CHAPTER 1

Automatic Speech Recognition of Ear-Recorded Speech

1.1 Introduction

The automatic recognition of human speech by a computer has been a subject of interest spanning decades. Humans first and foremost communicate their ideas via speech and human language, and teaching computers to be able to take verbal instructions would make interaction with them much easier for a majority of the population, particularly the elderly and disabled. Since this task has been a subject of much study for over half a century, and is only recently gaining much success, it is important to briefly discuss these successes and the challenges that still remain.

The present study proposes a new technique to be used in the advancement of noise-robust automatic speech recognition (ASR). The experiment in Chapter 2**??** collected ear-recorded speech data, which aimed to overcome the difficulty of accurately perceiving speech in a noisy environment, for recognition both by computer (ASR) or by human speech perception. This collected data will be used in an experiment utilizing the standard open-source ASR system Kaldi (Povey et al. (2011)) with the standard, freely-available acoustic model developed from the LibriSpeech corpus (Panayotov et al. (2015)).

1.2 Background

While there are more than 30 years of research involving ASR, and nearly as many working with speech in noisy environments, it is impossible to mention all or even most developed techniques within this very brief overview. Therefore, the discussion will involve only several areas of research in noise-robust ASR that pertain to the present study.

Equation 1.1 is often used to represent the combination of speech and noise,

$$x * h + n = y \tag{1.1}$$

where x is the clean speech signal, h is any convolution (eg. room reverberation, microphone channel warp, etc.) of the original speech signal, n is additive noise, and y is the noisy speech signal. In addition, the phase of the component signals contributing to the resulting signal y also has an effect. Taking phase into account means that additive noise is not always simply additive, as energy at frequencies with competing phases could cancel each other out to varying degrees, depending on phase. While there is some research that does take phase into account and demonstrates that by doing so, more accurate noise removal can be achieved, attempting to model phase adds significant complexity, and the - albeit problematic (many researchers admit) - assumption is frequently made that phase plays no role (Li et al. (2014)).

The categorization of noise-robustness techniques utilized by hundreds of researchers over the years is difficult, but, broadly, two domains can be outlined (Li et al. (2014); Zhang et al. (2017)). The first is the "feature-space" domain, which focuses on front-end processing of the signal y itself. The second group utilizes the "model space" domain, or the back-end processing that modifies the acoustic model to account for any noise in the signal y. Feature space is the part of the ASR process where an acoustic vector is transformed into 'features' describing salient parts the acoustic signal that the ASR model will receive as input. Model space, consequently, then, comes after feature space in the ASR process, and encompasses the acoustic model parameters, methods of training the model, etc.

Noise can be accounted for in either the feature space domain or the model space domain, or both. In the feature domain, noise is dealt with, and the signal is enhanced, prior to sending the features to the acoustic model for recognition. These modifications are made without altering the acoustic model parameters, resulting in low computation cost. In the model space, acoustic parameters themselves can be modified in accordance with the noisy signal. This generally results in high computation cost when training the acoustic model. There is normally a trade-off between the two domains: one of computation efficiency and performance. Model space alterations often yielding higher performance improvement, while feature space alterations are less computationally intensive (Li et al. (2014)). Advances in neural network technology have also lent itself to application in noise-robust ASR. These have provided additional methods of tackling the problem.

1.2.1 Feature Space Domain

In the feature space domain of noise-robust ASR processing, there are a number of broad techniques, including (a) noise-resistant features, (b) feature normalization, and (c) feature compensation. Noise resistant features are, quite simply, features in the acoustic signal which are not sensitive to environmental changes. Many researchers have proposed many methods of signal derivation that incorporate features of the human auditory system, including Perceptual Linear Prediction (PLP, Hermansky et al. (1985)), introducing the "auditory spectrum" and explicit formant information into ASR processing, and Relative Spectral processing (RASTA) applied to PLP (Hermansky et al. (1992)), making PLP less sensitive to slow varying speech information and more sensitive to the more rapid-varying transitions of speech, which is important in human speech perception (Willi (2017)). Kim et al. (1999) attempts to model functions of the cochlea and auditory nerve. These methods are quite effective at dealing with short-term, stationary, additive noise (Zhang et al. (2017)). More recent methods include SPARK (Fazel and Chakrabartty (2012), 1369), which is "neurobiologically inspired" by "auditory receptive fields" and "local competitive behavior", and that proposed by Moritz et al. (2015), which emulates the amplitude modulation found in mammalian auditory cortexes. These are just a selection of methods incorporating biologically based noise resistant features, which generally outperform "vanilla" MFCC methods. The fact that these features can be quite complex to generate, and the parameters difficult to set, makes it hard to utilize a combination of them, preventing widespread usage and incorporation with other techniques. A straightforward relation between the cleaned features and noisy speech is also difficult to derive due to the complexity involved in the feature calculation (Li et al. (2014)).

Feature normalization (b) generally involves normalizing cepstral feature vectors in the form of cepstral mean normalization (CMN) and cepstral mean and variance normalization (CMVN). CMN involves finding the mean values out of all cepstral vectors (Atal (1974)). All cepstral vectors are then normalized, such that the mean cepstral value becomes zero. CMN primarily eliminates reverberation and channel-related distortion, but signals with noise and no channel distortion also see improvement (Droppo and Acero (2008)). CMVN takes the mean and normalizes it together with the co-variance of the cepstral vectors, yielding improved performance on speech data with additive noise (Viikki et al. (1998)). These methods do not work in real-time, however, as they require cepstral vectors from the entire utterance in order to calculate means and variances.

Feature compensation (c) actually attempts to remove the noise from the noisy speech signal, allowing for use of traditional features. Spectral subtraction (Boll (1979) is an intuitive method of removing noise by taking a small window of the waveform, turning the linear signal into the spectral domain, and subtracting an existing or estimated noise spectrum n from a noisy speech spectrum y, leaving a spectrum of clean speech x. This is then converted back into the time domain, and the process is repeated all along the waveform. Spectral subtraction often estimates the noise by looking at sections of the observed signal that do not contain speech information.

This method still has several problems (Li et al. (2014)). First and foremost, the location of speech in the signal in a noisy environment is difficult to detect, which consequently affects the ability to accurately compute a noise average. It also requires relatively stationary, slow-variation noise; noise that changes quickly can have a different average spectrum during the portion of the signal containing speech than the portion of the signal in which the noisy spectrum was calculated. Furthermore, this is only an average of the noise, and not exactly the noise itself. This subtraction of the average can inadvertently have an additive noise effect by producing extraneous acoustic artifacts in the “clean” signal which were not there to begin with (Berouti et al. (1979)).

Wiener filtering (Lim and Oppenheim (1979)) is another method used to remove noise from a signal. As opposed to spectral subtraction, however, this is a linear filter that works without the need to convert the signal into spectra. However, this method also requires an estimation of the noise. Furthermore, it does not do well in very low SNR environments, as it generally results in suppression and dampening of the entire signal, and not just the noise (Li et al. (2014)).

More standard, is the “advanced front-end” (AFE) ensemble proposed in ETSI (2002). It yields more than 50% improvement over standard MFCC features alone, and has become a frequent baseline for comparison in noisy ASR research. It is composed of three separate ‘tools’: two-stage Mel-warped Wiener filtering, SNR-dependent waveform processing, and blind equalization (cf. Agarwal and Cheng (1999); Macho et al. (2002); Macho and Cheng (2001); Mauuary (1998)).

Most of the heavy work of the AFE ensemble is performed by the Mel-warped Wiener filtering (Agarwal and Cheng (1999); Li et al. (2014)). This filter differs from the more standard Wiener filter in that it uses the Mel-frequency power spectrum in the Wiener filter calculations, as opposed to the linear signal, the result of which is then converted back into the time domain. The filter is applied once, and then a second time to remove residual noise. SNR-dependent waveform processing (SDWP, Macho and Cheng (2001)) assumes that the noise remains relatively constant, whereas the speech signal causes variation in the amplitude of the signal. SDWP uses this assumption to dampen portions of the signal with a relatively constant and low SNR compared with the high SNR (ie, speech-less versus speech-bearing) portions of the signal, which are amplified. Blind equalization serves to eliminate convolutional (eg. reverberant) distortion from the signal (Mauuary (1998)).

Considered to be the best “general purpose” noise-removal tool (Zhang et al. (2017), 4) using traditional (non-neural network) techniques, the minimum mean square error (MMSE) magnitude modulation estimator (MME) was developed by

Paliwal et al. (2012), and based on the acoustic modulation estimator (AME) first proposed by Ephraim and Malah (1984). The approach utilizes the spectral modulation magnitude domain, rather than the spectral frequency domain (as is used in the AME method), which is where much of its success originates.

1.2.2 Feature Space Domain: Neural Networks

Broadly, there are two primary categories of utilizing neural networks to account for noisy speech: “mapping methods” and “masking methods” (Zhang et al. (2017)). Mapping methods involve finding the non-linear function that maps the noisy speech to the clean speech. In neural network terms, the noisy speech is the input to some type of neural network (eg. Deep Neural Network (NN), Convolutional NN, Recurrent NN, etc.) and the (intended) output is an approximation of the clean speech. Due to the complexity of speech in the temporal domain, the input to the neural network usually comes from one of the higher-processed input transformations, such as from the spectral or cepstral domains.

Masking-based approaches work similarly to a traditional filter, albeit learned via a neural network. A method using an ‘Ideal Ratio Mask’ will use a neural network to learn the ratio (value between 0 and 1) of the presence of clean speech to noise. This process is most beneficial when using spectral or cepstral features as inputs. These calculated masking ratio values are then multiplied element-wise to each spectral or cepstral feature (from the noisy signal y) at every time index, returning the estimation of the clean speech x as output (Zhang et al. (2017)).

1.2.3 Model Space Domain

The description of model space domain compensation techniques will be brief, as this is not the focus of the present study. This form of compensation usually involves adapting an existing acoustic model (presumably trained on relatively clean speech) to enable recognition of more noisy features. Variations of maximum likelihood linear regression (MLLR, Leggetter and Woodland (1995)) are often used to adjust the model means and co-variance parameters to account for differences in the signal that is introduced by noise. There are many variations and extensions of MLLR; one such variation, feature-space MLLR, or fMLLR, actually moves MLLR application into the feature domain (Gales (1998)).

There are also a few model-based approaches using neural networks for noise-robust ASR. Most widely used is multi-condition training (Seltzer et al. (2013); Zhang et al. (2017)), which, similar to multi-style training originally developed by Lippman et al. (1987), uses a collection of training data that exhibits a wide range of noise conditions. Another technique, similar to methods used in non-neural network approaches, involves adapting the already trained acoustic model with a small subset of noisy data. However, as doing so can inadvertently result in significant overfitting, Mirsamadi and Hansen (2015) have developed a technique unique to neural networks that - instead of slightly adjusting all weights, adds an additional layer to the neural network with its own weights. This largely avoids the issue of overfitting, while increasing the model's robustness to noise.

Weninger et al. (2013), among many others, also combine the modifications in the feature space domain with modifications in the model space domain, referred to as joint model training. Broadly, this takes the form of using the feature-based noise removal techniques to output feature-enhanced data which is then used as training data itself for the acoustic model.

1.2.4 Microphone Arrays

There are also techniques that employ multiple microphones as a method of source-separation to extract the speech source from any extraneous noise sources. Beamforming (Van Veen and Buckley (1988)), for example, has become a central technique to using microphone arrays for source separation (Hori et al. (2015); Zhang et al. (2017)). The direction of arrival of the different sound sources is calculated, taking into account the distance between the two (or more) microphones, and the time of arrival of the different sources in each signal recorded by each microphone. Recent work (cf. Heymann et al. (2015); Sivasankaran et al. (2015); Heymann et al. (2016))

has also employed neural networks to aid and enhance the beamforming process .

1.2.5 Summary

Most research over the past few decades has focused on feature space domain modifications. This is likely due to the intensive computation required by many model space domain techniques. Leading feature-space techniques include MMSE-MME (Paliwal et al. (2012)) and AFE (ETSI (2002)). Model space domain approaches include adjusting the acoustic model parameters, often using a form of MLLR (Leggetter and Woodland (1995)), which can also take the form of fMLLR in the feature space domain. In the last few years, the advent of neural networks has seen further improvement in both feature and model space domains. Other recent approaches have combined feature and model space modifications (joint model training), and the use of multiple microphones into a microphone input array has also become more mainstream. The recent CHiME challenges (CHiME Challenge (2016)) have incorporated the use of multi-channel ASR input as well as single channel input as part of its task.

Most of these employed feature and model space techniques that are used to account for noise are still forced to make estimations about noise type, SNR, and noise location in the signal. As would be expected, as SNR decreases, and as noise becomes more variable (non-stationary), these methods begin to falter. Chapter 2**??** proposes a method of collecting speech in noisy environments (recording speech from the inside of the ear canal) that is hypothesized to be largely immune to noise type and the stationarity (or lack thereof) of noise. It does not require any noise estimation or inference. It would be classified as a form of feature-space modification, affecting the signal in the temporal domain before it even reaches the microphone. The only drawback encountered is that the speech in the recorded signal is heavily low-pass filtered, where the highest speech frequencies observed in the signal are generally found near 2.7 kHz.

The only processing performed after the signal is recorded is pre-emphasis and low-pass filtering, which could be easily built into the recording mechanism itself.

It is hypothesized that this method of passive noise removal via recording speech from the ear canal, plus the minimal modification of pre-emphasis and low-pass filtering, will demonstrate similar gains over noisy speech achieved by many of the other techniques described above.

1.3 Experiment 3: ASR of Ear-Recorded and Noisy Mouth-Recorded Speech

While there are many proposed techniques, discussed in Section 1.2, that have been used to modify the acoustic features of noisy speech, or to modify the acoustic model to compensate for noise, noise-robust ASR is still imperfect and requires additional advances to ASR technology (Zhang et al. (2017)). This particular study proposes the new technique of using speech recorded from the inside of the ear canal. This would be classified as a feature space modification in the temporal domain, prior to any processing. Rather than using significant computation to achieve the noise reduction, this study employs purely passive mechanisms (ie. tissues in the head, earplug, ear muffs) to reduce noise.

As described previously in Chapter 2**??**, very simple signal enhancement techniques (ie. pre-emphasis and band-pass filtering)[1] are then applied to the recorded signal to produce an enhanced signal with relatively little noise and one that is very similar (below 2.7 kHz) to what could be recorded at the mouth.

1.3.1 Stimuli

Recordings from twenty speakers, ten male and ten female, from the data collection experiment in Chapter 2**??** comprises the test data for this experiment. This included 30 distinct sentences from each speaker, each with 5 different noise conditions (bus, cafe, pedestrian, street, factory) with 3 different noise levels (60dB, 70dB, 80dB), plus an additional 'clean' (no noise) condition. This results in 16 iterations of each distinct sentence (30), for each microphone location (2), resulting in 960

[1] These are simple enough to be hard-wired into an electrical chip to be performed in real-time, requiring no actual computation.

utterances for each of 20 speakers, totaling 19200 test utterances. There are 9600 ear-recorded utterances (6 hours, 55 minutes)[2], 9000 mouth-recorded noisy utterances (6 hours, 30 minutes), and 600 mouth-recorded clean utterances (26 minutes).

Two additional datasets were used. One was recorded from the mouth with the intention to obtain speech with a lower SNR - described in Chapter 4??, Section ??. The other was the dataset (recorded at the same time) of ear signals combined with the mouth signals from the dataset above; the low-frequency ear-recorded signal was combined with the high-frequency components of the mouth-recorded signal, described in Chapter 4??, Section ??. For these two datasets, each contained two speakers, one male and one female. The dataset of speech combined from ear- and mouth-recorded signals contains 80 distinct utterances, each repeated 5 times (the 'clean' speech condition was removed) by each of 2 speakers, totaling 800 utterances. The dataset of speech recorded at the mouth in noisy conditions contains 80 distinct utterances, each repeated 5 times (again, the 'clean' speech condition was removed) by each of 2 speakers, totaling 800 utterances. These datasets were recorded simultaneously at the ear and at the mouth, from the same two speakers.

1.3.2 Design

The existing deep neural network (DNN) acoustic model trained on 960 hours of speech from the LibriSpeech corpus (published and described in Panayotov et al. (2015)) was used to test the collected data. As described in Panayotov et al. (2015), the 960 hours of data, collected from the audio transcripts of books from the LibriVox website, was divided into two groups - 'clean' and 'other'. This designation was chosen by preliminarily running an acoustic model trained on Wall Street Journal data (WSJ, Paul and Baker (1992)) on the LibriSpeech utterances. The set was split down the middle, with the half containing lower WERs designated at the 'clean' set, and those with higher WERs designated as the 'other' set. Data from both sets were used in training the LibriSpeech acoustic model.

Additionally, the ASR setup utilized the language models from Panayotov et al.

[2] Time estimates are calculated from an average utterance duration of 2.6 seconds

(2015), trained on the LibriSpeech corpus. To verify the current set-up of the acoustic and language models, the experimenter performed a replication of Panayotov et al. (2015)'s experiment using LibriSpeech's "test-clean" and "test-other" datasets. Afterwards, the experimenter used the the same acoustic and language models to recognize the speech data collected for this study described in Chapter 2**??**. This primarily tested the performance between the ear-recorded and noisy mouth-recorded speech, but also between the different noise conditions and noise levels. The experimenter also tested the data collected for, and described in Chater 4**??**, containing very noisy mouth-recorded data and the data combination of mouth- and ear-recorded speech.

1.3.3 Results

The experimenter used LibriSpeech test data on the configured acoustic and language models specified above. Table 1.1 demonstrates that accuracy similar to the published results was achieved, and the acoustic and language model set up was verified.

Language Models	Clean Publ.	Clean Repl.	Other Publ.	Other Repl.
3-gram, thresh. 3e-7	8.02	9.15	19.41	23.76
3-gram, thresh. 1e-7	7.21	8.20	17.66	21.55
3-gram, unpruned	5.74	6.50	14.77	18.37
4-gram, unpruned	5.51	6.20	13.97	17.53

Table 1.1: The 'Language Models' column specifies the Language Model (LM) used in each row; two LMs are simply the 3-gram model which was pruned to the specified threshold. 'Publ.' columns list the performance listed in the published paper, and 'Repl.' columns contain the replication performance achieved in the present study. 'Clean' and 'Other' refer to the clean (2707 utterances) and 'noisy' (5968 utterances) LibriSpeech test datasets. All LMs utilize the 960-hour LibriSpeech DNN acoustic model. All values are given as WER.

The experimenter then used data collected in the present study, described in Chapter 2**??** with the same acoustic and language models. Table 1.2 shows these results. This table combines all noisy speech at the mouth into a single 'Noisy

Mouth' category and all speech collected at the ear into a single 'Ear Speech' category.

Language Models	Clean Mouth	Noisy Mouth	Ear Speech
3-gram, prune thresh. 3e-7	16.74	51.04	84.10
3-gram, prune thresh. 1e-7	14.01	49.11	83.55
3-gram, unpruned	9.52	44.49	82.29
4-gram, unpruned	9.38	44.31	82.47

Table 1.2: The 'Language Models' column specifies the Language Model (LM) used in each row; two LMs are simply the 3-gram model which was pruned to the specified threshold. "Clean Mouth" includes only the sentences recorded at the mouth with no noise, versus "Noisy Mouth", which includes all other [noisy] sentences recorded at the mouth. "Ear Speech" contains all ear-recorded utterances. All LMs utilize the 960-hour LibriSpeech DNN acoustic model. All values are given as WER.

Table 1.3 separates out the noisy mouth-recorded speech into its different noise types and noise levels.

	Bus	Cafe	Pedestrian	Street	Factory	Totals
60 dB	21.51	20.33	18.64	19.05	17.93	**19.49**
70 dB	41.74	32.93	32.02	36.49	29.96	**34.63**
80 dB	88.20	73.39	75.64	85.00	71.43	**78.73**
Totals	**50.48**	**42.22**	**42.10**	**46.85**	**39.77**	**44.28**

Table 1.3: These are only from the highest-performing (4-gram) language model, utilizing the 960-hour LibriSpeech DNN acoustic model. Each row is a different noise level, and each column a different noise type (excluding the 'clean' noise type). Totals (averages) for each category are given in the bottom-most row and the right-most column. The overall average differs slightly from the joint noise WER value given in Table 1.2, but the difference can be attributed to rounding estimations. All values are given as WER.

The experimenter performed another test using two sets of data created in Chapter 4**??**. One includes the very noisy speech collected for the experiment in Chapter 4**??**. This speech was re-recorded in a way that resulted in a lower SNR; for more details, refer to Section **??**.

Language Models	Ear/Mouth Combined	Extra Noisy Mouth
3-gram, prune thresh. 3e-7	85.18	98.32
3-gram, prune thresh. 1e-7	84.66	99.14
3-gram, unpruned	84.04	99.03
4-gram, unpruned	83.90	99.41

Table 1.4: The 'Language Models' column specifies the Language Model (LM) used in each row; two LMs are simply the 3-gram model which was pruned to the specified threshold. The 'Ear/Mouth Combined' column contains the results from the dataset using speech reconstructed from low-frequency ear-signal and high frequency mouth-signal parts. The 'Extra Noisy Mouth' column contains results from the re-recorded mouth speech (described in Section **??**), with a lower SNR. Note that the data used in these tests are different than those used in previous tests displayed in Tables 1.2 and 1.3. All LMs utilize the 960-hour LibriSpeech DNN acoustic model. All values are given as WER.

The other data set combined the low-passed ear-recorded speech (approximately 0-2.7 kHz) with the very noisy speech from the same utterance, high-pass filtered from approximately 2.7 kHz to 8 kHz. The WER results are given in Table 1.4.

1.4 Discussion

The results indicate, broadly, that the ear-recorded speech (cf. Table 1.2) (which has been pre-emphasized and low-pass filtered < 2.7 kHz) performs minimally better than two of the 80dB noise conditions - the bus and street background noises (cf. Tables 1.2 and 1.3) - but none of the other conditions. Performing better than other conditions achieving 80+% WER does not offer much in terms of benefit, as there is minimal recognition happening in either case. This is clear in Table 1.2, noting how little improvement is seen when more extensive language models are used. There is not much natural language that the acoustic model uncovers which can be improved by the language model.

The ear-recorded speech falls well below the performance of the ASR system on the speech in the 60- and 70- dB conditions. The 60 dB condition demonstrates that the SNR in this condition was very high, noting that its performance was only 10%

WER short of the clean-speech condition - occurring with no other noise-reducing measures taken. And this - the lack of other noise-reducing measures applied to the noisy speech - is yet another reason why the ear-recorded speech falls behind; after other noise-reduction transforms are applied, the noisy speech will be even more easily recognizable.

The noisy speech recorded at the mouth in Chapter 4**??** to obtain a lower SNR for the noisy speech condition (cf. Table 1.4) served its purpose. As expected, performance drops substantially to near zero accuracy (cf. Table 1.4), regardless of the language model. Again, this is without the application of any noise enhancement techniques.

The experimenter used the data from the human speech perception experiment, described in Chapter 4**??**, Section **??**, to test whether reintroducing noisy upper frequencies to the ear-recorded signal resulted in an improvement in ASR accuracy. Although using different speakers with a slightly different set of sentences, on its surface, the combination appears to offer no benefit over the low-pass filtered ear-recorded speech alone. Results indicate an approximate 1% increase in WER over the regular low-pass filtered ear-recorded speech (cf. Tables 1.2, 1.4).

However, rather than adding noisy upper frequencies to the ear-recorded speech, it can also be thought of as replacing the noisy lower frequencies of the mouth recorded speech with the lower frequencies of the clean, ear-recorded speech. To this degree, adding in clean lower frequencies results in an improvement of nearly 16% WER over the noisy mouth data (which was recorded simultaneously with the ear-recorded speech). This is accomplished without the application of additional noise-removal processing.

1.4.1 Limitations and Future Research

The low-pass filtered ear-recorded data was run using an ASR system trained for full frequency speech on data ranging from 0 to 8 kHz. The results make it clear that models trained on “normal” speech will not perform well on ear-recorded speech devoid of significant modification. If future researchers collected enough additional speech from the ear to amass a corpus of adequate size for adapting an existing acoustic model or training a new model, performance would be expected to improve. The degree to which performance would improve depends on the amount of critical speech information in the higher frequencies that do not make it into the ear-recorded signal. I hypothesize that there would be enough speech information below 2.7 kHz to substantially improve ASR performance with acoustic model adaption or retraining.

Considering the possibility of using a combination of low-frequency ear-recorded speech and high-frequency noisy mouth-recorded speech, existing noise-removal methods may be able to utilize the relatively clean low-frequency speech harmonics to "clean" the high-frequency, noisy harmonics. This would result in a much more uniform signal across the frequency spectrum and likely much lower WERs.

In many of the ear-recorded signals, one can observe a small amount of noise reaching the microphone and appearing in the signal. One advantage of using speech recorded from the ear (either in isolation or combination with high-frequency mouth-recorded speech) is that, due to affecting the signal at the lowest level of processing (in the temporal domain) prior to any level of higher processing (eg. in the spectral or power domains), many of the discussed feature enhancement techniques, such as AFE and MMSE-MME (discussed in Section 1.2), can be applied to this signal. Despite the noise in these signals, the ear-recorded signal retains a very high SNR, and I hypothesize that these noise removal techniques would work quite well.

Another potential method to further eliminate any noise that reaches an ear-recorded signal would be to use beamforming on input signals from two different microphones, one in each ear. This method can utilize the difference in temporal arrival of the same sound source in different microphones to tease apart the multiple sources and identify the speech source; the speech source will arrive at both ear canals at the same time (each canal is the same distance from the vocal tract), whereas noise sources will likely not arrive at both in-ear microphones simultaneously.

This study utilized noise that was by and large stationary in amplitude. This was intentional, to test the proof of concept and to test the extent of noise reduction the proposed method can accomplish. In theory, as has been shown in Chapters 2?? and 4??, that the noise does not have a dramatic effect on speech recorded from the ear, variations and modulations in the amplitude of the noise (and the SNR of the speech recorded at the mouth) should have little effect on the speech recorded at the ear. Nevertheless, any future research in this area should incorporate noise that fluctuates in both type and loudness. The recent CHiME Challenge (2016) has incorporated amplitude-varying noise into their task, and similar tests could be performed by collecting another data set of speech recorded from speakers' ears, with amplitude-varying background noise to determine if amplitude variance has an effect on ear-recorded speech.

1.4.2 Summary

Many methods of noise-reduction have been developed over the last several decades, and much improvement has been seen recently, particularly for signals with a higher SNR (Zhang et al. (2017)). Nevertheless, lower SNR speech,especially that with a negative SNR, still proves to be a challenge for modern ASR systems. The proposed method of recording the speech from the inside of the ear canal does offer a noise-reduction benefit, but also significantly filters out the higher frequencies within the speech signal. The results above demonstrate that enough speech information that the ASR acoustic model relies upon is filtered out of the higher frequencies to substantially reduce ASR recognition performance of ear-recorded speech, even more-so than noisy speech with a moderate SNR. The ear-recorded speech does outperform very noisy, low SNR speech data (prior to any noise-reduction methods performed on the noisy speech), but neither data's accuracy is suitable for any real-life ASR application.

However, adding low-frequency ear-recorded speech, which is mostly devoid of noise, to the noisy mouth-recorded signal (replacing the noisy low-frequency speech in that signal) does substantially improve performance. This improvement can be seen prior to the application of any further noise-reduction method or signal enhancement method.

It is also possible for future research, with substantial ear-recorded data, to adapt or train an acoustic model on that style of speech. If this could be accomplished with success, adequate recognition could be achieved solely through ear-recorded speech, without the need to use it in combination with the noisy speech signal from the mouth. It is clear that additional steps are necessary to determine the full extent of benefit that can be achieved by using ear-recorded signals for noise robust ASR systems.

REFERENCES

Agarwal, A. and Cheng, Y. M. (1999). Two-stage Mel-warped Wiener filter for robust speech recognition. In *Proc. ASRU*, volume 99, pages 67–70.

Atal, B. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Am.*, 55(5):1304–1312.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, volume 4, pages 208–211. IEEE.

Boll, S. F. (1979). Suppression of Acoustic Noise in Speech Using Spectral Subraction. *IEEE*, ASSP-27(2):113–120.

CHiME Challenge (2016). Chime speech separation and recognition challenge. `http://spandh.dcs.shef.ac.uk/chime_challenge/`. Online; accessed 02-18-2016.

Droppo, J. and Acero, A. (2008). Environmental robustness. In *springer handbook of speech processing*, pages 653–680. Springer.

Ephraim, Y. and Malah, D. (1984). Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121.

ETSI (2002). Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. *ETSI ES 202 050 Ver.1.1.1.*

Fazel, A. and Chakrabartty, S. (2012). Sparse Auditory Reproducing Kernal (SPARK) Features for Noise-robust Speech Recognition. *IEEE Trans. Audio, Speech, Lang Process.*, 20(4):1362–1371.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98.

Hermansky, H., Hanson, B., and Wakita, H. (1985). Perceptually based linear predictive analysis of speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 509–512. IEEE.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1992). RASTA-PLP speech analysis technique. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 121–124. IEEE.

Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R. (2015). BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 444–451. IEEE.

Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 196–200. IEEE.

Hori, T., Chen, Z., Erdogan, H., Hershey, J. R., Le Roux, J., Mitra, V., and Watanabe, S. (2015). The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 475–481. IEEE.

Kim, D. S., Lee, Y. S., and Kil, R. M. (1999). Auditory Processing of Speech Signals for Robust Speech Recognition in Real-world Noisy Environments. *IEEE Trans. Speech Audio Process.*, 7(1):55–69.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An Overview of noise robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777.

Lim, J. S. and Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. In *in Proc. IEEE*, volume 67, pages 1586–1604.

Lippman, P., R., Martin, E. A., and Paul, D. B. (1987). Multi-Style Training for Robust Isolated-Word Speech Recognition. *Acoustics, Speech and Signal Processing, IEE International Converence on ICASSP.*

Macho, D. and Cheng, Y. M. (2001). SNR-dependent waveform processing for improving the robustness of ASR front-end. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 305–308. IEEE.

Macho, D., Mauuary, L., No, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., and Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases. In *Seventh International Conference on Spoken Language Processing.*

Mauuary, L. (1998). Blind equalization in the cepstral domain for robust telephone based speech recognition. *EUSIPCO*, 1:359–363.

Mirsamadi, S. and Hansen, J. H. (2015). A study on deep neural network acoustic model adaptation for robust far-field speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Moritz, N., Anemueller, J., and Kollmeier, B. (2015). An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.

Paliwal, K., Schwerin, B., and Wjcicki, K. (2012). Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication*, 54(2):282–305.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Paul, D. B. and Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., and others (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Seltzer, M. L., Yu, D., and Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7398–7402. IEEE.

Sivasankaran, S., Nugraha, A. A., Vincent, E., Morales-Cordovilla, J. A., Dalmia, S., Illina, I., and Liutkus, A. (2015). Robust ASR using neural network based speech enhancement and feature simulation. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 482–489. IEEE.

Van Veen, B. D. and Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24.

Viikki, O., Bye, D., and Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. In *Acoustics, Speech and*

Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, volume 2, pages 733–736. IEEE.

Weninger, F., Geiger, J., Wllmer, M., Schuller, B., and Rigoll, G. (2013). The Munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks. In *Proceedings of the 2nd CHiME workshop on machine listening in multisource environments*, pages 86–90.

Willi, M. (2017). *The Perceptual Significance of a Relative Acoustic Representation of Speech*. PhD thesis, The University of Arizona.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., and Schuller, B. (2017). Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. *arXiv preprint arXiv:1705.10874*.