INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Investigating the Effects of Noisy and Reverberant Speech in Text-to-Speech Systems

David Ayllón, Héctor A. Sánchez-Hevia, Carol Figueroa, Pierre Lanchantin

ObEN Inc., CA, USA
david@oben.com

Abstract

The quality of the voices synthesized by a Text-to-Speech (TTS) system depends on the quality of the training data. In real case scenario of TTS personalization from user's voice recordings, the latter are usually affected by noise and reverberation. Speech enhancement can be useful to clean the corrupted speech but it is necessary to understand the effects that noise and reverberation have on the different statistical models that compose the TTS system. In this work we perform a thorough study of how noise and reverberation impact the acoustic and duration models of the TTS system. We also evaluate the effectiveness of time-frequency masking for cleaning the training data. Objective and subjective evaluations reveal that under normal recording scenarios noise leads to a higher degradation than reverberation in terms of naturalness of the synthesized speech.

Index Terms: text-to-speech, speech enhancement, speech synthesis.

1. Introduction

Recent advances in Text-To-Speech (TTS) technology are noteworthy and now synthesized voices sound closer to humans than ever. The next milestone in this technology is the generation of personalized voices of any person [1, 2, 3]. The biggest challenge in building personalized TTS systems is to obtain a high quality training corpus from a particular voice to either build a speaker-dependent model or a speaker-adapted model using a pre-trained base model [4, 5]. In any case, the quality of synthetic voices is highly affected by the presence of noise and reverberation in the training corpus [6, 7]. One alternative is to identify and discard corrupted data, but this solution is only feasible when a large amount of training data is available, which is not the typical case in TTS personalization [8]. A better alternative is the use of speech enhancement techniques to clean the training data. Traditional techniques based on spectral subtraction [9] or the Wiener filter [10] are effective in reducing noise, however, they introduce spectral distortions that are perceptually unacceptable in synthesized voices. Statistical model-based algorithms [11, 12] and subspace-based approaches [13, 14] find a compromise between speech distortion and noise reduction. Unfortunately, none of these methods aim at reducing reverberation. In recent years, Time-Frequency (TF) masking has attracted the attention of the speech enhancement community. The Ideal Binary Mask (IBM) [15] has demonstrated its ability to improve the intelligibility of noisy speech but it introduces some musical noise. The Ideal Ratio Mask (IRM) [16] has similar properties to the IBM reducing musical noise and improving the overall quality. The original formulation of both TF masks only considers noise but they can be reformulated to suppress reverberation. A recent approach for robust TTS is to use deep learning models to map acoustic parameters extracted from corrupted speech to those extracted from clean speech [17, 18, 19].

Many studies about the effects of noise and speech enhancement in Automatic Speech Recognition (ASR) can be found in the literature. However, there are not many studies about the effects of noise, reverberation, and the application of speech enhancement techniques for TTS. The most detailed study we found in the literature is [7], in which the authors evaluate the effects of noise and reverberation on a speaker-adapted TTS system and propose a TF masking method based on a Deep-Neural Network (DNN) to enhance the training data. The objective of this paper is to perform a thorough assessment of how noise and reverberation affect the different statistical models that compose the TTS system or are involved in its training: the Forced-Aligner (FA), the Acoustic Model (AM), and the Duration Model (DM). In addition, we evaluate the application of TF masking to enhance the training data. The evaluation is carried out by training a speaker-dependent DNN-based TTS system with a Korean speech corpus for which we have manually annotated the phone boundaries. We train several TTS systems adding different levels of noise and reverberation to the training data, and then we perform objective and subjective evaluations of how noise and reverberation affect the FA, AM, DM, and the whole system.

2. Database

2.1. Clean speech

For clean speech we use the Korean corpus SiTEC Synth-Male01 - *Read Sentences Corpus for Prosody Synthesis* [20] which contains 4392 sentences of a Korean male professional announcer recorded at 16 kHz. Professional native linguists manually annotated the phone boundaries of 2472 sentences, thus, we use approximately 56% of the whole corpus (5.2 h). For testing the TTS system we exclude 5% of these sentences from the training set.

2.2. Noise corpus

Noise recordings were taken from the DEMAND database [21], which contains 16-channel recordings at 48 kHz of 18 acoustic scenes that are divided into 6 categories: domestic, nature, office, public, street, and transport. Noise is added to speech with controlled signal-to-noise ratio (SNR) using the ITU-T P.56-B standard [22]. Noise segments are randomly selected among all the recordings of a given category (microphones and acoustic scenes). In addition to these 6 noise categories we generated speech-shaped noise by filtering white noise with the 6th order Linear Predictive Coding (LPC) coefficients obtained from the speakers included in the ACE corpus (9 male and 5 female) [23].

2.3. Reverberation

To generate realistic mixtures we selected 140 impulse response (IR) recordings from the ACE challenge corpus [23], in particu-

http://dx.doi.org/10.21437/Interspeech.2019-3104

lar the 2-channel, 3-channel, and 5-channel IR recordings. The selected IRs were recorded at 48 kHz and have a reverberation time (RT60) that ranges from 0.29 s to 1.33 s, and a direct-to-reverberant ratio (DRR) that ranges from of -2.27 dB to 14.63 dB. In our experiments we want to perform a parameter sweep (RT60 and DRR) which is not possible using real IR recordings. To adjust the reverberation parameters, we modified the IR recordings as follows: 1) an IR recording is randomly selected among the ten closest IRs in terms of RT60; 2) the selected IR is modified to approximately match one of the two reverberation parameters; and 3) the other reverberation parameter is modified. RT60 is artificially extended using time-stretching, whereas DRR is artificially controlled modifying the gain of the direct signal in the IR recordings. The direct signal segment is modified by applying a shifted and scaled Hanning window of length 15 ms around the direct signal peak, without modifying the rest of the IR. To estimate the reverberation parameters from the modified IRs we used the methods in [23]. Compared against the ground truth reported in the ACE corpus, our implementation has a Mean Average Error (MAE) of 0.025 s ± 0.024 for the RT60 and 0.198 dB ± 0.217 for the DRR. Reverberant speech is obtained by convolving clean speech with the generated IR.

2.4. Time-Frequency Masking

We also evaluate the effects of applying TF masking to enhance the corrupted speech signals before training. We consider two different TF masks: a modification of the IBM [15] and the complex IRM (cIRM) [24]. Let us define $Y(k,n) = Y_D(k,n) + Y_R(k,n) + Y_N(k,n)$, where n and k are time and frequency indices, respectively, $Y_D(k,n)$ is the direct-path speech signal, $Y_R(k,n)$ is the addition of all speech reflections, and $Y_N(k,n)$ is the additive noise. The original IBM assigns '1' to TF bins where the SNR is higher than a threshold (usually 0 dB), but it does not consider reverberation. We use the next modification thereof:

$$IBM(k,n) = \begin{cases} 1 & \text{if } |Y_D(k,n)| > |Y_R(k,n)| + |Y_N(k,n)| \\ 0 & \text{otherwise ,} \end{cases} \tag{1}$$

which compares the power of the direct-path speech signal with the addition of the power of the reverberation and noise signals. The direct-path signal is obtained by convolving the speech signal with a segment of the IR that goes from its beginning to 7.5 ms after the first peak. According to [25], the cIRM is obtained as $cIRM(k,n) = Y_D(k,n)/Y(k,n)$.

3. Text-To-Speech system

The TTS AM, which maps linguistic features to acoustic features, is based on a DNN trained using the Merlin Speech Synthesis toolkit [26]. FA is performed to obtain phone boundaries which are used to train the DM based on a Gradient Boosting Regression Tree (GBRT) [8]. Speech signals are synthesized from acoustic features using the WORLD vocoder [27]. A description of the forced-aligner and TTS system follow.

3.1. Forced-Aligner

The forced-aligner is based on a DNN-HMM speech recognizer. A tri-phone DNN-HMM acoustic model (FA-AM) is trained on top of a GMM-HMM system using Mel-Frequency Cepstral Coefficients (MFCCs) and delta features. The dimension of the feature vector is reduced to 40 using Linear Discriminant Analysis (LDA), and then Maximum Likelihood Linear Transform (MLLT) is applied for each speaker. Speaker Adaptive Training (SAT) [28] is performed. The DNN model is trained on top of the SAT model with the same set of training data. The DNN configuration and training algorithm are described in [29]. The Korean FA-AM was trained using the King-ASR-174 corpus [30], which contains approximately 290 h of spoken speech recorded by 200 different native speakers.

During alignment, a 2-pass decoding strategy is used, involving Feature-space Maximum Likelihood Linear Regression (fMLLR) speaker adaptation [31]. First, the LDA+MLLT acoustic features are extracted from input audio. Second, speaker-independent forced-alignment is performed. Then the FA-AM is adapted to the current speaker using the obtained phonetic alignment, and finally speaker-adapted FA is performed based on the Viterbi algorithm.

3.2. Acoustic and duration models

The input features for the AM consist of 479 linguistic features. Most features are binary answers to linguist context questions (e.g. quinphone identity, part-of-speech) of the current, preceding, and following phone. Other numerical features include the duration of the phone, and positional information. Further details can be found in [8].

The WORLD [27] vocoder was used to extract 60-dimensional Mel-frequency Cepstral Coefficients (MFCCs), 3-dimensional band aperiodicities (BAPs), and the fundamental frequency on log scale ($\log F_0$) at 5 ms frame intervals. The output features of the AM are: MFCCs, BAPs and $\log F_0$ with their deltas and delta-delta values, and a voiced/unvoiced binary feature.

The AM consists of a DNN with 10 feed-forward hidden layers. Each hidden layer has the corresponding number of hyperbolic tangent units 1024, 512, 512, 256, 256, 512, 512, 512, 1024, 1024. The output layer has a linear activation function. The AM was trained for 50 epochs, having the batch size set to 512, learning rate fixed at 0.001, warm-up momentum set to 0.3, and drop-out rate set to 0.02.

4. Objective evaluation

4.1. Effects on forced-alignment

The effects of noise and reverberation on the FA have been evaluated independently. To evaluate the effects of reverberation, reverberant speech has been generated with RT60 values that range from 0.2 s to 1.2 s in steps of 0.2 s, and DRR values that range from -5 dB to 20 dB in steps of 5 dB. To evaluate the effects of noise, the 7 different types of noise previously described have been added to clean speech with SNR that ranges from -5 dB to 20 dB in steps of 5 dB.

We evaluated a total of 78 acoustic conditions. For each condition, the degraded speech corpus was forced-aligned. The metric used to evaluate the quality of the alignments is the F-score, which reflects the number of phones matching inside a match window of 50 ms. A phone matches when the addition of the onset and offset mismatch with respect to the manual annotations is lower than 50 ms. Figure 1 represent the F-score as a function of the different acoustic conditions. According to the results, the FA is quite robust to any type of noise for SNRs higher than 10 dB, but the performance starts to decrease for SNRs lower than 5 dB. Comparing the different types of noise, the ones containing voices affect the FA the most. This result

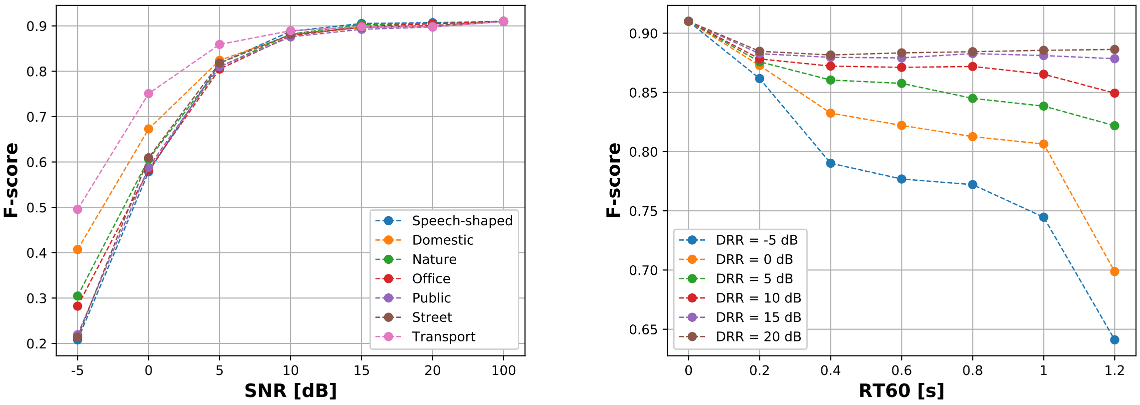

Figure 1: *Effects of noise (left) and reverberation (right) in Forced-Alignment.*

is expected since noises that overlap with speech frequencies are commonly the most problematic. Regarding reverberation, results show that the FA is affected more by low DRR levels than by high RT60. For DRR values above 5 dB the F-score remains practically constant for any RT60. However, for lower DRRs, the F-score gradually decreases as the RT60 increases.

4.2. Effects on TTS acoustic model

In this section we evaluate the effects of noise and reverberation in the AM by computing the vocoder parameter distortions: Mel-Cepstral coefficients distortion (MCEP) in dB, BAP distortion in dB, F_0 distortion in Hz, F_0 correlation and V/UV distortion in %. We also evaluate the effects of TF masking by processing the training set with the IBM and cIRM before TTS training. To evaluate how TF masking enhances training data, we compute the Perceptual Evaluation of Speech Quality (PESQ) [32] and the Short-Time Objective Intelligibility (STOI) [33] objective measures, which are correlated with speech quality and speech intelligibility, respectively. All objective measures have been calculated using clean speech as reference. In addition, the FA F-score of the enhanced training corpus is computed.

Unfortunately, AM training is very time consuming, therefore, we only evaluated five different acoustic scenarios that combine different levels of noise and reverberation (4 degraded scenarios in addition to clean speech). Degraded scenarios have been selected considering real scenarios where end users may record their voices. The acoustic parameters of the degraded scenarios are summarized in Table 1. In total we trained 18 TTS systems: for each of the 4 degraded scenarios we trained the system using the unprocessed (degraded) audios both with manual and forced alignments, and the audios enhanced by the IBM and cIRM with forced alignments. We also trained the system with the original clean audios both with manual and forced alignments. Table 2 includes the different training sets with their corresponding FA F-score, PESQ and STOI values. We can see how the original F-score obtained with clean audios (0.91) is affected by the different noise and reverberation levels of the considered scenarios (in the worst case F-score decreases to 0.77). Regarding TF masking, both IBM and cIRM introduce important improvements for all acoustic scenarios, cIRM being more effective than IBM in terms of PESQ and STOI.

Table 1: *Acoustic scenarios for TTS evaluation.*

Scenario	RT_{60} (s)	DRR (dB)	SNR (dB)	Noise Type
Domestic	0.5	10	10	Demand: domestic
Street	0.2	15	10	Demand: street
Restaurant	0.7	15	5	Demand: cafe+restaurant
Hall	1.2	10	20	Artificial: Speech-shaped

Table 3 contains the vocoder parameter distortions for the 18 trained AMs. Comparing AMs trained with manual (MA) and automatic alignments (FA), alignment errors affect more the F0 than MCEP and BAP, for any scenario. In general, vocoder parameters are notably affected by noise and reverberation in the four scenarios. The lowest MCEP distortions are obtained in Hall, and the highest in Domestic, which also has the highest F0 distortion. BAP distortions are quite similar for Domestic, Restaurant and Hall, and slightly smaller for Street. V/UV distortions are also equivalent for all scenarios except for Restaurant, where distortions are slightly higher. Considering the noise and reverberation levels of each scenario the results suggest that noise affects more than reverberation to MCEP, F0 and V/UV parameters, but reverberation influences more BAP.

5. Subjective evaluation

To evaluate the effect of degraded speech we conducted a listening test composed of four tasks designed to evaluate different effects on the TTS output. The full test contained 68 synthesized sentences [1] that were evaluated by 40 native Korean speakers. The first three tests were a ranking task composed of 12 sentences each. Participants had to order the sentences generated by different TTS models according to their naturalness. The fourth test was a rating task composed of 32 sentences. Participants were presented with a TTS sentence and asked to rate its naturalness from 1 to 7. Figure 2 shows the results of the first

Table 2: *F-score, PESQ and STOI for the different training sets.*

Training set	FA F-score	PESQ	STOI
Clean Speech	0.91	-	-
Domestic UN	0.79	2.27	0.85
Domestic w/IBM	0.83	2.99	0.92
Domestic w/cIRM	0.83	3.60	0.96
Street UN	0.81	2.28	0.89
Street w/IBM	0.85	3.31	0.95
Street w/cIRM	0.83	3.77	0.97
Restaurant UN	0.77	1.97	0.81
Restaurant w/IBM	0.78	2.97	0.94
Restaurant w/cIRM	0.83	3.59	0.97
Hall UN	0.79	2.10	0.82
Hall w/IBM	0.82	2.86	0.91
Hall w/cIRM	0.83	3.52	0.95

[1] Audio samples are provided as supplementary files.

Table 3: *Vocoder parameter distortions in the test set for the different TTS AM models.*

TTS AM	**MCEP** (dB)	**BAP** (dB)	**F0** (Hz)	**F0-corr**	**V/UV** (%)
Clean					
UN - MA	4.72	0.55	13.57	0.69	7.02
UN - FA	5.15	0.61	14.38	0.64	7.57
Domestic					
UN - MA	16.54	0.90	17.26	0.49	12.83
UN - FA	17.10	1.01	20.53	0.30	16.56
IBM - FA	8.09	0.93	22.81	0.22	13.88
cIRM - FA	6.21	0.73	15.08	0.60	8.56
Street					
UN - MA	13.45	0.83	15.89	0.52	12.18
UN - FA	13.51	0.89	17.99	0.42	15.36
IBM - FA	7.91	0.83	19.07	0.38	12.33
cIRM - FA	6.37	0.68	14.79	0.61	8.50
Restaurant					
UN - MA	14.31	0.97	16.38	0.50	14.24
UN - FA	14.48	1.04	21.11	0.21	18.93
IBM - FA	8.80	0.93	17.76	0.46	14.15
cIRM - FA	6.13	0.72	14.71	0.61	8.23
Hall					
UN - MA	8.79	0.94	16.01	0.54	12.44
UN - FA	9.84	1.00	19.49	0.35	15.59
IBM - FA	8.35	0.94	18.24	0.42	14.13
cIRM - FA	6.37	0.74	14.97	0.61	8.69

three tests; the fourth test results are in Figure 3.

(a) *Effects on the AM*: the sentences were synthesized using manual alignments and the AM trained on corrupted speech with FA. Results show that, mild noise levels are preferable to high reverberation times, however, the detrimental effect of noise is amplified when it is combined with reverberation. In general, low SNR seems like the worst scenario.

(b) *Effects on the DM*: in this case we combined the DM trained on corrupted speech with the AM trained on clean speech with manual alignments. Results show that the DM seems equally affected by large reverberation and moderate noise levels, which is in line with the results of Section 4.1. Distinction among noise types is not so clear in this case, but less reverberation is still desirable and low SNR is again the worst scenario.

(c) *Effects on the TTS system*: The third test represents the real scenario where both the AM and the DM are trained on corrupted speech. Results are very similar to those obtained in the first test, which points at the AM having more influence in TTS naturalness. Reverberation also influences more compared to the first test.

(d) *Effects of TF masking*: The fourth test compares the performance of TF masking on naturalness. Originally we intended to include the clean signals, but following an informal listening test, TTS generated with clean signals was clearly superior in terms of sound quality (the masked versions have less low frequency content and have some background noise), whereas differences in prosody were much less noticeable. Likewise, the TTS produced from any of the degraded signals was too inferior in quality for a direct comparison with the masked version. From the results it is clear that cIRM is superior to IBM as predicted by the objective scores. When separated by noise type, the results show a trend towards worse performance in more reverberant scenarios, especially for the IBM.

According to the results, noise tends to have a greater effect on perceived naturalness than reverberation, although their influence is cumulative and both of them are highly detrimental. Comparing the results of the first three tests, it seems like non-expert listeners give more importance to degradation of the AM rather than the DM. Regarding the effectiveness of speech enhancement, results from the masked versions of the degraded audio clearly show that better source quality relates to better TTS quality, thus, improving the quality of corrupted speech prior to TTS training should be a priority.

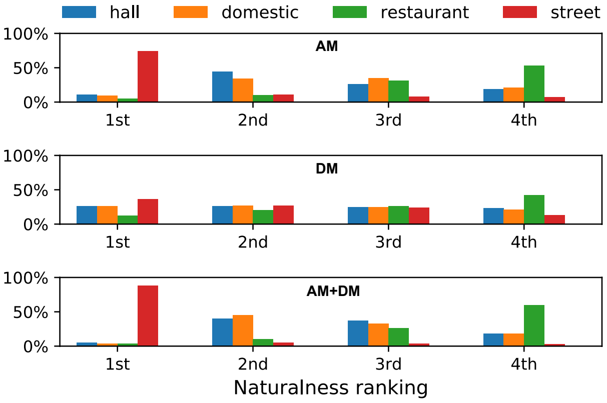

Figure 2: *Results of the listening test. From top to bottom: a) Effects on AM, b) Effects on DM, c) Effects on both AM and DM*

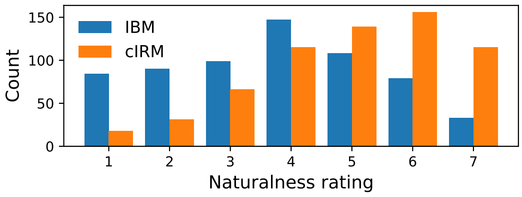

Figure 3: *Effects of Time Frequency Masking on Naturalness.*

6. Conclusions

In this work we have carried out a detailed study of how noise and reverberation affect the different statistical models that compose a TTS system. Regarding the FA, the model is quite robust to normal noise and reverberation levels. FA performance only drops for DRR values lower than 5 dBs and SNR values lower than 10 dBs. According to the vocoder distortions, both noise and reverberation have an important impact in the TTS AM, whereas the DM performance is highly correlated with the FA performance, as it is expected. Listening tests point at noise as being a more important issue for the perceived naturalness of speech than reverberation. Noise gets intertwined with the synthetic voice and tends to flatten and accelerate the speech rate, whereas reverberation adds a metallic quality to the synthesized voice and tends to elongate vowels. Subjective evaluation also reveals that vocoder distortions are not necessarily correlated with the naturalness of the synthesized speech perceived by listeners. Comparing the different types of noise, those that overlap with the speech spectrum (e.g, babble) affect the TTS system the most, as it was expected. We also demonstrated that TF masking can greatly improve the quality of the TTS and that PESQ and STOI measures are correlated with the naturalness of the synthesized voices.

Overall we can assert that, for normal recording scenarios, noise has a higher impact than reverberation on the TTS system. In the future we will expand this study using more speakers and languages as well as evaluating the impact of noise and reverberation on speaker-adapted TTS systems.

7. Acknowledgements

We would like to thank Dr.Iris Ouyang and Dr.Shukhan Ng for their help in the elaboration of the listening tests.

8. References

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

[2] T. Mills, H. T. Bunnell, and R. Patel, "Towards personalized speech synthesis for augmentative and alternative communication," *Augmentative and Alternative Communication*, vol. 30, no. 3, pp. 226–236, 2014.

[3] Y.-C. Huang, C.-H. Wu, and S.-L. Lin, "Personalized natural speech synthesis based on retrieval of pitch patterns using hierarchical fujisaki model," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7844–7848.

[4] 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 Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[5] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4475–4479.

[6] R. Karhila, U. Remes, and M. Kurimo, "Noise in hmm-based speech synthesis adaptation: Analysis, evaluation methods and experiments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 285–295, 2014.

[7] C. Valentini-Botinhao and J. Yamagishi, "Speech enhancement of noisy and reverberant speech for text-to-speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1420–1433, 2018.

[8] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, and I. Ouyang, "Data selection for improving naturalness of tts voices trained on small found corpuses," 12 2018, pp. 319–324.

[9] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[10] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[12] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 845–856, 2005.

[13] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[14] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[15] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 79–82.

[16] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[17] D. Gowda, H. Kallasjoki, R. Karhila, C. Contan, K. Palomäki, M. Giurgiu, and M. Kurimo, "On the role of missing data imputation and nmf feature enhancement in building synthetic voices using reverberant speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[18] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4390–4394.

[19] 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 *Interspeech*, 2016, pp. 352–356.

[20] SiTEC, "Synthmale01 - read sentences corpus for prosody synthesis," http://sitec.or.kr/#speechcopora.

[21] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[22] P. ITU-T, "Objective measurement of active speech level," *ITU-T Recommendation*, 1993.

[23] J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, N. D. Gaubitch, J. Eaton *et al.*, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1681–1693, 2016.

[24] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.

[25] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[26] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016.

[27] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, 2016.

[28] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker Adaptive Training," in *Proc. ICSLP*, 1996.

[29] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. ICASSP*. IEEE, 2014, pp. 215–219.

[30] Speechocean, "Korean speech recognition corpus (desktop)-sentences-200 speakers." [Online]. Available: http://kingline.speechocean.com/exchange.php?id=2959act=view

[31] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation Within the MLLR Framework," *Computer Speech and Language*, vol. 10, 1996.

[32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.