.

References

[1] 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.

[2] 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.

[3] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "Tundra: a multilingual corpus of found data for tts research created with light supervision," in *interspeech*, 2013, pp. 2331–2335.

[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] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *J. on Advances in Signal Processing*, vol. 2016, no. 7, pp. 1–19, 2016.

[12] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Proc. REVERB Challenge Workshop*, Florence, Italy, May 2014.

[13] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[14] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E.-S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Proc. REVERB Challenge Workshop*, Florence, Italy, May 2014.

[15] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll, "The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement," in *Proc. REVERB Challenge Workshop*, Florence, Italy, May 2014.

[16] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in neural information processing systems*, vol. 13, pp. 758–764, 2001.

[17] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment." in *Proc. Interspeech*, 2007, pp. 854–857.

[18] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. on Audio, Speech and Language Processing.*, vol. 16, no. 8, pp. 1433–1451, 2008.

[19] H. W. Lollmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. ICASSP*, 2009, pp. 3989–3992.

[20] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. on Audio, Speech and Language Processing.*, vol. 17, no. 2, pp. 231–246, 2009.

[21] K. Han, Y. Wang, D. Wang, W. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Trans. on Audio, Speech and Language Processing.*, vol. 23, no. 6, pp. 982–992, June 2015.

[22] 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 Communication*, vol. 27, pp. 187–207, 1999.

[23] 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*, San Francisco, USA, 2016.

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

[25] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. SSW*, San Francisco, USA, Sept. 2016.

[26] 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.

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

[28] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*. Antibes, France: IEEE, Sept. 2014, pp. 313–317.

[29] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ace challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2015.

[30] J. Y. Wen, N. D. Gaubitch, E. A. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*. Paris, France: IEEE, Sept. 2006.

[31] E. A. Habets, "Room Impulse Response Generator," International Audio Laboratories Erlangen, Fraunhofer, Tech. Rep., 09 2010.

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

[33] 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.

[34] Monty, "Postfish - Xiph SVN repository," https://svn.xiph.org/trunk/postfish/, 2005.

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

[36] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects*, vol. 10, 2010.

[37] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr 1984.

[38] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. MAVEBA*, Florence, Italy, September 2001.

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

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

[41] 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.

[42] K. Yanagisawa, J. Latorre, V. Wan, M. J. F. Gales, and S. King, "Noise robustness in HMM-TTS speaker adaptation," in *Proc. SSW*, Barcelona, Spain, August 2013, pp. 139–144.

[43] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive hmm-based text-to-speech synthesis," *IEEE Trans. on Audio, Speech and Language Processing.*, vol. 17, no. 6, pp. 1208–1230, Aug 2009.

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

[45] 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.

[46] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," in *NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, Dallas, USA, March 2010, pp. 4214–4217.

[48] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, Salt Lake City, USA, May 2001, pp. 749–752.

[49] *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.

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

[51] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: speech enhancement generative adversarial network," in *Proc. Interspeech*, Stockholm, Sweden, 2017.

[52] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," *CoRR*, vol. abs/1706.07162, 2017.

[53] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian Wavenet," in *Proc. Interspeech*, Stockholm, Sweden, 2017.

[54] S. W. Fu, T. y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. MLSP*, Sept 2017, pp. 1–6.

[55] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, Stockholm, Sweden, 2017.

[56] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverbertion," in *Proc. Interspeech*, Stockholm, Sweden, 2017.

[57] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465 – 494, 2011.

[58] F. Espic, C. Valentini-Botinhao, and S. King, "Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis," in *Proc. Interspeech*, 5 2017.

Cassia Valentini-Botinhao is a research associate at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. She graduated from the Federal University of Rio de Janeiro, Brazil, receiving the title of Electronic Engineer in 2006 and received an MSc from the University of Erlangen-Nuremberg in Germany in 2009, on Systems of Information and Multimedia Technology. As a Marie Curie Fellow, Cassia obtained her PhD in University of Edinburgh, UK, with the thesis "Intelligibility enhancement of synthetic speech in noise". Her research interests are speech intelligibility and signal processing for speech synthesis.

Junichi Yamagishi (SM'13) is an associate professor of National Institute of Informatics in Japan. He is also a senior research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. Since 2006, he has authored and co-authored over 100 refereed papers in international journals and conferences. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists Prize from the Minister of Education, Science and Technology, the JSPS prize in 2010, 2013, 2014, and 2016, respectively. He was one of organizers for special sessions on "Spoofing and Countermeasures for Automatic Speaker Verification" at Interspeech 2013, "ASVspoof evaluation" at Interspeech 2015 and "Voice conversion challenge 2016" at Interspeech 2016. He has been a member of the Speech & Language Technical Committee (SLTC) and an Associate Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He is a Lead Guest Editor for the IEEE Journal of Selected Topics in Signal Processing (JSTSP) special issue on Spoofing and Countermeasures for Automatic Speaker Verification.