Improving the intelligibility of synthetic speech over the telephone and in noise

an evaluation of near end listening enhancement algorithms

Jay Y. Y. Yeung

The University of Edinburgh

School of Philosophy, Psychology and Language Sciences

Master of Arts with Honours Linguistics

Supervisor: Professor Simon King

May 2020

8,139 words

Acknowledgements

First, I would like to thank my supervisor Professor Simon King for opening my eyes to the world of speech processing and speech synthesis and for his guidance and direction throughout my studies and investigation.

I also want to thank Carol Chermaz, who not only shared with me her wonderful creation – the ASE algorithm – but also her vast knowledge of NELE and sound engineering, from theory to obscure MATLAB functions.

I thank my friends Abby Riach and Erik Eveland for their loving support and encouragement.

Last but not least, I thank my mother Carina Lau for her love, her wisdom, and her unwavering belief in me.

Table of contents

1. Introduction

As the naturalness and intelligibility of text-to-speech technology (TTS) have gradually improved in recent years (King, 2014), the technology has begun to see a wider range of use over the telephone, for example in interactive voice response systems used by banks and other businesses, as well as in more advanced artificial intelligent (AI) applications such as Google Duplex, where an AI agent automatically makes restaurant and hairdressing appointments for its user over the telephone, using a TTS synthetic voice to communicate with human employees at the restaurant or hair salon (Leviathan & Matias, 2018).

In such commercial applications, the intelligibility of the TTS system is a crucial requirement. But despite advances in TTS technology, its intelligibility over the telephone line is still made difficult by three factors: First, the telephone as a channel reduces intelligibility, as speech signals may be band limited (International Telecommunication Union, 1988) and may also be degraded by speech coding and transmission errors (Friedman-berg, Allendoefer, & Deshmukh, 2009). Secondly, even though the synthetic speech itself should carry minimal noise since it is based on professionally studio recorded speech samples or generated by a vocoder, there is likely to be acoustical background noise on the listener's end, also referred to as the near end, that can reduce intelligibility (Morimoto, Sato, & Kobayashi, 2004). Finally, background noise, while already detrimental to natural speech intelligibility, has been shown to affect synthetic speech intelligibility to an even greater degree (King & Karaiskos, 2010).

One promising solution to this intelligibility problem is a type of algorithms called near end listening enhancement (NELE) algorithms that has been shown to improve the intelligibility of synthetic speech in noise (Valentini-Bontinhao et al., 2013), particularly a subset of these algorithms that are both usable automatically in real-time and do not require any knowledge of the noise signal (e.g. Zorila, Kandia, & Stylianou, 2012). However, these algorithms have not been designed for use over a narrowband telephone transmission and their effectiveness over the telephone is yet unknown. Additionally, it is unclear whether there are any negative interactions between the NELE algorithms, synthetic speech, and telephone transmission that are not present for natural speech.

The aim of this dissertation is to design a complete and realistic testing platform that simulates this specific listening scenario and using this platform, conduct a number of formal listening tests in order to evaluate the effectiveness of NELE algorithms in improving the intelligibility of synthetic speech that is subsequently transmitted via telephone and presented in near end background noise.

2. Literature Review

This literature review is structured into three parts. First, I will review the three intelligibility impediments—telephone degradation, background noise, and the interaction between synthetic speech and noise—that the use of synthetic speech over the telephone faces. Secondly, I will then review past research on speech enhancement techniques and their effectiveness on counteracting some of the intelligibility impediments explored. Finally, I will situate this dissertation and the motivations of this dissertation in the context of past research.

2.1 Impediments to speech intelligibility

2.1.1 Telephone audio encoding

Telephone transmission has been known to negatively impact intelligibility. For speech to transmit over the telephone, the analog speech signal must first be converted to a digital signal in a process known as audio encoding and then from a digital signal back to an analog signal for playback in a process known as decoding. The telecommunication industry employs a number of codecs—standard for this encoding and decoding process—which have been shown to decrease intelligibility in two ways. First, most audio codecs only pass audio signals in the range of 300-3400 Hz, discarding the rest of the frequencies. Secondly, some audio codecs employ lossy compression to reduce the bitrate—the amount of data being transmitted—further degrading its intelligibility.

First, I will review the detrimental effects of a limited bandwidth on speech intelligibility. Most audio codecs commonly used by telecommunication networks, including G.711, G.726, and G.729, are narrowband codecs that only pass audio signals in the range of 300-3400 Hz (International Telecommunication Union, 1988). However, research has shown that the human audible frequency range reaches up to 15 kHz for most adults and up to 20 kHz for children and younger adults, and certain speech sounds do in fact show significant spectral energy above the 3400 Hz upper limit of narrowband audio codecs (Monson et al., 2014). Hughes and Halle (1956) recorded two male and one female English speakers pronouncing voiced and voiceless fricatives with equipment that captured frequencies of up to 10 kHz and their results showed that the spectral peaks for certain fricatives were located

in frequencies above 3.4 kHz. In particular, the spectral peaks for alveolar fricatives /s/ and /z/ were located between 3 and 8 kHz, and those for labio-dental fricatives /f/ and /v/ were located between 8 and 9 kHz. A later experiment by Jongman et al. (2000), performed with 20 subjects, corroborated the results of Hughes and Halles and found spectral peak location to significantly distinguish place of articulation of fricatives. Research seems to suggest that the exclusion of higher frequencies, and consequently the removal of the spectral peaks of certain consonants, does impact intelligibility. Moore et al. (2010) found in formal listening tests that there was a statistically significant difference in intelligibility for normal hearing listeners when speech is low-pass filtered at different cut-off points, with speech low-pass filtered at 5 kHz being less intelligible than speech low-pass filtered at 7.5 kHz. Another research performed with children instead of adults found that normal hearing children required three times as much exposure to novel words to learn them when the words were low-pass filtered at 4 kHz compared to when the words were low-pass filtered at 9 kHz (Pittman, 2008). Therefore, by discarding frequencies outside of the range of 300-3400 Hz, narrowband telephone codecs are discarding information used to distinguish certain fricatives and consequently reducing intelligibility.

Next, I will discuss how data compression performed by audio codecs can reduce speech intelligibility over the telephone. In order for more calls to be transmitted under the same network capacity, audio codecs such as G.726 perform data compression to the speech signal to reduce its bitrate. For example, the G.726 codec employs adaptive differential pulse-code modulation (ADPCM), which varies the quantization step size—the definition or level of detail of the amplitude in a digital signal—to reduce bitrate (International Telecommunication Union, 1990). While ADPCM is able to significantly reduce the bitrate of a speech signal from 64 kb/s (as in G.711) to 32, 24, or 16 kb/s, it has also been found to reduce its intelligibility. In Friedman-berg et al. (2009), the investigators tested the intelligibility of different speech codecs using the modified rhyme test. Results (n=24) showed that speech encoded under the ADPCM-compressed, 16 kb/s G.726 codec had a small but significant reduction in intelligibility compared to speech encoded under the 64 kb/s G.711 codec, with word accuracy rate (WAR) dropping from 95% to 90%.

2.1.2 Background noise

Perhaps unsurprisingly, acoustical background noise has been found to affect the intelligibility of speech. While evident in our personal experience conversing in noisy

environments, this phenomenon has also been well documented in the literature through formal listening tests, in which subjects listen to stimuli under different signal-to-noise ratios (SNR), where the lower the SNR, the noisier the speech, and are then asked to reproduce the stimuli. For example, in Morimoto, Sato, & Kobayashi (2004), the investigators studied the effects of noise and reverberation on speech intelligibility by performing a series of listening tests with five SNRs and two reverberation times. Results (n=13) showed that when the stimuli are comparatively less noisy, at SNRs of 15 dB and above, WAR was almost 100%. But this percentage dropped to roughly 80% and 60% when the background noise became more significant at SNRs of 0 and -5 dB, respectively. Results showed that intelligibility was affected by noise level but not by reverberation, at the noise and reverberation levels tested. Similar results were obtained by George, Goverts, Festen, & Houtgast (2010) in a similar study.

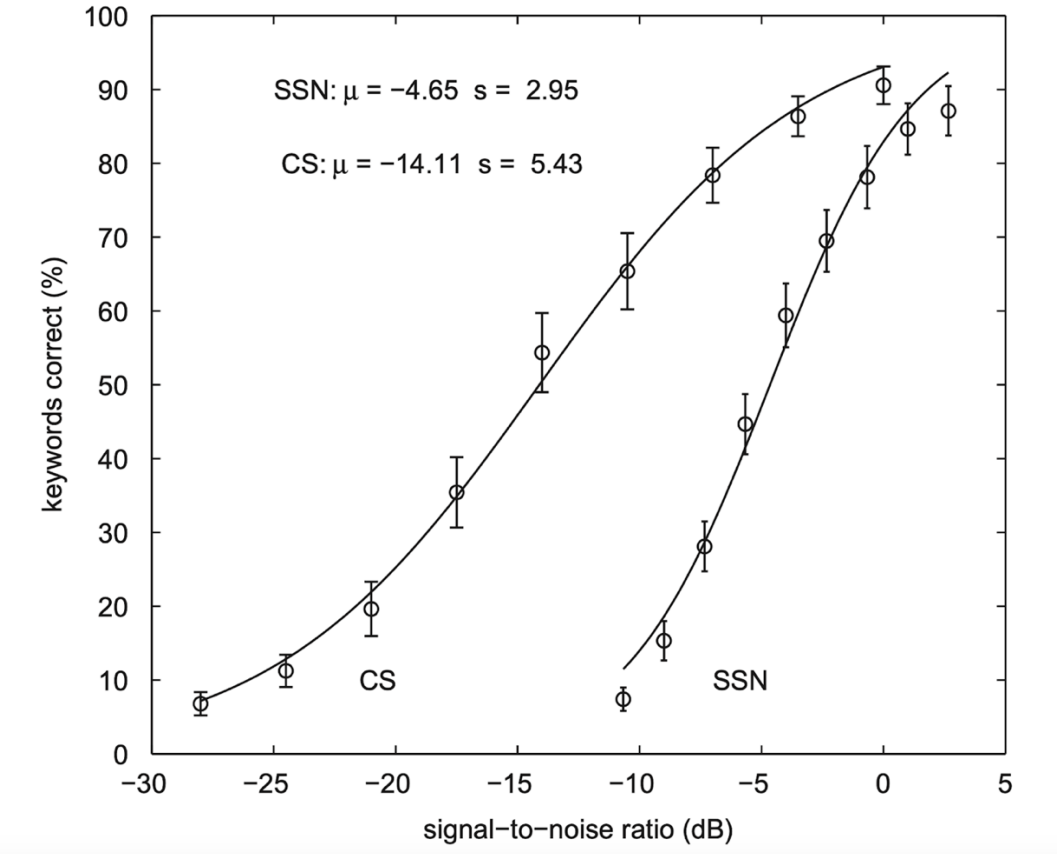

Fig. 2.1: Formal listening test results of intelligibility in 2 noise types, reproduced from Cooke et al. (2013). Listeners' mean WAR (open circles) as a function of SNR for competing speech noise (CS) and speech shaped noise (SSN).

Another interesting property of noise's effect on intelligibility is that this effect is non-linear, such that given a number of SNRs at fixed intervals (e.g. -10, -5, 0, 5, 10 dB), the

difference in intelligibility between the SNRs would not be equal. Rather, the intelligibility of speech in different SNRs can be fitted onto a psychometric function, where change in intelligibility is more acute near 50% WAR and less so near the two extremes. This has been shown in evaluation studies such as Cooke, Mayo, & Valentini-Bontinhao (2013) and Chermaz, Valentini-Bontinhao, Schepker, & King (2019), a figure from the former has been reproduced in Figure 2.1 to illustrate this phenomenon.

2.1.3 Synthetic speech in noise

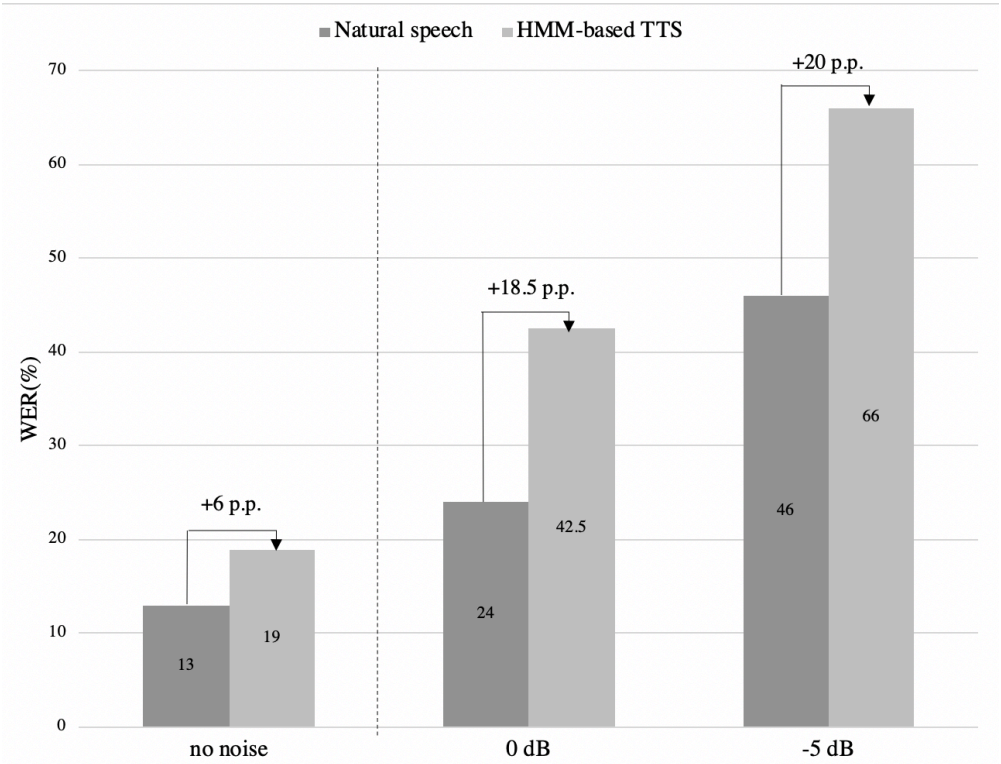

Fig. 2.2: Word error rates of natural speech and HMM-based synthetic speech in 3 noise conditions (SNR in dB) Difference in intelligibility shown in percentage points. Adapted from King & Karaiskos (2010).

Historically, synthetic speech has been less intelligible than natural human speech (Green, Logan, & Pisoni, 1986; Bennett, 2005), but this pattern began to change in the late 2000s, when a hidden-Markov-model-based (HMM) speech synthesis system achieved a WER statistically comparable to that of natural speech (Yamagishi et al., 2008) and in the 2010s with the advent of deep-neural-network-based systems (Zen, Senior, & Schuster, 2013), which now routinely achieves similar intelligibility to natural speech.

However, even though recent synthetic speech systems may be as intelligible as natural speech in quiet, evaluation studies have shown that in the presence of background

noise, synthetic speech suffers a bigger reduction in intelligibility compared to natural speech. In King and Karaiskos (2010), intelligibility of natural and synthetic speech was measured in multiple noise conditions: without additive noise, with additive noise at a signal-to-noise ratio of 0 dB, and at SNR of -5 dB. Results showed that in quiet, intelligibility was close between natural and synthetic speech, but the gap widened significantly at SNRs of 0 and -5 dB. For instance, without additive noise, the benchmark hidden-Markov-model (HMM) statistical parametric TTS system had a WER of 19% compared to 13% of natural speech, representing a 6 percentage point difference between the two. At an SNR of 0 dB, the gap widened reaching a 18.5 percentage point difference and at SNR of -5 dB, a 20 percentage point difference. This shows that while some modern synthetic speech systems may not hinder intelligibility in quiet, synthetic speech as a whole still presents a challenge to intelligibility due to its negative interaction with background noise as well as in cases where an older TTS system is used, for example with less popularly spoken, under-researched languages.

2.2 Near end listening enhancement

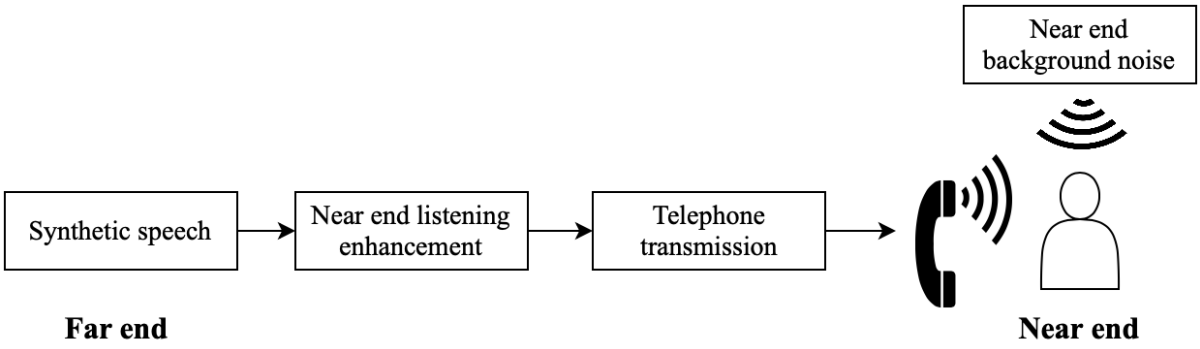

Fig. 2.3: A diagram illustrating the near end listening enhancement problem for our specific scenario of synthetic speech over the telephone. Far end refers to the original signal – the synthetic speech, while near end refers to the receiver or listener's end

Given that the three intelligibility impediments discussed in the previous section are inherent to the use case of TTS over the telephone, the question is therefore whether we can enhance the synthetic speech signal prior to its transmission over the telephone and playback in background noise, to counteract some of the effects of these impediments, making the speech more intelligible compared to an unenhanced baseline (Figure 2.3). This problem, known as the near end listening enhancement (NELE) problem, has seen research interest in recent years as it has broad implications in many speech production and playback scenarios

such as in public announcements in train stations and airports, in telephone conversations, or in radio communications.

2.2.1 NELE for near end noise

Perhaps the most well-studied intelligibility impediment for NELE is near end noise, partly due to how near end background noise is present in most speech production and playback scenarios. Due to the popularity of this research topic, many different NELE approaches with varying techniques and constraints have been proposed and tested over the years. In the following paragraphs, I will give a brief overview of these approaches.

One NELE approach for near end noise explored by phoneticians involved manually segmenting and labelling speech, then selectively increasing the relative intensity and/or duration of perceptually important sounds that are of low intensity (e.g. fricatives) and/or brief in duration (e.g. plosive release). Gordon-Salant (1986) studied increasing consonant duration and consonant-vowel intensity ratio and found in formal listening tests using nonsense consonant-vowel (CV) pairs that increasing the consonant-vowel intensity ratio alone was able to increase the syllable's intelligibility. Hazan and Simpson (1996) took a more fine-grained approach and differentiated between different types of consonants: plosives, fricatives, affricates, approximants, and nasals. For each type of consonant, they performed a specific set of amplification procedures, with different gains in intensity for different types of consonants and for different regions within a consonant. Results (n=13) from formal listening tests showed that this approach was effective in increasing the intelligibility of nonsense VCV syllables in speech shaped noise at SNRs of -5 and 0 dB. However, further experiments (n=12) that tested its effectiveness in a sentential setting using semantically unpredictable sentences instead of isolated nonsense VCV syllables showed mixed results, with plain unmodified speech scoring higher than modified speech at an SNR of 0 dB and vice versa at an SNR of 5 dB. For our particular use case, this approach presents two problems. First, its effectiveness in full sentences is yet unclear but more importantly, the manual procedures required of segmentation, labelling, and selective enhancement, while perhaps applicable for pre-recorded messages like public announcements or deterministic IVR utterances, would not be applicable for real-time TTS production.

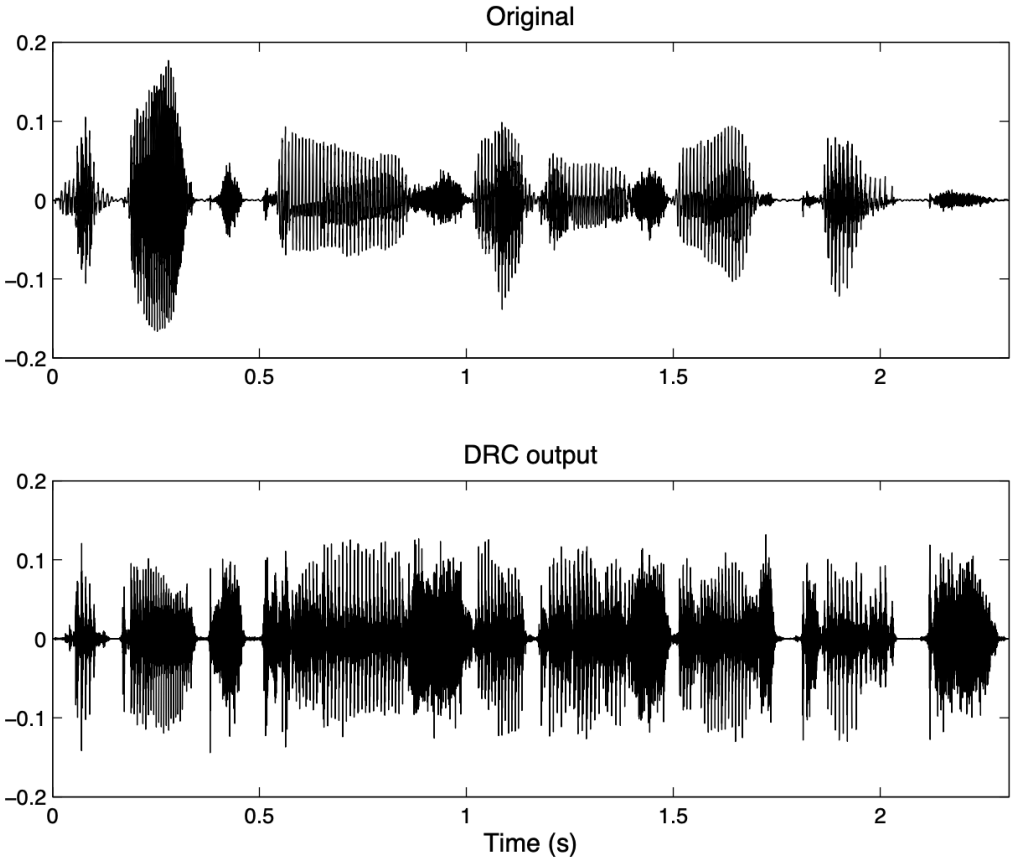

Fig. 2.4: The effects of compression – one of the techniques used in NELE to increase intelligibility in noise. Reproduced from Zorila, Kandia, & Stylianou (2012).

A more algorithmic approach to NELE in noise using signal processing techniques has also been explored by researchers. Niederjohn and Grotelueschen (1976) proposed one of the first NELE algorithms, which first applied a high-pass filter then a rapid amplitude compression to the speech signal. In formal listening tests using monosyllabic English words, at five different SNRs (-10, -5, 0, 5, and 10 dB), results (n=6) showed that this approach was able to increase WAR at all five SNRs tested. In particular, at SNRs of -5 and 0 dB, the algorithm was able to dramatically boost WAR from 30% to 85% and from 35% to 90% respectively.

More recent works motivated by Niederjohn and Grotelueschen's approach of reallocating energy in the frequency domain (high-pass filter) and the time domain (compression) include the spectral shaping and dynamic range compression (SSDRC) algorithm by Zorila, Kandia, & Stylianou (2012). In the frequency domain portion of the algorithm, SSDRC differs from Niederjohn and Grotelueschen's algorithm in two ways: first,

instead of one high pass filter, SSDRC first performs formants enhancement, then two pre-emphasis filters boosting energy above 1000 Hz, with the latter also reducing the energy of frequencies below 500 Hz. Secondly, instead of indiscriminately applying a high pass filter to all portions of the signal, SSDRC only performs the formants enhancement and the first filter if the frame is estimated to contain voicing, thereby preventing the introduction of artefacts from performing these operations to unvoiced speech. Next, in the time domain portion of the algorithm, dynamic range compression is performed, which similarly to Niederjohn and Grotelueschen's rapid amplitude compression, is able to reduce the dynamic range of the speech signal, making speech sounds previously low in amplitude (typically consonants) more prominent and vice versa, effectively increasing the consonant-vowel intensity ratio, similarly to the phonetic approaches explored above. In intelligibility tests using phonemically balanced Harvard sentences (e.g. "Oak is strong and also gives shade"), SSDRC is able to increase intelligibility across noise types and across SNRs. Compared to the manual approach above, SSDRC has been shown to be effective when applied to sentences and is also algorithmic, thus applicable to automatic real-time enhancement for TTS.

2.2.3 NELE for synthetic speech in noise

Given the success of NELE algorithms in improving intelligibility in noise, researchers have attempted to explore whether these approaches designed for natural speech would also be effective for synthetic speech. Valentini-Botinhao et al. (2013) tested the effectiveness of 6 NELE algorithms when applied to an HMM-based synthetic speech system, in three SNR conditions and two noise types. The list of NELE algorithms tested included the noise-independent SSDRC, a number of noise-dependent algorithms as well as algorithms combining the noise-dependent algorithms with SSDRC. Results (n=88) showed that SSDRC as well as algorithms combined with SSDRC were effective in increasing the intelligibility of the HMM-based system in all noise conditions. Results for the noise-dependent algorithms were mixed, with all systems able to improve intelligibility in stationary, speech-shaped noise while some systems performed equally to or worse than the unmodified baseline in competing speaker noise. Across all NELE algorithms, SSDRC was shown to be the most effective, across all noise conditions. This study showed that NELE algorithms, in particular SSDRC, were effective in improving the intelligibility of synthetic speech in noise.

2.2.2 NELE for telephone transmission and noise

While NELE algorithms for noise have been shown to be effective for both natural and synthetic speech, their effectiveness when transmitted through a narrowband telephone codec is still unclear. Examining SSDRC as an example, one of its pre-emphasis filters boosts all frequencies above 1 kHz, which would include frequencies above 3.4 kHz. Additionally, when broadband compression is performed, the intensity of low-intensity portions of the speech signal is boosted. As many consonants such as fricatives [s] and [f] have energy between 4-8 kHz and are often low in intensity compared to vowels, energies above 3.4 kHz would therefore be boosted. However, if the SSDRC enhanced speech is transmitted via narrowband telephone encoding that only passes frequencies between 300-3400 Hz, all the enhancement done to frequencies above 3.4 kHz would be completely discarded.

It is unclear whether the limited frequency range of narrowband telephone transmission would be detrimental to the effectiveness of SSDRC and the other NELE methods described above. The formal listening tests previously conducted to evaluate NELE performance have used stimuli with a much higher frequency range than the one used by narrowband telephone: the Hurricane Challenge presented stimuli at a sampling rate of 16 kHz, thus a Nyquist frequency of 8 kHz (Cooke et al., 2013) while Chermaz et al. (2019) presented stimuli at a sampling rate of 48 kHz. Further, past research on NELE specifically for narrowband telephone transmission, primarily by Jokinen and colleagues (Jokinen & Alku, 2017; Jokinen, Remes, & Alku, 2017) focused on enhancing the speech signal after telephone transmission and the accompanying degradations have already taken place, instead of enhancing the speech signal prior to transmission. As the aim of this dissertation is to explore ways to enhance synthetic speech prior to transmission and playback, Jokinen and colleagues' results are therefore not applicable.

2.3 Motivation for this study

As shown in the previous sections, the use of TTS over the telephone suffers from three impediments to intelligibility: telephone transmission, background noise, and the interaction between synthetic speech and noise. The effects of these impediments have been explored, but it is unclear how all three of these impediments interact. Similarly, I've reviewed how NELE algorithms have been shown to improve the intelligibility of speech, both natural and synthetic, in noise. Yet, it is unclear whether NELE still remains effective

when applied to speech prior to telephone transmission and indeed when all three impediments are present. In this dissertation, I will devise a complete testing pipeline to simulate this specific listening scenario and perform a series of formal listening tests in order to evaluate the effectiveness of NELE on synthetic speech, when applied prior to telephone transmission and playback in noise.

3. Experimental Design

As established in the previous chapter, the objective of this dissertation is to evaluate the effectiveness of NELE algorithms on synthetic speech, when applied prior to narrowband telephone transmission and playback in noise, by conducting a series of formal listening tests. In this chapter, I will first detail the materials and methods used to recreate this specific listening scenario and explain the design choices made. Then, I will outline the experimental design of the listening tests.

3.1 Materials

3.1.1 Selecting the corpus and dataset

The Blizzard Challenge 2011 dataset (King & Karaiskos, 2011) was chosen to be used in the experiments. The Blizzard Challenge is an annual challenge in speech synthesis between various research groups. The dataset includes two main sets of speech data -- the training set and testing set. The training set only includes human speech data, which was used by the synthetic speech systems as training data. The testing set, on the other hand, contains speech samples of the same sentences produced by a human speaker (same speaker as the training set) as well as synthetic speech systems (which are based on the same human speaker who recorded the training and testing data), allowing for direct comparison to be made between natural and synthetic speech. In particular, 100 sentences from the "news" corpus of the testing set is used. These are sentences, 9 to 10 words in length, selected from the Glasgow Herald newspaper, with either a journalistic, expository prose or a conversational prose, with the latter coming from interview quotes in the news articles.

Police said there were no suspicious circumstances surrounding the death.
The 30-day time limit looks increasingly optimistic and unrealistic.
"I found them to be a really lovely family."
Ellie was an inspiration to her friends and family.
"I felt compromised and in some sort of trap."
This is a most unusual and most distressing case.
The restructuring proposals will effectively block that power play.

Inspector Charles Rankin praised the valiant efforts of his officers.
New evidence has emerged that heavier babies have higher intelligence.
So it's a great pressure to have in some ways.

Table 3.1: 10 sentences from the Scottish news corpus used in this study, from King & Karaiskos (2011).

The natural speech recordings had been made at a sampling rate of 48 kHz and at 16 bits per sample, of a female speaker with a general American accent. For synthetic speech, speech samples were generated by the HTS hidden-Markov-model (HMM) statistical parametric system (Zen et al., 2007). The synthetic speech samples had been generated at the same sampling rate and bit depth as the natural speech.

An HMM statistical parametric system was chosen over a state-of-the-art deep neural network (DNN) statistical parametric system due to the ease of access to the former, as it is available in the Blizzard dataset along with the corresponding natural speech, while still being comparable to DNN systems due to the theoretical similarities between the two, as both generate artificial waveforms based on statistical parametric representations. It is also due to this comparability reason that an HMM system was chosen over a concatenative synthesis system.

We expect the HMM system to be marginally less intelligible than natural speech, based on listening test results from King & Karaiskos (2011), where the HMM system was found to have a word error rate (WER, lower the better) of 20% and 14% in semantically unpredictable sentences and when reading addresses, compared to natural speech WER of 16% and 13% respectively.

3.1.2 Selecting the NELE algorithms

Two NELE algorithms, SSDRC (Zorila, Kandia, & Stylianou, 2012) and ASE (Chermaz, 2020) have been chosen for the experiments. These two algorithms were selected for three main reasons: first, both of these algorithms are considered to be the state of the art in NELE research, with SSDRC scoring amongst the highest in NELE evaluation studies including the first Hurricane Challenge (Cooke et al., 2013) and Chermaz et al. (2019), and ASE scoring amongst the highest in the preliminary results of the second Hurricane Challenge (Cooke, Mayo, & Valentini-Botinhao, 2020). Secondly, they are both noise-

independent. Noise-independent algorithms, unlike their noise-dependent counterparts (e.g. Schepker, Rennies, & Doclo, 2013), do not require any knowledge of the noise signal at the receiver's end, such as its intensity or spectral characteristics. As we are attempting to enhance speech prior to telephone transmission, we would have no knowledge of the near end noise signal, therefore making this noise-independent constraint necessary. Finally, the two algorithms represent two differing approaches to NELE. SSDRC is designed to maximize intelligibility while ASE is designed to achieve a balance between increasing intelligibility and preserving speech quality. This is reflected in the techniques used, with SSDRC taking a broader and simpler signal-processing approach and ASE performing a more fine-grained approach inspired by techniques used by sound engineers and audio producers in the entertainment industry, as will be detailed in the following paragraphs.

SSDRC consists of two parts: energy reallocation in the frequency domain and in the time domain. In the first portion of the algorithm, SSDRC performs formants enhancement, then a pre-emphasis filter boosting energy from 1100 Hz, and finally a second pre-emphasis filter boosting energy between 1000 to 4000 Hz. The formants enhancement and the first filter are only applied if a frame is estimated to contain voicing, thereby preventing the introduction of artefacts from performing these operations on unvoiced speech. Next, in the time domain portion of the algorithm, a fixed broadband compression is performed, reducing the dynamic range of the speech signal, making speech sounds previously low in amplitude (typically consonants) more prominent and vice versa.

In ASE, the signal is divided into six frequency bands, with the frequency range for the bands based on state-of-the-art mixing consoles, and each band is compressed individually using a fully automatic compressor. ASE's approach to compression differs from SSDRC in two main ways. First, compression is applied to each frequency band individually instead of to all frequencies at once. Secondly, the compressor is guided by statistical measurements of the signal instead of using a static set of parameters. With these two features, ASE avoids performing compression if the signal already appears compressed and avoids harmonic distortion from excessive compression. Subsequently, the signal is equalized by analyzing the power in the different frequency bands. One band is taken as a reference and the other bands are scaled in respect of this. As opposed to using predetermined pre-emphasis filters as in SSDRC, this equalization approach avoids excessively boosting certain frequencies if they were already at a comparatively high level. Finally, broadband

compression and limiting are performed, emulating the mastering procedures performed in the entertainment industry.

3.1.3 Simulating telephone transmission

A simulation pipeline was set up to emulate a compressed narrowband telephone channel at 16 kb/s, using the ITU-T G.191 software tools for speech and audio coding standardization (International Telecommunication Union 2019). In this pipeline, the speech signal is first downsampled from 48 kHz to 8 kHz, then its active speech level is set to -26 dBov and filtered according to the telephone bandpass defined in ITU-T recommendation G.712. Subsequently, it is encoded according to the G.711 A-law codec, then encoded and decoded according to the G.726 codec at 16 kb/s, and decoded using the G.711 codec. Finally, the decoded signal is filtered according to ITU-T recommendation P.830 and its active speech level set to -26 dBov.

3.1.4 Simulating near end background noise

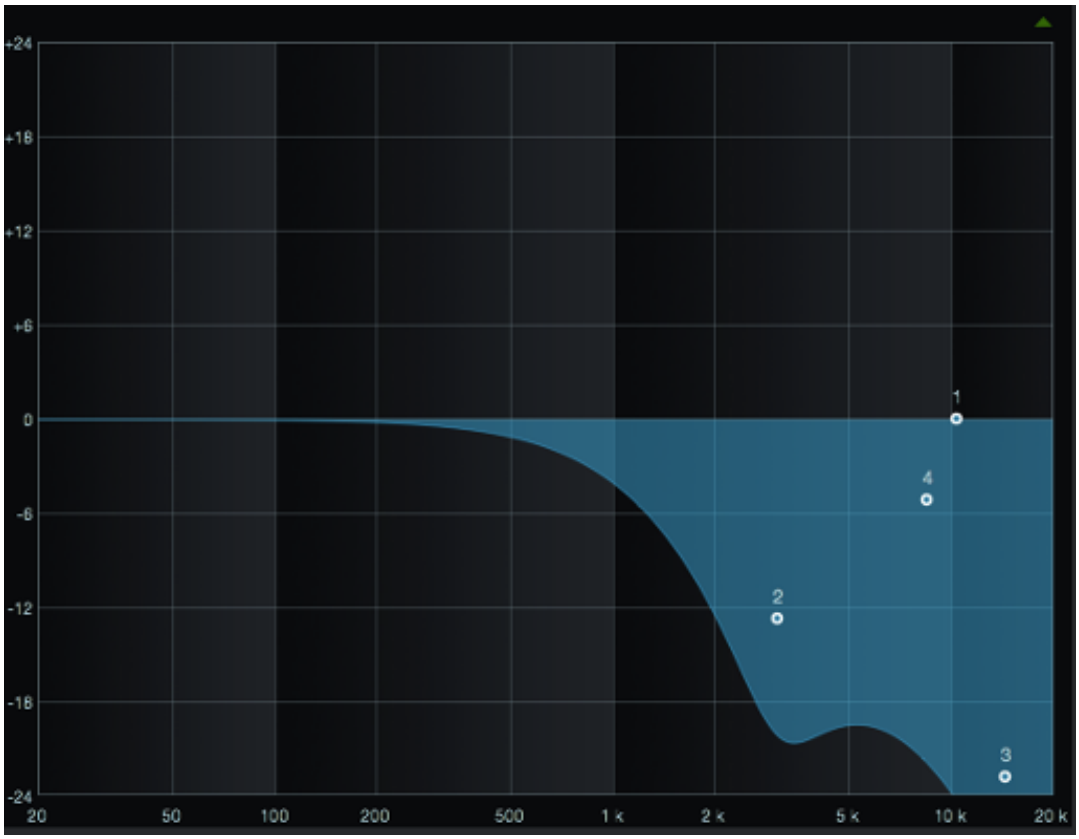

Fig. 3.2: Equalizer emulating the effects of earbuds blocking the ear canal resonance and the higher frequencies. The horizontal axis shows frequencies in logarithmic scale and the vertical axis shows amplitude in dBFS.

For the near end background noise, the cafeteria noise environment recorded by Kayser et al. (2009) was chosen. It has been recorded at 48kHz, at 24 bits per sample, in two channels and is representative of a large public space with relatively steady-state noise – a large number of individuals speaking at the same time in the background. The cafeteria noise recording was chosen because it had been previously used in evaluation studies of NELE algorithms (Chermaz et al, 2019), and is representative of a common use case of TTS telephone calls, for instance when a restaurant employee receives a reservation telephone call from an automated AI system equipped with TTS, such as Google Duplex.

This noise recording was further processed by an equalizer created using a popular digital audio workstation software Cubase to simulate the effects that wearing earbuds has on near-end noise. First, since the earbuds are plugged into the listener's ear canal, they cancel out the ear canal resonance at 2-4kHz. Secondly, the earbuds physically block out higher frequencies. The precise equalizer settings can be found in Figure 3.2.

3.2 Listening Test Design

A series of three listening tests were conducted in total: first, a pre-experiment calibration test was performed to establish the SNRs that should be used for the main experiments, followed by a first experiment using natural speech, NELE, telephone simulation, and noise to establish a baseline, and finally a second experiment performed with the same conditions as the previous experiment, but done using synthetic speech.

3.2.1 Pre-experiment calibration test

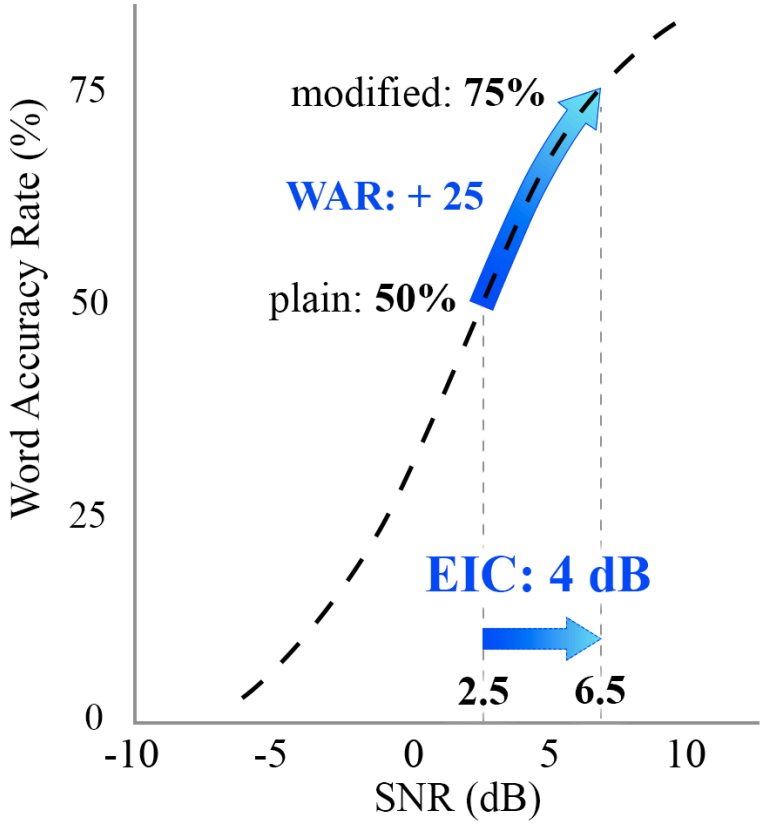

Fig. 3.3: an illustration of equivalent intensity change (EIC) where a speech modification was able to improve WAR from 50% to 75%, at an SNR of 2.5 dB. Given the psychometric curve, we are able to see that without any modifications, it would have required an SNR improvement of 4 dB to achieve the same level of WAR. Therefore we can conclude that the modification brings an equivalent intensity change of 4 dB. Reproduced from Chermaz et al. (2019).

The pre-experiment calibration test had two objectives: first, to determine the SNRs to be used later in the two main experiments. For the two main experiments, we needed to find the three SNRs that would correspond to WARs of 25%, 50%, and 75% for plain, unenhanced natural speech over the telephone. This would allow us to determine the effectiveness of the NELE algorithms at different noise levels. Secondly, the calibration test was needed to find the psychometric curve so that any intelligibility improvements brought on by the NELEs can be expressed in terms of equivalent intensity changes (EIC) in decibels.

As previous NELE research has found, intelligibility in noise lies on a psychometric curve and this curve can be estimated using logistic approximation (Cooke et al. 2013). With a psychometric curve, we would be able to determine the SNR that would correspond to any

given WAR and vice versa, allowing us to determine the three SNRs for the main experiments. This would also allow us to determine the EIC for a given NELE algorithm, defined as the amount of SNR improvement (in dB) that the unmodified speech would have required in order to achieve the same WAR as the modified speech. An illustration of this can be found in Figure 3.3.

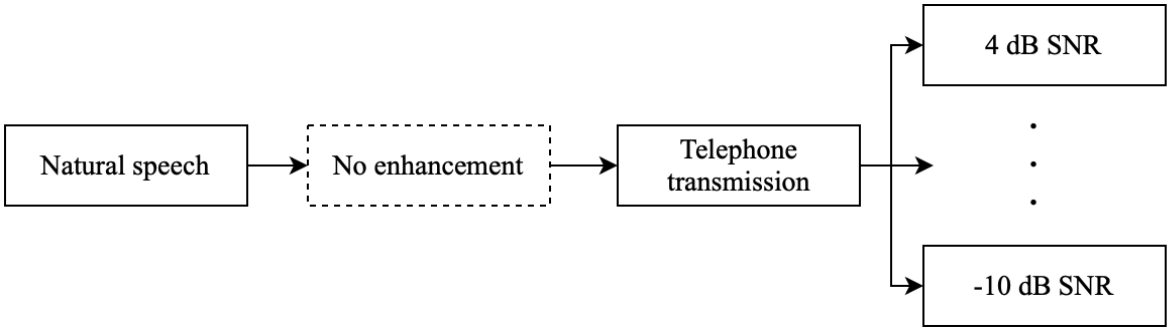

Fig. 3.4: A flowchart illustrating the pipeline for creating the stimuli used in the pre-experiment calibration test

7 native English speakers with self-reported normal hearing were recruited for the calibration test. The participants were all undergraduate students at the University of Edinburgh, recruited online through social media.

A total of 90 sentences were used in the listening test, with 80 sentences being test sentences and 10 being practice sentences. The 80 test sentences were divided into 8 blocks of 10 sentences, with each block corresponding to one of eight SNRs (-10, -8, -6, -4, -2, 0, 2, 4 dB). The 10 practice sentences had at least one of each SNRs. The 90 stimuli were taken from the natural speech news corpus of the Blizzard Challenge 2011, then processed with the telephone transmission simulation, and mixed with the noise signal. The speech signal was monaural presented diotically, emulating a telephone call taken with earbuds, while the noise signal was binaural.

The listening test was conducted in sound-attenuated booths, each equipped with a computer, mouse and keyboard as well as a pair of Beyerdynamic DT 770 headphones connected through a Focusrite iTrack Solo audio interface.

Participants began the test with the 10 practice sentences, with which they became familiarized with the keyboard and the interface of the MATLAB program used to collect responses. Each practice sentence was played only once, and participants were asked to type out what they had heard. At this stage, participants were allowed to change the volume at

which the stimuli were played, something they could no longer do once the practice session ended.

After the practice session, participants listened to the 8 blocks of 10 sentences, with each sentence played only once, and were asked to type out what they had heard. The ordering of the blocks and of the sentences within the blocks were pseudo-randomized using the MATLAB function `randperm` to avoid any unintended positive or negative effects arising from the sentence or block order.

WARs for the content words of the sentences were computed for each SNR block using an automated scoring script, which divided the number of correct content words by the total number of content words in the sentences, after removing non-content-words "a", "the", "in", "to", "on", "is", "and", "are", "of", "for".

With the WAR results from the 8 SNR blocks, we used the MATLAB `glmfit` function and the `logit` link function to estimate the psychometric curve, by finding the best-fitting logistic function as shown in Cooke et al. (2013). Using the psychometric curve, SNRs corresponding to WARs of 25%, 50%, and 75% were determined for use in the next experiments.

3.2.2 Experiment 1: natural speech over the telephone and in noise

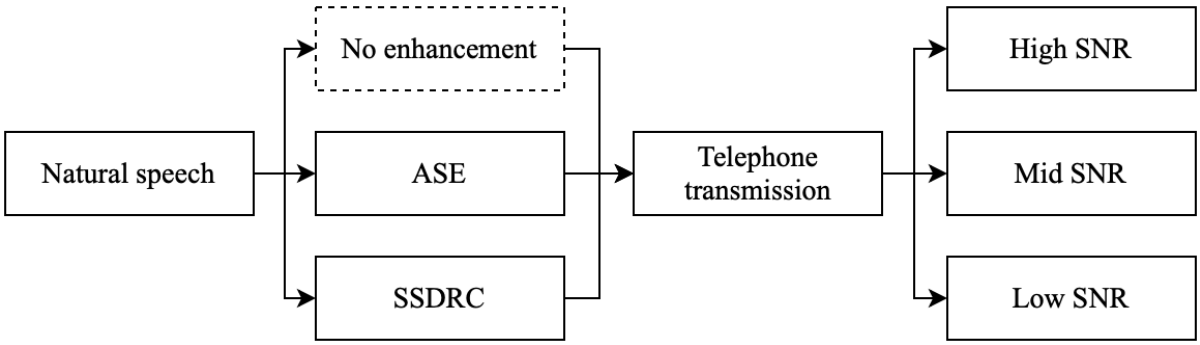

Fig. 3.5: A flowchart illustrating the pipeline for creating the stimuli used in experiment 1.

Using the low, mid, and high SNRs obtained from the calibration test, an experiment was performed to test the effectiveness of NELE algorithms on natural speech, when applied prior to narrowband telephone transmission and playback in noise.

Data was collected from 20 native speakers of English with self-reported normal hearing. These participants were a separate group from the ones who participated in the calibration test to avoid any familiarization to the test sentences used. They were however recruited in the same way and therefore shared similar demographic features.

A total of 100 sentences were used, with 10 being practice sentences. The 10 practice stimuli were plain, unenhanced speech with medium SNRs while the 90 test sentences were used to generate stimuli for each of the 9 conditions (3 NELE conditions * 3 SNR conditions).

The experiment had a practice phase similar to the calibration test where participants familiarized themselves with the testing interface and adjusted the volume level. After the practice session, a random permutation of the 90 test sentences was obtained using the MATLAB function `randperm`, randomizing the ordering of the sentences. Then, they were played in 3 blocks of 30 sentences, each block corresponding to an SNR. The ordering of the SNR blocks and the ordering of the NELE conditions within each block were randomized. Each sentence was played once and participants were asked to type out what they had heard. Due to randomization, a sentence is not tied to a NELE condition or an SNR condition, and indeed appears in different variations across participants.

Finally, WARs for the content words of the sentences were computed for each of the 9 conditions using the same automated scoring script used in the calibration test.

3.2.3 Experiment 2: synthetic speech over the telephone and in noise

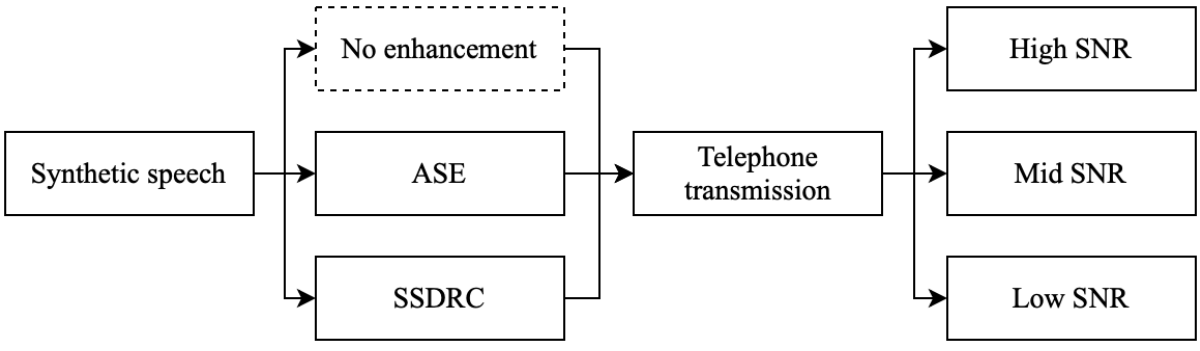

Fig 3.6: A flowchart illustrating the pipeline for creating the stimuli used in experiment 2.

Experiment 2 was performed using the same conditions as the previous experiment except for the natural speech stimuli being replaced by synthetic speech stimuli generated by

the HMM statistical parametric system, in order to test the effectiveness of NELE algorithms on synthetic speech.

Data was collected from 20 native speakers of English with self-reported normal hearing. These participants were again different from the ones who participated in the calibration test or experiment 1 to avoid any familiarization to the test sentences used. They were however recruited in the same way and therefore shared similar demographic features.

The method by which the stimuli were generated, the procedure followed as well as the scoring methodology all mirrored the previous experiment.

4. Results

4.1 Calibration Test

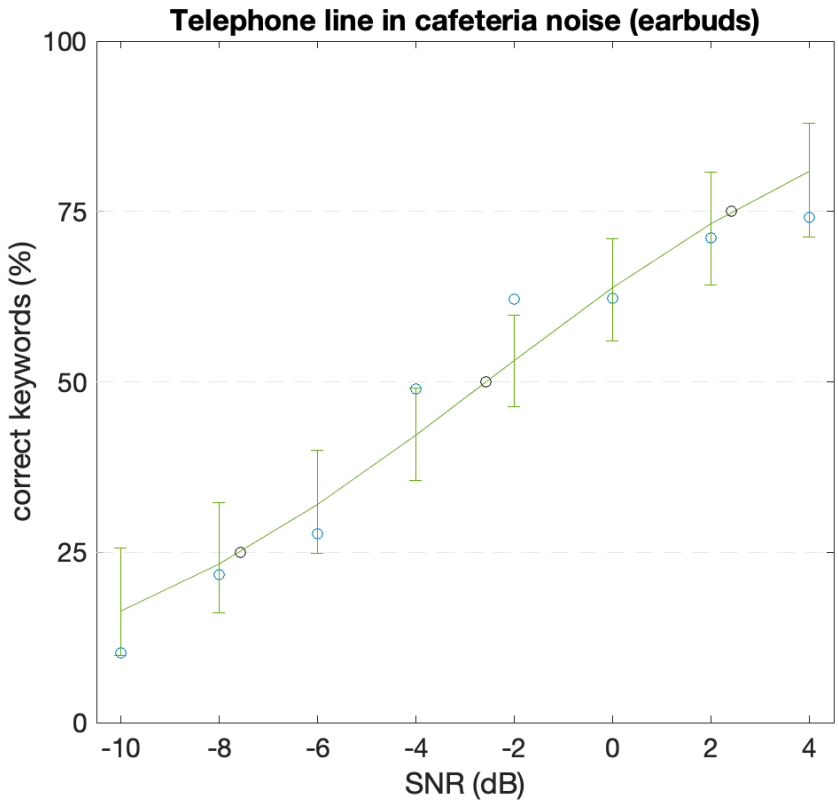

Fig 4.1: Psychometric curve for plain speech over the telephone. Listeners' (n=7) mean WAR (open blue circles) as a function of SNR. Error bars represent 95% confidence intervals. Estimated SNR values for WARs of 25%, 50% and 75% are shown as open black circles.

Results from the pre-experiment calibration test (n=7) were used to find the psychometric curve for plain natural speech over the telephone and in noise, as shown in Figure 4.1. Using this psychometric curve, SNR values of -7.8, -2.9, and 2.1 dB were found to correspond to WARs of 25%, 50%, 75% respectively and were subsequently used for the two main experiments.

4.2 Experiment 1: Natural speech over the telephone and in noise

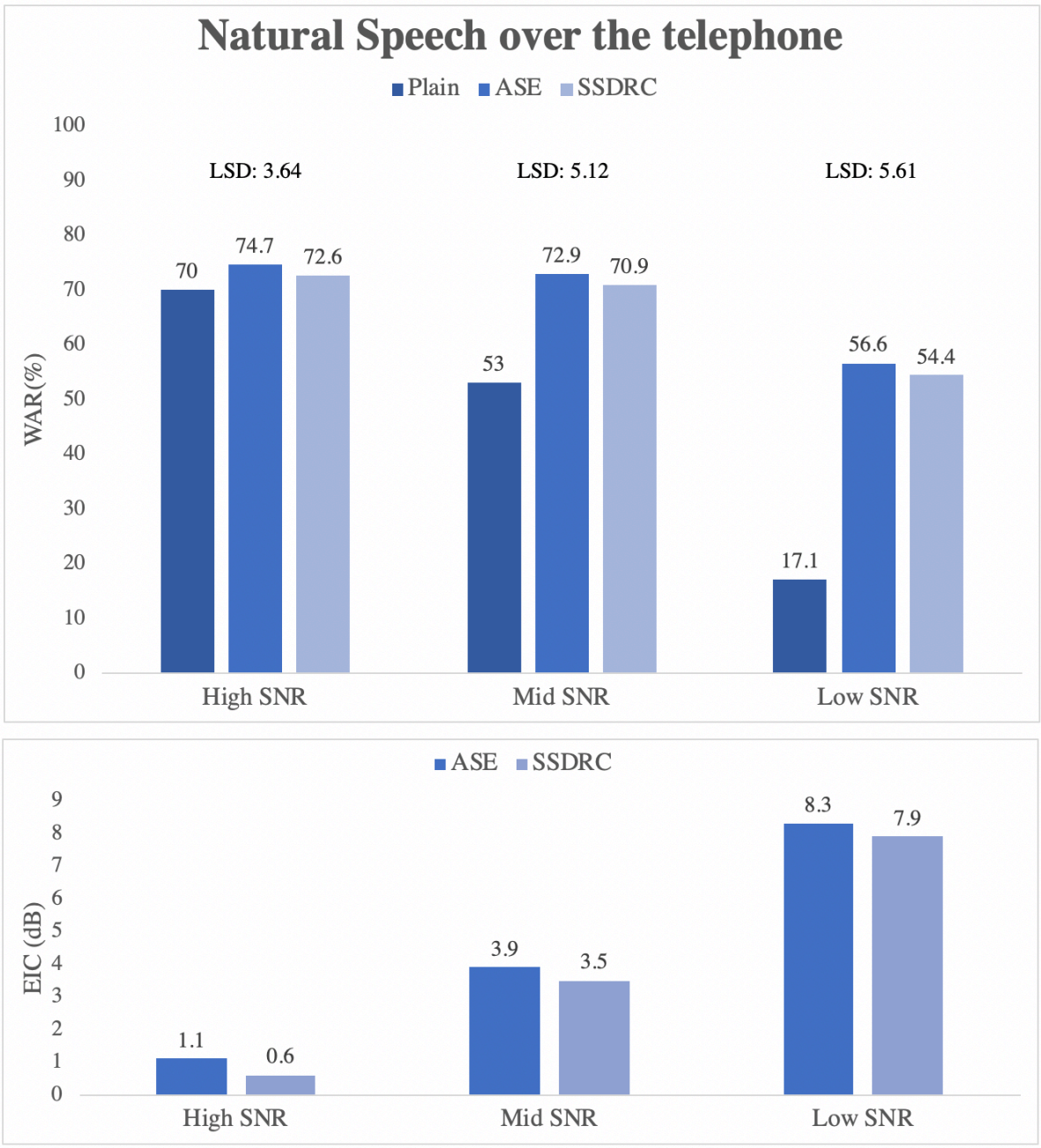

Fig 4.2: WAR and EIC for the three NELE conditions, in the three SNRs. LSD = Fisher's least significant difference.

Results (n=20) from experiment 1 are shown in Figure 4.2, with speech intelligibility reported in terms of WAR and the corresponding EIC. On average, both ASE and SSDRC provided intelligibility gains across all SNR conditions. ASE achieved EIC of 1.1, 3.9, and 8.3 dB, while SSDRC brought EIC of 0.6, 3.5, and 7.9 dB, from high to low SNR.

Intelligibility gains were highest at the noisiest, low SNR condition, followed by mid SNR, and then the least noisy, high SNR condition. On average, ASE provided higher gains than SSDRC.

An analysis of variance (ANOVA) was performed for each SNR to analyze whether NELE conditions had an effect on intelligibility, and Fisher's least significant differences (LSD) were computed at a confidence level of 95%. Results showed that ASE was able to significantly increase intelligibility in all three SNRs, while SSDRC was able to do so in the noisier conditions of mid and low SNRs. Even though on average, ASE appeared to outperform SSDRC, this difference was found to be not statistically significant and in fact, the two algorithms had very similar performance in all three SNRs.

4.3 Experiment 2: Synthetic speech over the telephone and in noise

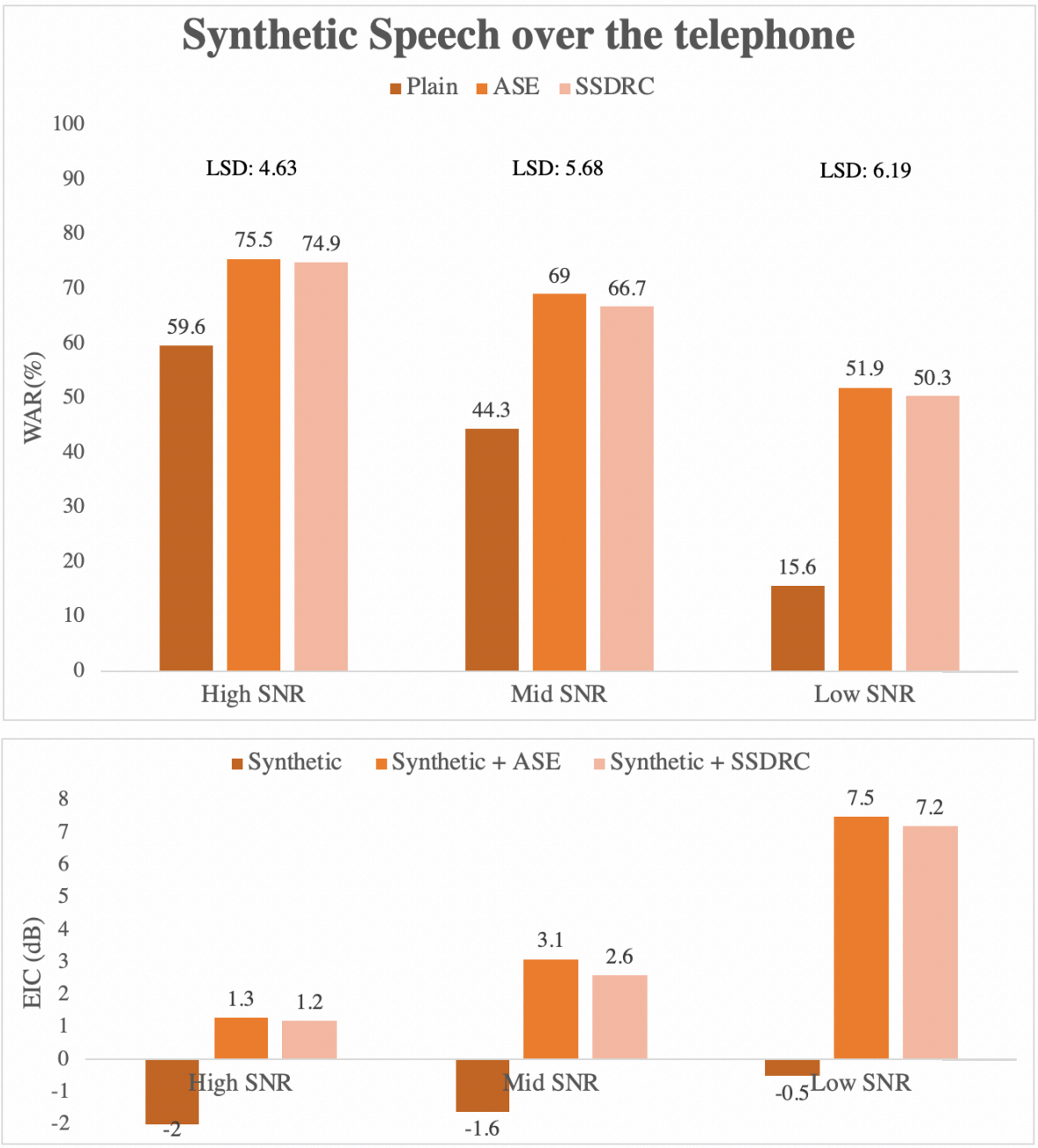

Fig 4.3: WAR and EIC for the three NELE conditions, in the three SNRs. LSD = Fisher's least significant difference.

Results (n=20) from experiment 2 are shown in Figure 4.3, with speech intelligibility reported in terms of WAR and the corresponding EIC. Since the same SNRs from the

previous experiment were used, EICs were calculated from plain natural speech and not from unmodified synthetic speech.

On average, unmodified synthetic speech was less intelligible than unmodified natural speech, this is reflected by the negative EICs of -2, -1.6, and -0.5 dB in high, mid, and low SNRs, relative to plain natural speech. However, despite this, both ASE and SSDRC again provided intelligibility gains across all SNR conditions, with ASE achieving EIC of 1.3, 3.1, and 7.5 dB, while SSDRC brought EIC of 1.2, 2.6, and 7.2 dB, from high SNR to low SNR. The same pattern of intelligibility gains being highest in the noisiest condition and lowest in the least noisy condition was again present.

The intelligibility of NELE-enhanced synthetic speech appeared to be comparable to NELE-enhanced natural speech. In fact, synthetic speech enhanced by both NELE algorithms had higher WARs, on average, than enhanced natural speech in high SNR. In mid and low SNRs, WARs appeared comparable, but with enhanced synthetic speech having lower WARs than enhanced natural speech. ASE-enhanced synthetic speech achieved WARs of 75.5%, 69%, and 51.9% from high SNR to low SNR, compared to 74.7%, 72.9%, and 56.6% for natural speech. For SSDRC, the WARs were 74.9%, 66.7%, and 50.3% compared to 72.6%, 70.9%, and 54.4%.

An ANOVA was performed for each SNR and results showed that both ASE and SSDRC were able to significantly increase intelligibility in all three SNR conditions. Similar to natural speech, no significant difference in performance was found between the two NELE algorithms.

5. Discussion and conclusion

In the previous chapters, I've devised a complete testing pipeline to evaluate whether NELE algorithms are effective in increasing the intelligibility of synthetic speech that is subsequently transmitted through the telephone and presented in additive noise. Results showed that both ASE and SSDRC were effective for both natural and synthetic speech degraded by telephone transmission, across SNR conditions. Additionally, both algorithms were found to be equally effective with no significant difference in performance found between the two.

5.1 Results

The success of NELE algorithms in improving synthetic speech intelligibility over the telephone reflect past studies on NELE algorithms' effectiveness on synthetic speech that did not include telephone degradation in their experiments (Valentini-Botinhao et al. 2013). While all of our experiments included the telephone line, and we therefore cannot make comparisons of the intelligibility difference with or without the telephone line, we can nevertheless make the following observations: (1) telephone transmission does not catastrophically impact the effectiveness of NELE, as intelligibility boosts were observed for both natural and synthetic speech, across all SNR conditions, (2) telephone transmission does not negatively impact NELE effectiveness on synthetic speech more so than on natural speech, as post-enhancement, natural and synthetic speech had comparable intelligibility.

The lower intelligibility of unmodified synthetic speech compared to natural speech was to be expected, given that the same HMM-based system was found to be less intelligible than natural speech in quiet (King & Karaiskos 2011) and given that background noise has been found to have a more severe effect on synthetic speech intelligibility compared to natural speech (King & Karaiskos 2010).

The performance of ASE confirms its effectiveness shown in the preliminary results of the second Hurricane Challenge (Cooke et al. 2020) and shows that it, like SSDRC, is effective for both natural and synthetic speech (Valentini-Bontinhao et al. 2013). The fact that ASE achieved the same intelligibility improvements as SSDRC is impressive due to the fact that SSDRC was designed to maximize intelligibility, often at the expense of speech

quality or naturalness, as shown in Tang, Arnold, & Cox (2017) where SSDRC-modified speech was given a subjective mean opinion score (MOS) of 3 out of 5 in quiet (n=10), an entire point lower than unmodified speech. On the other hand, ASE was developed to achieve a balance between increasing intelligibility and preserving speech quality. In informal listening tests with experts in this area of research, speech processed with ASE was said to have a higher quality and to suffer less from distortion or other quality degradations compared to SSDRC. If this is further corroborated through formal listening tests, then ASE should be considered a better overall NELE algorithm.

5.2 Limitations and future extensions

5.2.1 Materials

In terms of the materials used in the experiments, two main limitations exist. First, the synthetic speech used was generated by the HTS speech synthesis system (Zen et al., 2007) which used an HMM-based statistical parametric approach instead of the current state-of-the-art DNN-based statistical parametric approach (e.g. Zen, Senior, & Schuster, 2013). With a DNN-based system, intelligibility would likely be higher than in the current experiments. However, we believe that the results of this dissertation are still applicable to DNN-based systems as both approaches operate by predicting speech parameters and generating waveforms based on those parameters, instead of using stored samples of natural speech in a database. Therefore, we expect the patterns of intelligibility improvements to be similar with a DNN-system, but with higher WARs across SNRs, tending closer towards the WARs of natural speech. Future studies could use a DNN system in a similar testing pipeline to confirm this prediction.

Secondly, only one type of noise was tested in this study. In most intelligibility evaluation studies, two types of noise are typically used – a fluctuating noise such as a competing speaker and a steady noise such as speech shaped noise (Cooke et al., 2013; Valentini-Bontinhao et al., 2013). In this dissertation, only one type of noise was used – a cafeteria noise environment representative of relatively steady-state noise. We expect SSDRC and ASE would also be effective in fluctuating noise for synthetic speech over the telephone, given their performance in this study and the fact that they have been shown to work in both noise types for natural speech (Cooke et al., 2013; Cooke et al., 2020). However, the pattern of intelligibility gains (e.g. higher gains the noisier the condition), the degree of intelligibility gains, and the SNR levels at which the algorithms are effective would likely be different.

Future studies would have to be conducted to determine these values and to compare the performance of NELE on synthetic speech over the telephone in both types of noise.

5.2.2 Experimental design

In terms of the experimental design, a few limitations and possible extensions have also been noted. First, the psychometric curve for intelligibility over the telephone in noise was obtained for only natural speech and not synthetic speech. By using the same SNRs to test both natural and synthetic speech, we were able to directly compare the absolute intelligibility in terms of WARs, before and after enhancement, for both types of speech. However, since natural and synthetic speech have different psychometric curves, we cannot make any direct comparisons on the degree of intelligibility improvement in EICs between the two types of speech. A future study would need to conduct calibration tests for both types of speech, under the same conditions, in order to obtain their psychometric curves and make EIC comparisons.

Secondly, the demographics of the participants may present confounds. Given that a local Scottish news corpus was used in this study, non-British participants who are nevertheless native speakers may perform worse in identifying unfamiliar British terms such as “Home Secretary”, “ministers”, or “the Tories”, which British participants are more frequently exposed to. This could be prevented by either using a corpus without a geographical or dialectal bias, such as the Harvard sentences or semantically unpredictable sentences, or by collecting more detailed demographics data from participants that could be used to perform statistical analysis with country of origin as a variable.

Finally, a further study can be performed to evaluate the quality of the modified speech. As suggested in the previous section, while ASE and SSDRC achieved similar results in intelligibility, ASE was designed with preserving speech quality as an explicit objective while SSDRC was not. ASE has been judged by expert listeners to be better than SSDRC in preserving speech quality in normal usage and also for synthetic speech over the telephone. This can be tested through two possible means. The first way is through measuring a subjective opinion score of the perceived speech quality and the second way is through measuring objective proxies of listening effort, such as reaction time and recall. This has serious implications for industry applications of NELE over the telephone. For instance, if NELE algorithms are shown to severely impact speech quality and significantly increase listening effort of speech over the telephone, companies may be reluctant to adopt them even

though their effectiveness on intelligibility has been shown in this study, due to the possibility of causing dissatisfaction in their customers. Further, if two NELE algorithm are shown to provide similar intelligibility improvements, but further studies show that one has higher speech quality and requires less listening effort, this would facilitate the selection process of NELE algorithms for companies.

5.3 Industry applications

The two NELE algorithms used in this study have been chosen with realistic industry applications in mind. First, they are noise-independent, thus do not require any knowledge of the background noise signal on the receiver's end of the telephone. Secondly, they are both designed to be usable in real-time, and can therefore work with real-time TTS systems and not just with pre-recorded or pre-generated speech.

The positive results of this study show that the use of NELE algorithms is promising for the use case of synthetic speech over the telephone. As discussed in the previous section, if further studies are able to show that a NELE algorithm can increase intelligibility while preserving speech quality both in quiet and in noise and does not require additional listening effort from the receiver, then it would be an ideal addition to most businesses already using TTS over the telephone.

Further, NELE algorithms may be able to accelerate industry adoption of TTS over the telephone, as NELE algorithms are able to remove one of synthetic speech's primary drawbacks in this scenario—its negative interaction with noise—and bring intelligibility close to natural speech. For banks and other businesses that use pre-recorded messages or even human operators to redirect phone calls, TTS would be a much more economic option and if combined with NELE, it may be able to do so without one of its major drawbacks.

References

Bennett, C. L. (2005). Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005. In *Ninth European Conference on Speech Communication and Technology*.

Chermaz, C. (2020). The automatic sound engineer: a near end listening enhancement algorithm for speech in noise. Unpublished manuscript.

Chermaz, C., Valentini-Botinhao, C., Schepker, H., & King, S. (2019). Evaluating Near End Listening Enhancement Algorithms in Realistic Environments. *Proc. Interspeech 2019*, 1373-1377.

Cooke, M., Mayo, C. & Valtini-Bontinhao, C. (2020). Intelligibility-enhancing speech modifications: the second hurricane challenge. Unpublished raw data.

Cooke, M., Mayo, C., & Valentini-Botinhao, C. (2013, August). Intelligibility-enhancing speech modifications: the hurricane challenge. In *Interspeech* (pp. 3552-3556).

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., & Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, *55*(4), 572-585.

Friedman-Berg, F., Allendoerfer, K., & Deshmukh, A. (2009). *Voice over Internet protocol: Speech intelligibility assessment* (No. DOT/FAA/TC-TN-09/04).

George, E. L., Goverts, S. T., Festen, J. M., & Houtgast, T. (2010). Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners. *Journal of Speech, Language, and Hearing Research*.

Gordon-Salant, S. (1986). Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. *The Journal of the Acoustical Society of America*, *80*(6), 1599-1607.

Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects. *Language and Speech*, *43*(3), 273-294.

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *The journal of the acoustical society of America*, *28*(2), 303-310.

International Telecommunication Union. (1988). Recommendation G.711: Pulse code modulation (PCM) of voice frequencies.

International Telecommunication Union. (1990). Recommendation G.726: 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM).

International Telecommunication Union. (2019). Recommendation G.191: Software tools for speech and audio coding standardization.

Jokinen, E., & Alku, P. (2017). Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network. *The Journal of the Acoustical Society of America*, *141*(4), EL327-EL330.

Jokinen, E., Remes, U., & Alku, P. (2017). Intelligibility enhancement of telephone speech using Gaussian process regression for normal-to-Lombard spectral tilt conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(10), 1985-1996.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252-1263.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., & Kollmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, *2009*(1), 298605.

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, *1*(1), 006.

King, S., & Karaiskos, V. (2010). The Blizzard Challenge. *Proc. Blizzard Challenge workshop 2010*.

King, S., & Karaiskos, V. (2011). The Blizzard Challenge. *Proc. Blizzard Challenge workshop 2011*.

Leviathan, Y., & Matias, Y. (2018). Google Duplex: an AI system for accomplishing real-world tasks over the phone. *Google AI blog*, *8*.

Monson, B. B., Hunter, E. J., Lotto, A. J., & Story, B. H. (2014). The perceptual significance of high-frequency energy in the human voice. *Frontiers in psychology*, *5*, 587.

Moore, B. C., Füllgrabe, C., & Stone, M. A. (2010). Effect of spatial separation, extended bandwidth, and compression speed on intelligibility in a competing-speech task. *The Journal of the Acoustical Society of America*, *128*(1), 360-371.

Morimoto, M., Sato, H., & Kobayashi, M. (2004). Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces. *The Journal of the Acoustical Society of America*, *116*(3), 1607-1613.