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 seemingly simple task has been a subject of much study for over half a century, and is only recently gaining much success, it will be important to briefly discuss the reasons for the challenges, traditional ways of dealing with them, and more recent successes.

Despite these successes, challenges still remain when there is noise in the signal. As before, it is important to understand the mechnics and acoustics of why this proves to be a challenge for automatic speech recognition (ASR), and traditional methods of dealing with this as well as more modern techniques.

In the previous chapter, data was collected using a novel technique 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 using a standard open source ASR system with a standard, open source acoustic model.

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 touch on all techniques in this very brief overview. Therefore, only several areas of research in noise-robust ASR will be discussed pertaining to the present research.

[Present the problem with function using x_t and y_t, mention reverberation. Can mention LATER that ear recording elimnates reverberation, introduces OE 'reverberation' (though not really reverberation in the same sense), but (c.f. Ch2) this is highly predictable. Also mention here that phase is rarely accounted for, and is normally assumed out of the equation.]

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

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

where x is the clean speech signal, h is any convolution (ie. room reverberation, microphone channel warp), n is additive noise, and y is the noisy speech signal.

Multi-style training, what it is (training with multiple types of noise) - it is ineffective. Older - paper cited was written in 1987. Li et al. (2014), **?**.

Noise robustness techniques may be categorized by (a) the kind of observation models used, (b) according to whether an explicit or an implicit distortion model is used, and (c) according to whether or not prior knowledge about distortion is employed to learn the relationship between x_t and y_t (Li et al. (2014)).

Feature space is the part of the ASR process where an acoustic vector is transformed into 'features' about the acoustic signal that the ASR model will receive as input. Model space, imperatively, 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 or both domains. In the feature domain, noise is dealt with prior to the sending the features to the acoustic model. These noise reduction processes intend to enhance the signal before it reaches the model. This is mostly done 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 of computation and performance, with Model Domain space alterations generally yielding higher performance improvement (Li et al. (2014)).

In the feature domain space, there are a number of 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 methods of deriving these features have been proposed that incorporate features of the human auditory system, including Perceptual linear prediction (Hermansky et al. (1985)), which aims to replicate **XXXX**, and Relative Spectral processing (RASTA) applied to Perceptual Linear Prediction (Hermansky et al. (1992)), making perceptual linear prediction less sensitive to slow varying speech information, and more sensitive to the quick-varying transitions of speech (**?**). **ALSO Kim et al. (1999), mimicking XXX**. 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)), which mimics **XXXX**, (Moritz et al. (2015)), which mimics **XXXXX**. These methods generally outperform normal MFCC methods. Difficulties utilizing these methods include that it is difficult to determine the parameters. Also the generation of these features can be quite complex, preventing widespread usage with other techniques, and difficulty deriving a relation between clean and noisy speech (which makes explicit distortion modelling difficult).

Feature normalization generally involves normalizing cepstral feature vectors in the form of cepstral mean normalization (CMN) and cepstral mean and variance normalization (CMVN). CMN invovles 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. This is primarily used to eliminate reverberation and channel-related distortion, but signals with noise and no channel distortion also see improvement (Droppo and Acero (2008)). “CMVN normalizes the mean and covariance together” (Li et al. (2014), 752), yielding improved performance on speech data with additive noise. These methods do not work in real-time, however, as they require cepstral vectors from the entire utterance in order to calculate a mean.

Feature compensation actually attempts to remove the noise from the noisy speech signal, allowing for use of traditional features. Spectral subtraction (Boll (1979) is an older and intuitive method of removing noise by turning the linear signal into the spectral domain, subtracting a noise spectrum from a noisy speech spectrum, leaving the clean speech spectrum. This is then converted back into the series of samples that derive the clean speech spectrum. This is performed all along the waveform. The noise can be estimated by looking at sections of the observed signal that do not contain speech information.

This method still comes with several problems (Li et al. (2014)). First and foremost, it is difficult to detect the location of speech in the signal in a noisy environment, which consequently affects the ability to accurately compute a noise average. It also requires relatively stationary, slow-variation noise, as noise that changes quickly very possibly has a different average spectrum during the portion of the signal containing speech than the portion of the signal in which it was calculated. Furthermore, this is only an average of the noise, and not exactly the noise itself; this subtraction 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))

Weiner filtering (**?**) 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 enviroments, as it generally results in suppression and dampening of the entire signal (Li et al. (2014)).

More standard, is the "advanced front-end" (AFE) ensemble proposed in ets (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 Weiner filtering, SNR-dependant waveform processing, and blind equalization (cf. **??**, respectively.

Most of the heavy work is performed by the Mel-warped Weiner filtering (Li et al. (2014)). This differs from the more standard Weiner filter in that it uses

Mel Weiner filtering???, which is performed once, and then a second time to remove residual noise. SNR-dependant waveform processing (SDWP) assumes that the noise is relatively constant, whereas the speech signal will cause variation in the amplitude of the signal. SDWP uses this assumption to dampen portions of the signal with a relatively low SNR (ie, speech-less) compared with the high SNR (ie, speech-bearing) portions of the signal, which are boosted. Blind equalization serves to eliminate convolutional (ie. reverberant) distortion from the signal.

Model domain compensation usually involves adapting an existing acoustic model (presumably trained on relatively clean speech) to enable recognition of more noisy features. Most popularly, variations of maximum likelihood linear regression (MLLR, Leggetter and Woodland (1995)) are used to adjust the gaussian component vector means and the covariance parameters. A slight variation in the caculation allows these transforms to be applied to the features themselves (feature-space MLLR, or fMLLR, Gales (1998)), moving this method into an on-line feature space tranformation.

Another tool used is the exploitation of any prior knowledge about the distortion (Li et al. (2014)); this is prior knowledge that is utilized during the training stage, not knowledge about the noise during the testing stage. Some methods include learning the mapping between noisy and non-noisy pairs of acoustic signals. This mapping is then extended to novel noisy utterances during testing. This is used in feature domain space to enhance a noisy speech feature to then send to the model.

Other methods utilze mutliple acoustic models, each trained on data from different environments and different noises and different SNRs. The means and covariance matrices of each of these models are stored, and during recognition, the most appropriate model is chosen to use to decode the signal in question. Either of these tactics, though, do require prior knowledge about the noise. As with multi-style training, explained above, it is very difficult to ensure that all noises, SNRs, etc, are adequately accounted for during training in order to be prepared for what is seen during testing.

Explicit distortion modelling uses a “physical model” which allows for high performance with few distortion parameters. An example of an explicit distortion model would be spectral subtraction, discussed earlier. It seems obvious that spectral subtraction, when matched with an agreeable signal that best utilizes its noise removal abilities, would result in more accurate speech recognition. Consequently, other noise reduction methods that *explicitly* specify the distortion tend to perform well.

[**Read Paliwal et al. (2012), find out where to put it in above discussion, introduces MMSE (minimum mean square error). Zhang et al. (2017), 4 that it is still considered to be the best "general purpose" noise-removal tool.**]

There are two primary categories of utilizing neural networks to account for noisy speech, "mapping methods" and "masking methods". Mapping methods primarily fall underneath the umbrella of feature-space methods. This involves 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. DNN, CNN, RNN, 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 later input transformations, such as from the spectral or cepstral domains (Zhang et al. (2017)).

Masking-based approaches work similarly to a traditional filter, albeit learned via a neural network. These are also performed as a feature-space tool. An 'Ideal Ratio Mask' will learn the ratio (value between 0 and 1) of the clean speech to the noisy speech. The function can be learned - via an neural network architecture - between the noisy speech signal and the masking value. Most beneficial when using spectral or cepstral features as input, a masking value is learned for each set of features. These masking values are then multiplied elementwise to each feature in the set, at that time index within the signal, ie. a method very similar to that of a traditional filter (Zhang et al. (2017)).

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 **?**, uses an collection of training data which exhibits a wide range of noise conditions. Another method, 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) has 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, instead of modifying all weights. This largely avoids the issue of overfitting, while increasing the model's robustness to noise.

Much work is also turning to combine modifications in the feature space domain, with modififications in the model space domain (Weninger et al. (2013)). Most simply, this takes the form of using the feature-enhanced data output from feature-based noise removal as training data itself for the acoustic model.

There are also techniques that employ multiple microphones as source-separation technique to separate the speech source from any extraneous noise sources. The study at hand does not use multiple microphones, and so this will be discusses as only a reference. Beamforming (Van Veen and Buckley (1988)) has become a central technique to using microphone arrays in source separation (Hori et al. (2015); Zhang et al. (2017)). This is done by calculating the direction of arrival of the different sound sources, 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 the microphone. Neural networks have also been employed to aid and enhance the beamforming process (**?**Sivasankaran et al. (2015); **?**).

1.3 Experiment 2: 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 acheive 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 to what could be recorded at the mouth (below 2.7 kHz).

1.3.1 Stimuli

Recordings from twenty speakers, ten male and ten female, from the data collection experiment in Chapter 2?? were used as test data for this experiment. This included 30 distint sentences from each speaker, each with 5 different noise conditions (bus, cafe, pedestrian, street, factory), 3 different noise levels (60dB, 70dB, 80dB), and a ‘clean’ (no noise) condition. This results in 16 iterations of each distict sentence, for each speaker, totalling 480 sentences per speaker, 4800 sentences for each gender group, and 9600 total test sentences.

1.3.2 Design

The existing **OPEN SLR** acoustic model will be used to test the collected data. **This acoustic model has not been trained on significantly noisy data**. This will primarily test the performance between the ear-recorded and noisy mouth-recorded speech, but also between the different noise conditions and noise levels. Both the ear-recorded and noisy mouth data will then be enhanced using the well-established advanced front end (AFE, ets (2002)) technique, and will be retested on the same, unchanged acoustic model.

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

It is quite possible that, due to the ear-recorded speech only containing information below 3 kHz, that the existing acoustic model in its current state will have poor recognition of these sentences. If this is the case, the same acoustic model will be adapted with ear-recorded and low-pass filtered additional setnences (not from the 30 test sentences, nor from any of the speakers being tested). A total of **XXX** distinct sentences from **XX** additional speakers were used for adaptation of the acoustic model, totalling **XXX** sentences used for adaptaiton. The same 9600 sentences from the same 20 speakers were used again for test data.

1.3.3 Procedure

1.3.4 Results

1.4 Discussion

1.4.1 Limitations and Future Research

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 (amplitude) of noise the proposed method can handle. In theory, as has been shown in Chapter 2**??** 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 hence the SNR of the speech recorded at the mouth) should have no effect on the speech recorded at the ear. Nevertheless, this should be investigated. 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 participants' ears, with amplitude varying noise.

REFERENCES

(2002). ETSI. *Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms.*

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.

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

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.

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.

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.

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.

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.

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*.