



Comparing Speech Enhancement Techniques for Voice Adaptation-Based Speech Synthesis

Nicholas Eng¹, C. T. Justine Hui¹, Yusuke Hioka¹, Catherine I. Watson²

¹Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland, New Zealand

²Department of Electrical, Computer and Software Engineering, University of Auckland, New Zealand

neng668@aucklanduni.ac.nz, justine.hui@auckland.ac.nz, yusuke.hioka@ieee.org, c.watson@auckland.ac.nz

Abstract

This study investigates the use of speech enhancement techniques in creating text-to-speech voices with degraded or noisy speech. A number of synthetic voices were created using speech that was first degraded by different noise types at various signal-to-noise ratios (SNRs), then enhanced through four speech enhancement algorithms: Subspace, Wiener filter, SEGAN and a DNN-based method. Subjective listening tests show that the quality of the synthetic voices produced by subspace and the DNN-based method enhanced speech outperforms the quality of the voices created using Wiener filter or SEGAN enhanced speech at low SNRs, and speech enhanced by the subspace method results in higher quality synthetic speech at higher SNRs.

Index Terms: speech synthesis, speech enhancement, speech quality, subjective listening tests

1. Introduction

Speech synthesis has many applications in today's life such as personal assistants in mobile phones and GPS navigation systems. Creating a synthetic voice usually requires speech data which is both high in quality and quantity, however, obtaining high quality speech and/or a large quantity of speech can be difficult in a number of instances. Some examples include creating voices from found data (such as for someone who has lost their voice), whereby there may only be a few noisy recordings of the individual, or creating personalised voices, where one cannot expect the target speaker to record their voice in a quiet acoustical environment or record for long periods of time.

In order to create voices from a low quantity of recordings, speech adaptation techniques, whereby a target speaker's voice model is adapted from an average speaker's voice model with a low number of sentences, can be used [1]. However, when creating voices from low quality data, regardless of using speech adaptation, it is known that voices made with clean speech are preferable to voices made with noisy speech [2]. Therefore, the issue of using low quality speech samples to create synthetic voices still needs to be addressed, such as by pre-processing using speech enhancement techniques.

Speech enhancement is a heavily researched area, and consequently, there are a multitude of approaches to remove noise from speech. Traditional methods such as Wiener filter [3] or spectral subtraction [4] have been used for speech enhancement for decades, but recently more modern techniques such as utilising deep neural networks (DNN), recurrent neural networks (RNN) or convolutional neural networks (CNN) [5–7] have be-

come the main focus for research in this field, and have shown to perform comparably to traditional methods.

However, it is important to note that the vast majority of speech enhancement techniques are not designed with speech synthesis in mind [8]. Different methods of speech enhancement on noisy speech can distort the desired speech signal whilst removing the noise, and/or may result in undesirable artefacts, which when made into a synthesised voice can be detrimental to the overall quality of the synthesised voice as well as affecting the speaker's personal characteristics of the voice. As a result, speech enhancement techniques that result in high-quality speech in a subjective manner may not lead to high-quality synthetic speech.

Although comparisons between differing speech enhancement techniques for natural speech are plentiful, there are very few studies that investigate the quality of synthetic speech made from noisy speech enhanced by speech enhancement techniques. In this paper, we investigate the use of both traditional and modern speech enhancement techniques in enhancing noisy speech used to create synthetic voices through voice adaptation, and compare the quality of the synthetic voices through a subjective listening test.

2. Experimental design

To evaluate the performance of the speech enhancement methods in the creation of synthetic voices, a subjective listening test was conducted whereby participants were asked to judge the quality of synthetic voices created with enhanced speech.

2.1. Speech Enhancement Methods

We evaluate four well known speech enhancement methods to enhance speech to be used for speech synthesis - subspace, Wiener filter, SEGAN and a DNN-based method.

The subspace method decomposes the vector space of the noisy signal into a signal subspace (containing both signal and noise components) and noise subspace (containing only noise components) [9]. Enhanced speech is obtained by nulling the noise subspace and removing the noise component in the signal subspace. This implementation has built-in prewhitening and thus is suitable to use for coloured noise.

The Wiener filter implementation is based on an *a priori* signal-to-noise estimation method [10]. This method places constraints on the clean and noise distortions, and estimates the *a priori* SNR ζ_k through a weighted combination of past and future estimates of ζ_k for each frame in the audio sample.

Speech Enhancement Generative Adversarial Network

(SEGAN) utilises Generative Adversarial Networks (GANs) to enhance speech [11]. It consists of a generator, a convolutional neural network that is trained to clean the noisy speech, and a discriminator, which is a binary classifier that learns to distinguish true clean speech (from the training data) and fake clean data (from the generator output). The networks learn by competing with one another - the generator tries to create clean speech similar to the training data in order to fool the discriminator, which in turn improves in distinguishing between the fake clean speech and the true clean speech.

The DNN-based method as proposed in [5] utilises a DNN to map from noisy to clean speech features in two stages. In the first stage, log-power spectral features are used to train a DNN-based regression model using pairs of noisy and clean speech data. In the second stage, noisy speech features are processed by the trained DNN model to predict the clean speech features. Additionally, a post-processing step is applied to equalise the global variance of the enhanced speech to that of the training data, reducing the effect of over-smoothing which would muffle the speech. The waveform is then reconstructed frame-by-frame from the log-power spectral features of the post-processed clean speech.

2.2. Synthetic voice creation

To create the synthetic voices, we used an HMM-based speech synthesis framework [12], utilising speaker adaptation [1]. Firstly, an average voice model was created from 2400 sentences of clean speech from the Mansfield corpus [13] from two male and two female speakers of New Zealand English (600 sentences per speaker). This average voice model was then adapted using 150 sentences (ten minutes) of speech from the Mansfield corpus from a male target speaker, which were first contaminated by adding different types of noise, then processed by speech enhancement techniques. All audio signals used throughout this process had a sampling rate of 16 kHz.

Four noise types - pink noise, babble noise (in a cafeteria), ambient noise (in a busy park) and music (rock music) were added at two SNRs - 0 dB and 10 dB, for a total of eight conditions. The pink noise was generated using MATLAB's "pinknoise" function, music was obtained from the "fma" folder of the MUSAN database [14], and the babble and ambient noise was obtained from the first channel of PCAFETER and NPARK respectively from the DEMAND database [15].

The four speech enhancement techniques: Subspace (SS) [9], Wiener filter (WF) [10], SEGAN [11] and the DNN-based approach (DNN) [5] were evaluated in this study, as well as no speech enhancement. Both the subspace and Wiener filter used Loizou's MATLAB implementations (klt.m and wiener_as.m respectively) [16], and both the DNN-based method and SEGAN used pre-trained models provided by their respective authors [5, 11].

The sets of speech enhanced sentences were then used to create a synthetic voice through voice adaptation. A total of 41 synthetic voices were created - one voice for each combination of noise type (four), SNR (two) and speech enhancement method (four), a reference voice using the clean 150 sentences set which would act as the high quality baseline, and eight voices using the noisy speech without any speech enhancement, one voice for each noise/SNR combination.

2.3. Subjective evaluation

A MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [17] test was created to evaluate the quality of the

synthetic voices, utilising the webMUSHRA framework [18]. This method of testing has shown to be suitable for comparing speech synthesis systems [19–21] as it allows for multiple comparisons of stimuli, which in turn reduces the impact of the participant and utterance bias [22]. The test consisted of four blocks, each containing four sets of MUSHRA panels. For each set in the test, participants were asked to first listen to a sentence synthesised using the reference voice, then judge the quality of each of sentences synthesised using the voices made with noisy or enhanced speech, as well as a hidden reference voice on a scale of 0 (worst quality) to 100 (best quality) in comparison to the quality of the reference, as per the MUSHRA ITU-R recommendations [17]. For each question, the participants were required to assign at least one of the stimuli the maximum score of 100. A lower anchor was not included for this test as it was difficult to determine which voices could be considered the lowest quality, similar to the studies in [19, 20].

Each set evaluated all the speech enhancement conditions (subspace, Wiener filter, SEGAN, DNN, and no enhancement, as well as the hidden reference, for a specific noise type/SNR level. Each block only consisted of synthetic voices for a specific noise type - music, pink noise, babble or ambient. Two of the sets for each block contained voices which were made with noise added to the speech at SNR 0 dB, and the other two at SNR 10 dB. The order of the four sets within the block were randomised for each participant, however the order of the blocks were fixed in the order of music, pink noise, babble and ambient. Also included at the start of each block was one training question for the specific noise type of the block.

The sentence generated from each of the synthetic voices in each set was the same. Sentences were obtained from the IEEE Harvard sentences list [23] and were synthesised by the synthetic voices under test. To ensure each sentence was evaluated for each noise condition, four separate tests were created. The sentences used for each block in the test was used for another block in another test. For example, sentences used in block A of test 1 was used in block B of test 2 and so on. As there were four tests, containing four blocks with four sets each (training set excluded as it was not evaluated), and each set evaluated four speech enhancement methods plus no enhancement, a total of 320 utterances were evaluated, of which there were 16 unique sentences. For each unique sentence, one additional utterance was synthesised for the reference/hidden reference.

In total, 32 English speaking individuals (21 male, 11 female) participated in the test. Each participant was assigned evenly to one of the four tests. However, three of the results (2 male, 1 female) were discarded as they were consistently unable to determine the hidden reference. As per the MUSHRA ITU-R recommendations, participants results were excluded if they were unable to assign the hidden reference above a score of 90 more than 15% of the time [17].

All participants self-reported to have normal hearing and were using headphones. Due to New Zealand government restrictions in response to COVID-19 at the time the experiment was held, all tests were run online. The tests ran for an average duration of 39 minutes.

3. Results and Discussion

Linear mixed-effects models (LME) were utilised for the analysis using R through the lmerTest [24] and lme4 [25] packages, with the fixed effect being the enhancement type and the random effect being the participant number. In total, eight models, one model for each noise and SNR combination, were anal-

Table 1: p -values for the post-hoc pairwise comparison of speech enhancement methods for each noise type at SNR 0 dB (upper off-diagonal) and SNR 10 dB (lower off-diagonal). SS - subspace, WF - Wiener filter, SE - SEGAN, DNN - DNN-based method, None - no enhancement. Boldface signifies significant pairs, with green squares showing the enhancement method on the left hand side significantly outperforms the enhancement method on the top side with respect to their means, red squares are vice-versa. The bottom two rows for each table show the chi-squared, degrees of freedom and p -value for each LME.

	Ref	SS	WF	SE	DNN	None
Ref		<0.01	<0.01	<0.01	<0.01	<0.01
SS	<0.01		<0.01	<0.01	0.97	<0.01
WF	<0.01	<0.01		0.97	<0.01	<0.01
SE	<0.01	<0.01	0.05		<0.01	0.05
DNN	<0.01	0.25	0.50	<0.01		<0.01
None	<0.01	<0.01	0.17	0.99	<0.01	
0 dB	$\chi^2 = 573.41, df = 5, p = < 0.0001$					
10 dB	$\chi^2 = 141.99, df = 5, p = < 0.0001$					

(a) Music, SNR 0 dB (top-right), SNR 10 dB (bottom-left)

	Ref	SS	WF	SE	DNN	None
Ref		<0.01	<0.01	<0.01	<0.01	<0.01
SS	<0.01		<0.01	<0.01	<0.01	<0.01
WF	<0.01	<0.01		<0.01	0.21	<0.01
SE	<0.01	<0.01	<0.01		<0.01	<0.01
DNN	<0.01	<0.01	0.84	0.03		0.46
None	<0.01	<0.01	<0.01	0.99	0.12	
0 dB	$\chi^2 = 488.97, df = 5, p = < 0.0001$					
10 dB	$\chi^2 = 219.81, df = 5, p = < 0.0001$					

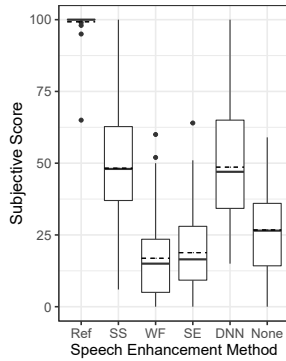
(b) Pink Noise, SNR 0 dB (top-right), SNR 10 dB (bottom-left)

	Ref	SS	WF	SE	DNN	None
Ref		<0.01	<0.01	<0.01	<0.01	<0.01
SS	<0.01		0.30	0.46	0.37	0.49
WF	<0.01	<0.01		<0.01	<0.01	<0.01
SE	<0.01	<0.01	0.08		1.00	1.00
DNN	<0.01	0.16	0.07	0.65		1.00
None	<0.01	0.13	0.08	0.71	1.00	
0 dB	$\chi^2 = 358.77, df = 5, p = < 0.0001$					
10 dB	$\chi^2 = 91.15, df = 5, p = < 0.0001$					

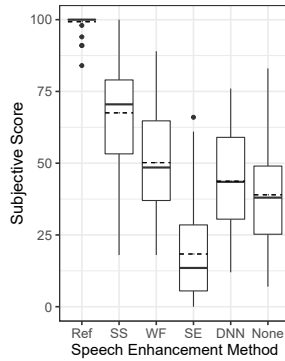
(c) Babble, SNR 0 dB (top-right), SNR 10 dB (bottom-left)

	Ref	SS	WF	SE	DNN	None
Ref		<0.01	<0.01	<0.01	<0.01	<0.01
SS	0.20		<0.01	0.04	0.97	<0.01
WF	0.17	1.00		<0.01	<0.01	0.60
SE	<0.01	<0.01	<0.01		0.24	0.07
DNN	<0.01	0.54	0.60	<0.01		<0.01
None	<0.01	<0.01	<0.01	<0.01	0.30	
0 dB	$\chi^2 = 211.99, df = 5, p = < 0.0001$					
10 dB	$\chi^2 = 152.33, df = 5, p = < 0.0001$					

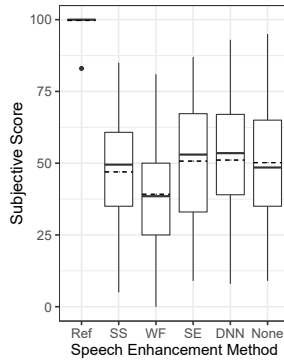
(d) Ambient, SNR 0 dB (top-right), SNR 10 dB (bottom-left)



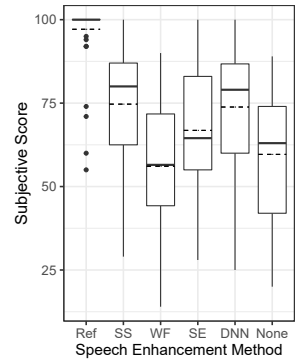
(a) Music, SNR 0 dB



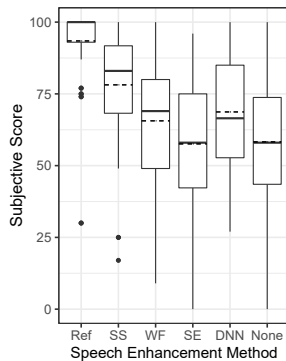
(b) Pink Noise, SNR 0 dB



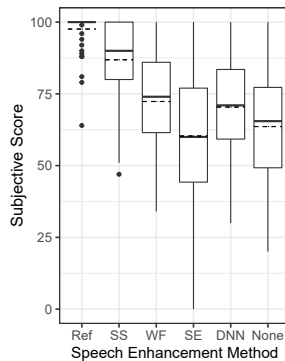
(c) Babble, SNR 0 dB



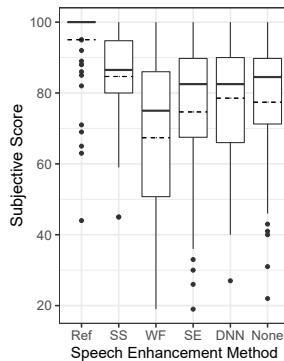
(d) Ambient, SNR 0 dB



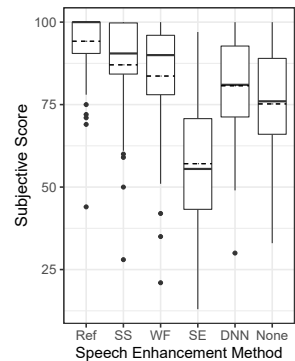
(e) Music, SNR 10 dB



(f) Pink Noise, SNR 10 dB



(g) Babble, SNR 10 dB



(h) Ambient, SNR 10 dB

Figure 1: Comparison of subjective scores for a male synthesised voice created from noisy speech after processing by different speech enhancement methods.

ysed individually, as we are comparing how each enhancement type performs against the other enhancement types under the same noise conditions. To determine if there is a significant difference between speech enhancement types for each noise and SNR combination, post-hoc pairwise comparisons of the LME models were used with the Tukey method for adjustment of the p -values, through the emmeans package [26]. p -values < 0.01 were considered significant.

Table 1 shows the p -values obtained from the post-hoc analysis for each of the noise/SNR conditions, where the values in the upper off-diagonal show the p -values for SNR 0 dB, and the bottom off-diagonal show the p -values for SNR 10 dB. Figure 1 shows boxplots of the 29 participants' responses for each noise/SNR combination, where the solid and dashed lines represent the medians and means respectively. A high subjective score represents closeness in terms of subjective quality between a specific enhancement method and the best voice in each set, which in the majority of cases was the hidden reference voice (denoted by "ref"). As mentioned in Section 2.3, the hidden reference was utilised as a post-screening method to ensure a participant has correctly done the test.

At SNR 0 dB, both SS and DNN obtained the highest mean scores for music and ambient, and for pink noise, SS alone obtained the highest mean score. For babble, all the enhancement methods performed similarly to no enhancement with the exception of WF, which performed worse. WF performed the worst of the four enhancement methods for all noise conditions except for pink noise, performing worse than no enhancement for music, pink noise and babble. SEGAN performed poorly for the music and pink noise conditions, scoring lower than no enhancement, but was comparable to enhancement for babble and ambient noise. For all noise types, none of the speech enhancement methods were comparable to the reference.

At SNR 10 dB, SS consistently scored the highest for all noise types. Interestingly, WF performed much better at SNR 10 dB as opposed to SNR 0 dB, posting comparable results to SS for ambient noise. DNN obtained similar results to SS for music, ambient and babble. Additionally, both SS and WF performed comparably to the reference for ambient noise. In all cases SEGAN performed lower than no enhancement, though there were no significant differences between SEGAN and no enhancement for any of the noise types except ambient noise.

Overall, these results illustrate that the SS method has shown to be the best of the four speech enhancement methods to enhance speech for speech synthesis, with DNN being a close second. Several reasons could explain the results. It is known that when creating speech synthesisers through voice adaptation, there is some inherent noise reduction effects in the final synthesised voice [2]. Although both the WF and SS methods result in musical noise artefacts after enhancement, the artefact from WF is more prominent than from SS, whilst the speech distortion is relatively the same. As a result, especially at lower SNRs, the effect of the musical noise is still present in the final synthesised voice for WF, whereas for SS much of the artefact has been removed. However, SEGAN and the DNN method each produce different kinds of artefacts after speech enhancement. Even though SEGAN was able to remove much of the background noise, it tended to over-filter the speech signal in some samples, in particular at SNR 0 dB, resulting in choppy speech. This is illustrated in Figure 2, where we can see a noisy natural speech file after processing by SEGAN (d) has varying degrees of attenuation of the signal, unlike SS (c) which follows the waveform of the clean speech (a). This issue may have propagated when the signals were used to train the synthetic voice,

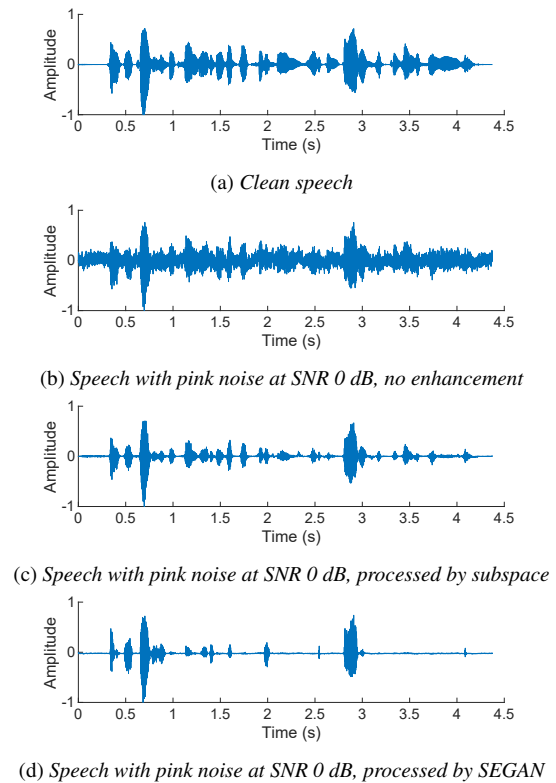


Figure 2: Waveforms of natural speech before and after speech enhancement. (a) shows the reference (clean) speech, (b) shows the noisy speech without enhancement, (c) - (d) show the noisy speech after processing by SS and SEGAN respectively.

resulting in large volume variations in the synthesised speech. It should be noted that although both models had been trained at many SNR levels, including 0 dB and 10 dB, the pretrained model for SEGAN was trained on only ten noise types from the DEMAND database [11], whereas the pretrained model for the DNN method was trained on over 115 noise types [5]; the lack of training material used for SEGAN may have reduced its ability to generalise to foreign noise/SNR conditions, which may have contributed to the poor results by SEGAN overall.

4. Conclusions and Future Work

This study presented an evaluation of two traditional (Wiener filter and subspace) and two modern (SEGAN and a DNN-based approach) speech enhancement methods in de-noising speech used for speech synthesis. Synthetic voices were created through HMM speaker adaptation utilising sets of 150 sentences, where each set was a combination of different noise types, SNRs and speech enhancement methods. The results of a subjective test shows that utilising the subspace method resulted in higher quality synthetic speech, comparable to voices made with clean speech at higher SNRs, and both subspace and the DNN-based methods were able to create higher quality voices when compared to no enhancement at lower SNRs, but did not score as highly as voices made with clean speech. For future work, we intend to extend this study for female synthetic speech, and investigate the effect of speech enhancement on neural network-based speech.

5. References

- [1] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.
- [2] 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, 2013.
- [3] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.
- [4] 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.
- [5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.
- [6] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.
- [7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.
- [8] F. Del Prado, Y. Hioka, and C. Watson, "The effect of speech enhancement in voice adaptation when building synthetic voices," *Acoustical Science and Technology*, vol. 39, no. 2, pp. 150–153, 2018.
- [9] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE transactions on speech and audio processing*, vol. 11, no. 4, pp. 334–341, 2003.
- [10] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.
- [11] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.
- [12] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.
- [13] C. I. Watson and A. Marchi, "Resources created for building New Zealand English voices," in *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, 2014, pp. 92–95.
- [14] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [15] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.
- [16] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.
- [17] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.
- [18] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.
- [19] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [20] T. Merritt, B. Putrycz, A. Nadolski, T. Ye, D. Korzekwa, W. Dolecki, T. Drugman, V. Klimkov, A. Moinet, A. Breen *et al.*, "Comprehensive evaluation of statistical speech waveform synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 325–331.
- [21] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," *arXiv preprint arXiv:1902.09574*, 2019.
- [22] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Interspeech*, 2017, pp. 3976–3980.
- [23] E. Rothausser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.
- [24] A. Kuznetsova, P. B. Brockhoff, R. H. Christensen *et al.*, "lmerTest package: tests in linear mixed effects models," *Journal of statistical software*, vol. 82, no. 13, pp. 1–26, 2017.
- [25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
- [26] R. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2020, r package version 1.5.0. [Online]. Available: <https://CRAN.R-project.org/package=emmeans>