On the Use of Spectro-Temporal Features in Noise-Additive Speech

by

Suman Ravuri

A thesis submitted in partial satisfaction of the
requirements for the degree of
Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nelson Harold Morgan, Chair
Andreas Stolcke, Ph.D.

Spring 2011

On the Use of Spectro-Temporal Features in Noise-Additive Speech

Abstract

On the Use of Spectro-Temporal Features in Noise-Additive Speech

by

Suman Ravuri

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Nelson Harold Morgan, Chair

Most extant features attempt to model how humans hear in some way. MFCCs, for instance, mimic the actions of the basilar membrane by triangle filter integration on the mel frequency scale, and emulate the hair cell stages with log compression. Most of this auditory inspiration has been limited to very early stages of human hearing, since beyond that point requires understanding parts of the brain, which up until this point has proven impossible.

Recently, however, a number of experiments of the primary auditory cortex of ferrets (which share many of the same features as the human version) have shown that individual neurons are tuned to fire at one particular spectral and temporal modulation. Incorporating these features could help automatic speech recognition, since previously this sort of parameterization has not been considered in existing features.

The use of these spectro-temporal features has been shown to improve speech recognition performance in a variety of tasks. In this work, I investigate the performance of spectro-temporal features under noisy conditions. In particular, I study the performance of spectro-temporal features in mismatched conditions, in which the training set consists of only clean data and the test set comprises of speech added to a variety of noises.

Although spectro-temporal features can be shown to improve recognition performance in this task, there exist a number of hurdles one must overcome in order to incorporate these features into a speech recognition system. One major problem is high dimensionality - consisting of possibly tens of thousands of dimensions - of spectro-temporal features, which

is not compatible with the standard HMM recognizer. Part of this problem is circumvented by performing discriminative training via multi-layer perceptrons. Since the number of dimensions is too high to perform adequate training on a multi-layer perceptron, one must determine an alternative method to fixing this problem. One way to handle high dimensional input, which is used in this work, is to incorporate many MLPs.

By using many multilayer perceptrons, one must determine both how to segment the features prior to discriminative training, and combine the outputs of the multilayer perceptron. For segmentation, I study two methods. The first is segmentation into multiple spectral and temporal modulations feature streams, called "multi-modulation spectral features", that are either consistent with biological findings or are shown to perform well on a given task. Moreover, I also introduce spectro-temporal mel-scaled cepstral coefficients and show how they outperform the standard multi-modulation spectral feature segmentation. To combine stream, I investigate various static and dynamic frame-level combination methods. On the Aurora2 corpus, compared to Advanced Front End features, spectro-temporal features combined with Advanced Front End processing reduced WER by over 50% and 18% in clean and noisy cases respectively. The improvement over MFCCs is even more dramatic, reducing WER by 50.0% and 68.2% in clean and noisy cases respectively.

To Sarah Downs, Sreenivas Ravuri, and Padmaja Ravuri

Contents

List of Figures

List of Tables

Acknowledgments

Although only my name is ultimately on this report, there are many who contributed to this work so deeply that this report would not have been finished without their support. First and foremost, I would like to thank Prof. Morgan for all the helpful technical advice and ideas. Moreover, I would like to thank Andreas Stolcke and Adam Janin for patiently answering my patently dumb questions. Finally, I would like to thank Arlo Faria for helping me understand how to run a speech recognition system.

I would be remiss if I did not also acknowledge those who provided strong moral support. I am especially indebted to Sarah Downs, who with bodhisattva-ish patience listened when I complained about the differences in HTK MFCC vs. Motorola MFCC calculation and its performance on the HTK speech recognizer (riveting stuff, I know). I must also thank my parents, Padmaja and Sreenivas Ravuri, who did not try to have me committed when I rambled on like a crazed man about my research. Their moral support truly is "sine qua non".

Chapter 1

Introduction

Most front-end research has focused on generating one carefully-crafted feature that balances incorporating what we know about the human auditory system with adhering to the structure that offers the best performance with the standard decoder. Out of such research came features such as the now ubiquitous mel-scaled cepstral coefficients (MFCCs) ([6]), perceptual linear prediction (PLPs) ([14]), and many others. In very clean, close microphone, read speech, these features perform very well. Moreover, since each of these features capture characteristics of human hearing slightly differently, we can often improve performance by combining features or systems that each use a different feature.

Despite the high performance of these systems in limited cases, in a number of other cases, such as noisy speech, speech captured from a far-field microphone, or conversational speech, performance of ASR systems is at best middling. Work on modifying these features for particular cases have enjoyed some success, but typically these features are not robust to different conditions. What seems to have worked, however, is again combining feature sets, since certain features may be robust to certain adverse conditions while others may perform better in other cases. Generally, at most 8 to 10 of these feature sets are included in a high-performance ASR system.

A new paradigm that is emerging in the speech community is using tens or hundreds feature streams that while not as carefully crafted as its predecessors, hope to achieve better robustness. Instead of a one-size fits all paradigm, one hopes to find some subset of features that are not corrupted by noise, reverberation, or other issues that have plagued normal

features. Moreover, these features incorporate new information on what people have learned about the auditory system. In this paper, I describe a certain set of "biologically-inspired" features called spectro-temporal features.

Implicit in this work is that we must balance what characteristics we would like to add in these features with creating a structure that gives us the best performance on the ASR task. As we will see in the next section, simply inputting raw spectro-temporal features into a standard HMM-GMM system is impossible because of the high dimensionality, but in trying to decoct the spectro-temporal features and transform them into a form useful for ASR, we break some ancillary characteristics that we think a priori are important. In this case, we take the view that better machine performance is more important than better intuitive understanding.

Finally, this report is by no means meant to be an exhaustive report on spectro-temporal features. Instead, we focus on a particular class of problems, that of mismatched noise in training and test. That is, when we only have clean training data and noisy test, we illustrate how spectro-temporal features can improve robustness. Moreover, as explained in this report, we show how specto-temporal features improve robustness when combined with another noise-robustness algorithm based on Wiener filtering.

1.1 Spectro-Temporal Features

Experiments in the 1990s and 2000s with the primary auditory cortex of mammals such as ferrets, cats, and monkeys have changed the way researchers view the way humans hear. In particular, researchers in have noticed the spectro-temporal receptor fields (STRFs) of neurons of these mammals, which share the same auditory cortical structure as humans, are particularly sensitive to certain spectro-temporal patterns of sound. Figure 1.1 shows one particular STRF.

These spectro-temporal receptor fields were calculated in one of two ways. One method involved measuring the output of a neuron of the primary auditory cortex when the auditory stimulus was a dynamic ripple, parameterized in the form:

$$S(t, \hat{f}) = \Delta A \sin(2\pi \cdot \omega_t \cdot t + 2\pi \cdot \omega_{\hat{f}} \cdot \hat{f} + \Phi)$$

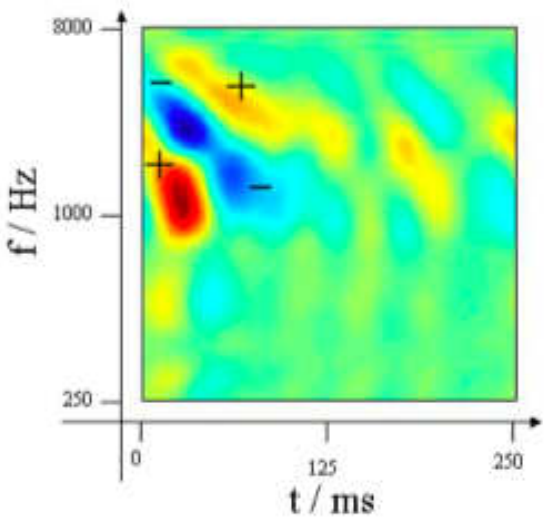

Figure 1.1: *An example of a spectro-temporal receptive field (STRF) of a neuron from the primary auditory cortex of a ferret. The red and yellow regions(+) represent active regions while blue regions(-) represented inhibited response. Taken from [22].*

Spikes were measured at different spectral and temporal modulations (represented by $\omega\hat{f}$ and ω_t respectively, where $\hat{f}$ is a the log-scale) at different amplitude levels (denoted by ΔA). The position on the graph represents the different spectral and temporal modulations, while the intensity depended on the amplitude. [8], [32], [3], and [23] performed these type of experiments on ferrets.

Another way to measure the STRFs is to use an arbitrary signal, measure spike outputs, and reverse correlate the output with the input to generate a transfer function that represents the STRF. [21] and [24] performed this type of experiment on ferrets, while [7] did that for monkeys.

One interesting thing to note about these STRFs is that they capture truncated diagonal slices of the time-frequency plane and these active regions seem to occur in ripples. This sort of processing, however, is not captured by standard MFCC or PLP features, so perhaps one can improve speech recognition by including this information.

An more important property of this model is that at each particular time, there are thousands of STRFs capturing different spectral and temporal modulations. Figure 1.2 illustrates this point. In typical features such as MFCCs or PLPs, the rate and scale dimensions do not exist.

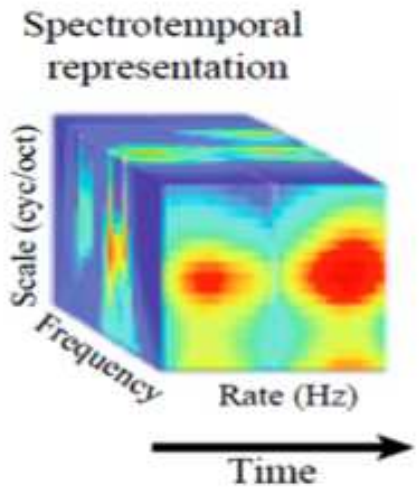

Figure 1.2: *An example of a spectro-temporal receptive field (STRF) of a neuron from the primary auditory cortex of a ferret. The red and yellow regions(+) represent active regions while blue regions(-) represented inhibited response. Taken from [27].*

In trying to implement this type of response, we must first determine how to approximate the STRFs with filters. While the filtering is indeed different from that done by traditional features, there already exists two parts of the filtering in extant features that closely approximate STRFs, the mel-warping in the frequency domain and the log amplitude filtering. Much of the filtering performed by mel-scaled cepstral coefficients is similar to the spectro-temporal feature processing in this respect and in fact, I use the same log mel-filterbanks used in MFCC calculation for the proposed spectro-temporal features.

For the MFCC, the processing from the log mel-filterbank to the final feature is a simple inverse discrete cosine transform. This processing, however, does not produce the diagonal ripple response needed to approximate the STRFs. [22] proposed approximating the STRF using a 2-D Gabor filter on the log mel-filterbank. The parameters of the filter are as follows: ω_t (the temporal modulation), ω_f (the spectral modulation), σ_t (the time variance of the Gabor parameter), and σ_f (the frequency variance of the Gabor parameter). The output is as follows:

$$S(f_0, t_0, \omega_f, \omega_t) = \sum_t \sum_f C(f, t; \omega_f, \omega_t) G(f, t; \sigma_f, \sigma_t) \tag{1.1}$$

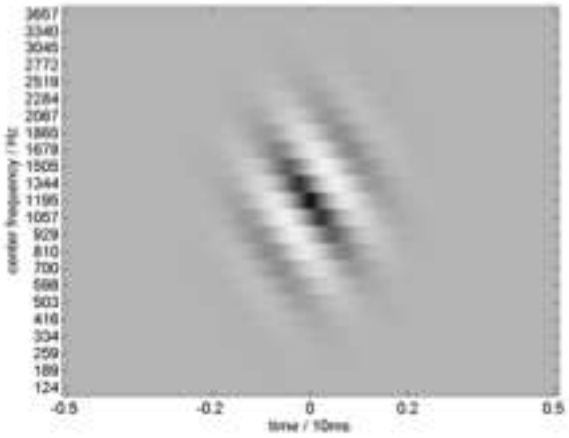

Figure 1.3: *The real part of a 2-D Gabor filter, taken from [22].*

where

$$C(f,t;\omega_f,\omega_t) = \exp(i(\omega_f(f-f_0)+\omega_t(t-t_0)) \tag{1.2}$$

$$G(f,t;f_0,t_0,\sigma_f,\sigma_t) = \frac{1}{2\pi\sigma_t\sigma_f}\exp(\frac{-(f-f_0)^2}{2\sigma_f^2}+\frac{-(t-t_0)^2}{2\sigma_t^2}) \tag{1.3}$$

According to the proposal, one can take the real part, imaginary part, or magnitude of the output of the Gabor filter. Also, in practice, we truncate the summation in time and frequency.

With this 2-D Gabor filter, we are able to approximate the spectro-temporal receptor fields from above. Figure 1.3 shows a particular Gabor filter function. Since I use Gabor filters to approximate STRFs, I will use the term Gabor features interchangeably with spectro-temporal features.

1.2 Related Work

Experimentally, these features have successfully been applied to a number of speech recognition and discrimination tasks [22, 26, 9, 17, 34, 35]. In particular, [35] demonstrates that spectro-temporal features perform quite well in automatic speech recognition under

noisy conditions. I surmise that the spectro-temporal feature calculation, which filters the log mel-spectra to emphasize many different spectral and temporal modulations, is able to emphasize components of the time-frequency plane that are usable for speech recognition, even if other sections are corrupted. This framework tends to generate many more features than are typically used in ASR, many of which may be highly correlated with one another.

One challenge in incorporating these spectro-temporal features is that their extremely high dimensionality may cause difficulties in the GMM observation structure of standard HMM-based systems. The example given above was for only one spectral and temporal modulation, but in theory there could be thousands of these features. So, there must be some way of bridging the gap between the biological inspiration, which leads to extremely high dimensional features, and the statistical framework of the HMM-GMM system, which as a rule of thumb can only handle dimensionality of at most 100.

Many people have proposed different methods to reduce the dimensionality of spectro-temporal features for different tasks. [22] suggested applying a feature-finding neural network for feature selection. In this algorithm, a neural network learns an optimal feature set by replacing an input feature with a randomly-drawn one until the net finds the feature with the smallest increase in classification error. [9] employed a winner-take-most approach which suppresses the least active spectro-temporal neurons in their speech recognition system. A completely different method, used in [26], has been quite successful in automatic speech/non-speech discrimination; Mesgarani et al. extended classical Principal Components Analysis (PCA) to multidimensional tensors in order to reduce feature dimensionality.

An alternate approach partitions spectro-temporal features into different streams, individually processes each stream, and then merges the processed streams prior to inputting them into the decoder. [17, 34, 35, 33] use this approach successfully with spectro-temporal features; each of their systems decrease WER in ASR in clean and noisy conditions. This approach has also been used in speech recognition systems prior to the advent of Gabor features, such as in multi-band systems (see [2, 15]) and in systems incorporating temporal critical bands and PLP (see [29]).

The advantage of using this multi-stream approach is that streams can be divided according to psycho-acoustic and physiological findings. Moreover, since the streams are considered independent, we can process the streams in parallel. The disadvantage of this approach is

that in general the dimensionality of the streams must be reduced in some way and a principled way of merging the processed streams must be determined.

The neural-network-based Tandem approach, originally proposed in [16], proves an effective way of reducing the feature dimensionality of the stream. Merging streams, however, still remains an open question. Previous work has considered two different options: weighting the feature streams such that the weights are static across the frames; and weighting the streams dynamically based on either some metric or learning algorithm. [35, 28] weighted streams through inverse entropy and learned weights using a weight-generating multilayer perceptron, while [33] used a hierarchical neural-network-based weighting system to combine features.

Chapter 2

Exploratory Work

This section contains some preliminary experiments that motivate later work. These sets of experiments do not necessarily use the same data set as the later sections in order to try to reduce tuning on the test set. Moreover, these experiments are by no means exhaustive, but demonstrate the impetus for certain design choices in following sections.

2.1 Advanced Front-End vs. Spectro-Temporal Features

Using redundant features as in the spectro-temporal framework is one way of reducing noise; it is, however, certainly not the only method. An alternative approach is the two-stage mel-warped Wiener filter first introduced in [1] to denoise the input signal prior to feature computation, an improved version of which was later implemented in an MFCC-like feature called "Advanced Front End" (AFE), described in [12]. In addition to a more refined two-stage mel-warped Wiener filter (known simply as "noise reduction" in the original document), AFE includes a waveform processing component that incorporates waveform smoothing and peak picking, a mel-cepstrum calculation, and a blind equalization step. AFE can be considered an MFCC feature with extra denoising algorithms. The top pane of Figure 2.2 represents a vastly simplified diagram of the AFE processing. The bottom pane illustrates the stages of the cepstrum calculation for later reference of the description of the

proposed system.

Work on the AFE processing can be traced backed to [1], which first introduced a two-stage mel-warped Wiener filter for reducing noise in the signal. On the Aurora1 data set with clean training and noisy test sets (which contained 4 different noises at 7 noise conditions each), the MFCCs extracted from this denoised signal reduced the average WER by 32.5% relative to MFCCs calculated on the original signal. The AFE in [12], mentioned in the previous section, employed an improved version of the two-stage mel-warped Wiener filter in its processing. AFE's implementation differs from that of previous work by applying a more sophisticated noise estimation algorithm in the first stage and adding in the second stage an extra step to more aggressively filter purely noisy frames and less aggressively filter speech frames. On the Aurora2 corpus with clean training and noisy test sets, AFE reduced the average WER by 49.1% compared to MFCCs.

In this experiment, spectro-temporal features are compared to AFE. The spectro-temporal features are divided into four streams which include features with different spectral and temporal modulations, and the magnitude of the output of spectro-temporal filters was taken. The spectral modulations for these streams range from 0.04 inverse channels to 0.5 inverse channels while the temporal modulations range from $\pm$6Hz to $\pm$50Hz. Table 2.1 summarizes the breakdown of the four streams.

The output of the spectro-temporal processing is segmented into the four streams and each stream is inputted into a multilayer perceptron. The input layer contains 9 frames of context, so the size of the input layer is 9 times the number of features in the stream. The neural network also contains 160 hidden units, and 56 output targets (each corresponding to an English phone).

The outputs of the MLP stream provide an estimate of the posterior probability distribution for phones. We combine each of these phone probability estimates across streams by equal weighting per frame. We then apply Karhunen-Loève Transform to the log-probabilities of the merged MLPs to reduce the dimensionality to 32 dimensions and orthogonalize those dimensions. We then mean and variance normalize the features by utterance. Finally, we append this spectro-temporal feature vector to the MFCC feature. This serves as the observation to HTK decoder, the setup of which is described in Chapter 5. The data set for this task is Aurora2 noisy digits corpus, details of which are also outlined in Chapter 5.

Feature Stream	No. of Features	Spectral Mod.(chan^{-1})	Temporal Mod.(Hz)
Stream 1	506	0.04, ... ,0.5	±50
		0.04	± 25
		0	20, 25, 33.3, 50
		0.04, . . . , .14	0
Stream 2	506	0.13, ..., 0.5	±25
		0.04, 0.13	±14
		0	11.1, 12.5, 14.3, 16.7
		.16, . . . , .26	0
Stream 3	506	0.24, 0.36, 0.5	±14
		0.04, 0.13, 0.24	±9
		0	7.7, 8.3, 9.1, 10
		.28, . . . , .38	0
Stream 4	529	0.36, 0.5	±9
		0.04, ..., 0.5	±6
		0	6.2, 6.7, 7.1
		.4, . . . , .5	0

Table 2.1: *Range of spectro-temporal modulation frequencies captured by each of the 4 feature streams.*

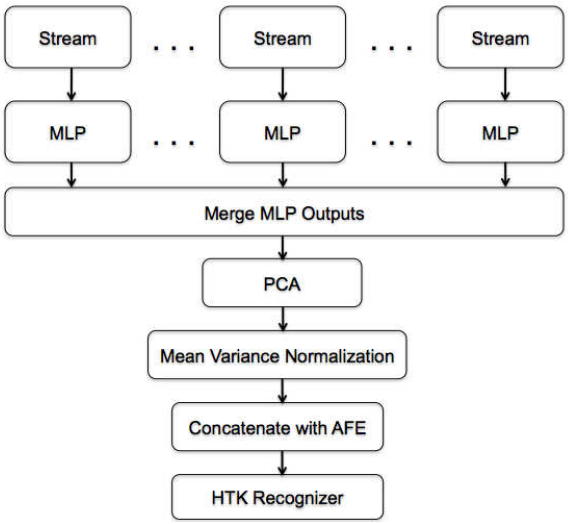

Figure 2.1: *Diagram of processing of the MLP streams.*

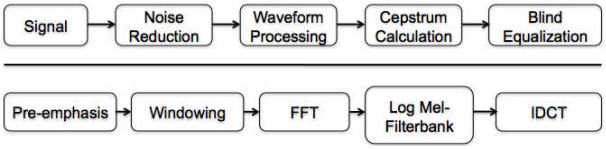

Figure 2.2: *Top Pane: Diagram of AFE feature calculation steps. Bottom Pane: Diagram of the steps in the cepstrum calculation.*

As seen in Table 2.2, while appending spectro-temporal tandem features to the MFCC[1] reduced the WER by 34.9% relative in the 20dB to 0dB SNR average WER, AFE features reduced WER by 65.7%. This result suggests that the Gabor features show some promise, but that the Advanced Front End features significantly outperform spectro-temporal features.

One caveat with the numbers, however, is that AFE is highly tuned to the noise-added conditions of the Aurora2 test set and the parameters of the HTK recognizer. To illustrate this point, I compared the same set of features (with the exception that the spectro-temporal features are averaged in the log-probability domain) on the NUMBER95 database. This corpus contains 32 numbers extracted from telephone conversations of American-English speakers. The training set consists of 3 hours of clean data, while the test set for the clean and each of the noisy conditions comprises 1 hour and 10 minutes test data respectively.

[1]Using only spectro-temporal features significantly increases the WER of the system.

SNR	MFCC (baseline)	AFE	MFCC+ Spectro-Temporal
clean	1.60%	1.63%	1.16%
20dB	5.33%	2.59%	2.45%
15dB	14.77%	4.25%	5.56%
10dB	36.59%	8.27%	15.75%
5dB	66.65%	18.74%	41.45%
0dB	86.98%	42.18%	75.19%
-5dB	94.01%	73.12%	96.68%
20-0dB Avg.	42.06%	15.20%	28.08%

Table 2.2: *Comparison of recognition results of MFCC baseline to AFE features and spectro-temporal features appended to MFCCs. The spectro-temporal features in this experiment are equally weighted per frame.*

SNR	MFCC (baseline)	AFE	MFCC+ Spectro-Temporal
clean	2.94%	4.65%	2.22%
20dB	4.72%	4.97%	4.59%
15dB	9.22%	8.72%	7.72%
10dB	9.95%	9.32%	8.31%
5dB	24.74%	19.77%	22.07%
0dB	39.43%	33.58%	35.70%
20-0dB Avg.	17.66%	15.32%	15.72%

Table 2.3: *Comparison of recognition results of MFCC baseline to AFE features and spectro-temporal features appended to MFCCs on the Numbers95 dataset. The spectro-temporal features in this experiment are geometrically weighted per frame.*

Noise was added at different signal-to-noise ratios from the RSG-10 collection using the same methodology as [18] for creating Aurora2 task. The recognizer used for this setup is SRI's DECIPHER. The MFCC and AFE features were VTLN normalized, while all features are mean and variance normalized per speaker. The acoustical model uses gender-independent, within-word triphone Hidden Markov Models (HMMs).

The results for this experiment (shown in Table 2.3) demonstrate that MFCCs seem unnecessarily impaired in the Aurora2 task using the HTK recognizer, that AFE exhibits much more modest improvements than in the Aurora2 task, and that MFCCs+spectro-temporal features outperforms Advanced Front End features in all but the noisiest conditions.

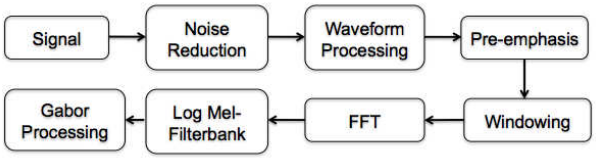

Figure 2.3: *"AFE" and Gabor processing of the input signal.*

This suggests that spectro-temporal features exhibit some robustness to a variety of test sets.

Although the NUMBERS95 setup uses a recognition system more likely to be used in the real world, many report results using this Aurora2 setup. Due to this standard, I use the Aurora2 setup for this thesis. Moreover, since AFE processing is complementary to the spectro-temporal one, one could combine much of the AFE processing with spectro-temporal feature extraction and sidestep the issue of using MFCCs in a hampered environment. Figure 2.3 illustrates this proposed feature calculation. We execute the entirety of the noise reduction and the waveform processing steps of the AFE algorithm; then carry out all but the inverse DCT of the cepstrum calculation, thereby leaving us with log mel-spectrogram for use with Gabor filtering; finally, we apply Gabor filtering on the log mel-spectrogram. By incorporating this filtering after execution of AFE processing, we combine two vastly different methods for reducing noise. In the next two chapters we will outline the spectro-temporal divisions and Tandem method used for this paper.

Chapter 3

Spectro-Temporal Divisions

One major hurdle for incorporating spectro-temporal processing into a HMM recognition system is the high dimensionality of the features. In particular, as is shown in Figure 1.2, each time step generates a three-dimensional block of features. This block could possibly contain hundreds of thousands of dimensions while canonical features such as MFCCs use only 23 (corresponding to the 23 mel-channels calculated from a speech sample). Since a Tandem approach is employed (described in the following chapter), one cannot necessarily use the whole block as input to a single multilayer perceptron; thus, features must be split into multiple components. How one actually divides such streams is an open question. In this work, two alternative methods for dividing spectro-temporal features are proposed. The proposed feature divisions try to balance incorporating all the new knowledge gained from biological findings with what actually works in a speech recognition system.

3.1 Multi-modulation Spectral Features

One way to divide spectro-temporal features is to separate them into streams that match some prior knowledge of the structure of the spectro-temporal features or perform well on some speech recognition task. The division used for this paper is shown in Table 2.1. This division was taken from [34], which found the range of spectral and temporal modulations used from experiments in [3] and [19] and segmented streams because such a segmentation performed well on the aforementioned Numbers95 task compared to simple spectral or tem-

poral modulation division. In [34], only the magnitude of the output of the spectro-temporal features were investigated, while in this study, we investigate the performance or real and imaginary components as well.

This multi-modulation spectral setup certainly weighs adherence to biological findings more heavily than computational concerns, since the number of feature dimensions in each stream is far higher and those dimensions are far more correlated with each other than can be used in an HMM recognizer. The hope in this setup is that discriminative training with MLPs will bypass these problems.

3.2 Spectro-Temporal MFCCs

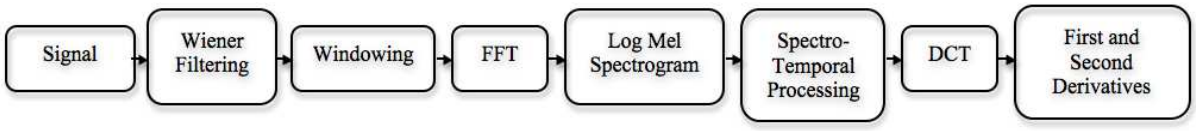

Figure 3.1: *Diagram of steps of spectro-temporal MFCC processing.*

One problem with the multi-modulation stream approach, especially in [34, 35, 30], is that modulations included in a particular stream are generally hand-tuned and the structure of streams that perform well for one task may not be optimal for another. Moreover, multi-modulation streams cannot take advantage of a number of normalizations (such as vocal-tract length normalization and cepstral-mean subtraction) that have been developed for standard features such as MFCCs or PLPs. Furthermore, the structure of the multilayer perceptrons used to train such streams is far different than in previous Tandem approaches (since the current structure has a large input layer compared to the hidden layer, while previous approaches have a smaller input layer compared to the hidden layer), and the resulting features may be significantly suboptimal. Finally, from a more abstract point of view, even though spectro-temporal features share much of the same feature calculation as MFCCs, one cannot reproduce this canonical feature under the spectro-temporal framework, suggesting that the current spectro-temporal framework is perhaps not as flexible as originally thought.

Spectral Mod.(chan^{-1})	Temporal Mod.(Hz)
0.04, 0.13 , 0.24, 0.36, 0.5	$\pm$6, $\pm$9, $\pm$14.2, $\pm$25, $\pm$50
0.04, 0.06, 0.08, ..., 0.48, 0.5	0
0.00	6, 6.7, ..., 33.3, 50

Table 3.1: *Range of spectro-temporal modulation frequencies used for spectro-temporal MFCCs.*

Spectro-temporal MFCCs address these aforementioned problems. To motivate the idea of spectro-temporal MFCCs, consider spectro-temporal processing as shown in Figure 3.1. If we were to consider spectro-temporal processing at one spectral and temporal modulation, one could consider the output prior to the DCT as a bandpass-filtered version of the log mel-spectra. Removing mean subtraction at the output of spectro-temporal processing (not shown in the diagram) allows us to preserve original the log mel-spectrogram from original MFCCs when the spectral and temporal modulations are only calculated over one window in the time and channel domains. Taking a discrete-cosine transform results in features that are "MFCC-like" in their structure. As such, they can be used as drop-in replacements for MFCCs in order to determine which spectral and temporal modulations perform well for different types of noise. Figure 3.2 shows performance of spectro-temporal MFCCs at different SNRs under Airport noise, while Appendix A includes these figures for all noises included in this study. Moreover, since one expects that different spectral and temporal modulations perform better for different types of speech, I will introduce a method to combine many such these spectro-temporal MFCCs to lead to more robust features. In this framework, each feature set contains only one spectral and temporal modulation. Thus, instead of four multi-modulation streams, there are 89 multi-modulation streams that are each discriminatively trained by a multilayer perceptron.

The downside to this sort of approach is that no single multilayer perceptron has information to more than one spectral and temporal modulation. This may lead to suboptimal predictions of the correct phone in any given MLP, but the hope is that the combination of the outputs will lead to better predictions and better features.

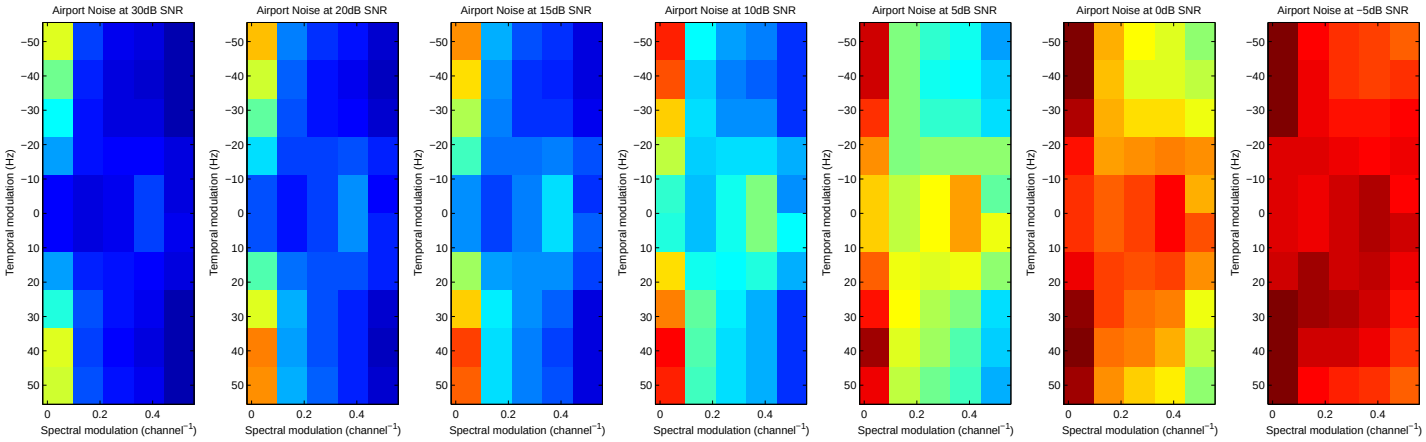

Figure 3.2: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Airport noise.*

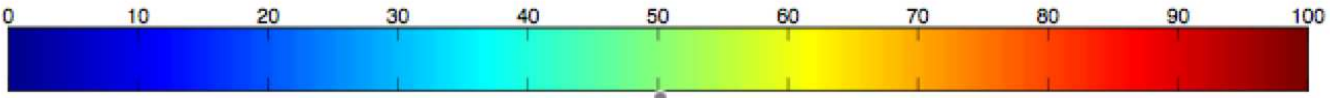

Figure 3.3: *Colorbar for WER for different spectro-temporal MFCCs.*

Chapter 4

Tandem Setup

A major problem for the spectro-temporal framework is that the structure of the features generated do not match the structure which works well for the standard HMM system. The standard system typically handles 13-100 dimensional features, while the proposed feature setup requires any system to handle possibly hundreds of thousands of dimensions of features. Moreover, typical HMM systems also require the observations to not be too highly correlated with each other, while spectro-temporal features may be very correlated with each other.

We sidestep this problem by adding an extra layer of feature computation by using discriminative training with multilayer perceptrons. In this approach, some set of raw features are used as an input to the MLP, and those frames are discriminatively trained on 56 phone targets. This sort of discriminative training provides the dimensionality reduction required for an HMM system, while performing the Karhunen Loève transform on the output of the MLP provides the necessary decorrelation. This Tandem approach has been used with many other features with good success.

Despite the advantages of the Tandem system, some additional hurdles still exist. The first is that the number of input features is generally too high for a single multi-layer perceptron. Roughly 60,000 input dimensions are possible in the both the stream and spectro-temporal MFCC framework, and this number could grow exponentially in future work since only a small subset of possible filters are considered in this study. Given this problem, one must divide the features in some way, perform discriminative training on each of these divisions, and combine their outputs. Chapter 3 addressed the first issue, while this chapter

aims to address the second and third.

4.1 MLP structure

Depending on the method of spectro-temporal division, the structure of the multilayer perceptron vastly differs. In the stream setup of multilayer perceptrons, there are roughly 4,500 input units for the multilayer perceptron, corresponding to the roughly 500 features per stream with 9 frames of context. Since the dataset used contains only 6 hours of training data, the size of the hidden layer is relatively small, 160 units. Completing the structure is 56 output targets, with each target corresponding to an English phone.

For the spectro-temporal MFCCs, the multilayer perceptron structure is more standard. Each feature consists of 13 cepstral features plus first and second derivatives. Nine frames of context make the input layer size 378 units. 480 hidden units and 56 output units round out the structure.

4.2 Combining Real, Imaginary, and Magnitude Spectro-Temporal Features

Recall from Equation 1.1 that the output of a spectro-temporal filter gives us a complex output, and either the real part, the imaginary part, or the magnitude of the output is taken. If one is to incorporate more than one type of output, how to structure the multilayer perceptrons is not entirely clear. In this work, two different approaches are investigated. The first is to combine outputs at the input layer. For example, if I were to combine real and imaginary parts of the spectro-temporal features, I would include both as input into a multilayer perceptron, doubling the input layer. This is called early fusion (EF) in this work and two separate systems (combining real and imaginary; and real, imaginary, and magnitude) are studied in this thesis. The other method is to train a separate MLP for each type of feature, and try to combine the outputs of the multilayer perceptrons in some way. This is denoted as late fusion (LF).

4.3 Combination Methods

In this work, six frame-level combination methods are studied. Three of these methods are static in that the weight for each stream does not change for each frame, while the other half is dynamic, since the weights for each stream may change depending on the frame. The static methods are rather straightforward; the combinations studied are the arithmetic, geometric, and harmonic means of the outputs of the MLPs. Since the outputs of the MLPs can be considered an probabilistic estimate of a phone for a given frame, these static methods provide a way of combining the probability estimates of the phones. The output of arithmetic mean can also be considered a probability distribution, whereas the outputs of the geometric mean and harmonic mean are no longer probability distributions. Combined with processing described in the next section, however, geometric and harmonic means can be considered, up to a constant, to be arithmetic means in the log-probability and inverse-probability domains.

In addition to the static methods, three methods for dynamic weighting (i.e. weights that change at each frame) are studied. The first of these is inverse entropy weighting. For each stream i, an entropy of the output posteriors at frame f, denoted as $entropy_{if}$ can be calculated. Then, the weight for stream i at frame f, w_{if}, is calculated as

$$w_{if} = \frac{1/entropy_{if}}{\sum_{j=1}^{n} 1/entropy_{if}}$$

Ideally, we would like to pick out the best or best set of streams in a given frame. Inverse entropy weighting serves as a proxy for picking out the best stream, since if posteriors provide an accurate probability estimate of a given phone, then inverse entropy weighting more highly weights better streams. Unfortunately, MLPs sometimes incorrectly identify the correct phone, and in those cases do so with high confidence. A better way would be to try to predict, using a learning algorithm, the best stream given the information such as entropy.

A weight-generating MLP is one way to approach this problem. In this framework, a weight-generating MLP tries to predict the best stream given the entropies of each of the streams and AFE input (with 9 frames of context). Here, the output stream is calculated by $\mathbb{P}(phone|input) = \sum_{stream} \mathbb{P}(phone|stream, input)\mathbb{P}(stream|input)$, where $\mathbb{P}(stream|input)$ is the probability of the best stream learned by a weight-generating MLP. In order to train

such a multilayer perceptron, labels for the best stream are needed. Since this is not known a priori, labels are generated by picking the frame which had the highest posterior of the correct frame. Ties are broken by the highest product of utterance and total performance, although in reality ties affect very few frames (under 400 of 1.4M frames).

The structure for the weight-generating MLP with stream divisions was 387 input units (corresponding to 4 entropies and 13-dim AFE features with first and second derivatives, both with 9 frames of context), 2 hidden units, and 4 output units (1 for each stream), while that for the spectro-temporal division was 1143 input units (for 88 entropies and AFE features), 40 hidden units, and 88 output units. Since the output probabilities of the weight-generating MLP only indicate guess to the best stream, but not necessarily the best form of the structure of the stream, the weights generated were used to combine features in both the probability and log-probability domains.

4.4 Post-Processing

Processing post-combination is rather straightforward. First, the logarithm of the posterior probabilities is taken (except for MLP weighting on the log-probabilities), and the feature dimensions are reduced from 56 to 32 dimensions by the Karhunen-Loève transform. Finally, the outputs are normalized by utterance and appended to AFE features . These observations serves as the input to the HTK decoder. Figure 2.1 illustrates the processing of the spectro-temporal features.

Chapter 5

Experimental Setup and Results

5.1 Experimental Setup

To test the improvement of recognition performance using spectro-temporal features, I use the Aurora2 data set, a connected digit corpus which contains 8,440 sentences of clean training data and 56,056 sentences of clean and noisy test data. This corpus is the standard for comparing the noise-robustness of algorithms, and hence I am using this data set. The test set comprises 8 different noises (subway, babble, car, exhibition, restaurant, street, airport, train-station) at 7 different noise levels (clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB), totaling 56 different test scenarios, each containing 1,001 sentences. Since we are interested in the performance of the spectro-temporal features in mismatched conditions, all systems were trained only on the clean training set but tested on the entire test set. The baseline system for this paper uses only AFE features.

The parameters for the HTK decoder are the same as that for the standard setup described in [18], with the exception that I parallelized training and decoding in order to speed up experiments. The setup uses whole word HMMs with 16 states with a 3-Gaussian mixture with diagonal covariances per state; skips over states are not permitted in this model. This is the setup used in the ETSI standards competition, from which AFE was created, and is not at all optimized for spectro-temporal features. More details on this setup are available in [18].

5.2 Results

Typical results on the Aurora2 test set using the ETSI setup report accuracies (or mean accuracy) across the 8 noises at 7 noise conditions. We do not report accuracies here for two reasons. The first and rather mundane reason is that reporting hundreds of results will hinder understanding of the overall performance of the system. Second, and perhaps more importantly, I do not think that reporting accuracies in general (even with a reduced table) is properly illustrative of the performance of the system. Consider, for instance, a table consisting of results for two systems in two noise conditions, one clean and one extremely noisy. If the baseline achieved a 98% accuracy rate on the clean test and 3% accuracy on the noisy one, and the proposed system achieved a 99% and 1.9% accuracy on the clean and noisy conditions, respectively, one would clearly choose the latter system as that system reduced over half the errors on the clean test while performing roughly similarly on the noisy one (that is, neither really worked in noise). If we simply look at mean accuracy, however, we see that the baseline actually outperforms the compared system. The reduction in errors corresponds fairly well to the common costs of using a system (for instance, how often a system must retreat to a human operator). For this reason, we report WER results, which for many years have been the standard for most speech recognition tasks.

For this paper, we average WER across noises and report scores for each noisy condition. This type of reporting appears in Appendix B, in which we compare the Gabor systems to the AFE baseline. Unfortunately, since the sheer number of results hinder the overall understanding of the systems, results in this chapter are averaged across clean and 20dB-0dB noise-added test conditions. Tables 5.2 and 5.5 present results of clean test conditions for multi-modulation spectral feature and spectro-temporal MFCC divisions, respectively, while those for noisy test conditions are shown in Tables 5.3 and 5.6. Improvement of the best spectro-temporal systems relative to the AFE baseline are shown in 5.4 and 5.7. The best individual performance (for the multi-modulation spectral and spectro-temporal MFCC divisions) on clean and noisy test conditions are italicized, while systems that performed well in both clean and noisy conditions are highlighted in bold. Finally, results of the best performing system are significant compared to the AFE baseline with a p-value less than 0.01 using the differences of proportions significance test.

System	MFCC	AFE
Clean	1.60%	1.63%
Noisy	17.66%	15.32%

Table 5.1: *Performance of MFCC and AFE baseline systems.*

Output	AM	GM	HM	IE	MLP	Log-MLP
Magnitude	1.13%	1.00%	1.03%	0.97%	0.97%	1.19%
Real	1.06%	1.06%	1.13%	1.05%	1.02%	1.14%
Imaginary	*0.88%*	1.50%	2.83%	1.02%	1.01%	1.02%
Real+Imag EF	0.94%	1.05%	0.99%	1.01%	1.05%	1.09%
Real+Imag+Mag EF	1.03%	1.10%	0.95%	**0.93%**	1.07%	1.17%
Real+Imag LF	1.02%	1.04%	2.94%	1.02%	0.90%	0.94%
Real+Imag+Mag LF	0.97%	0.95%	3.50%	0.96%	0.95%	*0.88%*

Table 5.2: *Performance of ASR systems based on multi-modulation spectral division of spectro-temporal features in clean conditions. AM refers to arithmetic mean combination, GM to geometric mean combination, HM to harmonic mean combination, IE to inverse entropy weighting, MLP to MLP weighting of posteriors, and Log-MLP to MLP-weighting of log posteriors. EF refers to early fusion combination while LF refers to late fusion combination.*

5.3 Discussion

Results indicate that adding spectro-temporal features in almost any framework (whether using multi-modulation spectral or spectro-temporal MFCC division, or using different combination methods) improved Advanced Front End features, and in some cases quite dramatically. The one exception to this case, at least in many instances, is harmonic mean averaging. This is rather unsurprising, however, since harmonic mean averaging followed by the logarithm transformation essentially places more weight on low posterior streams rather than high ones. A priori, this would seem like (and is) a poor choice of weight combination.

Compared to AFE, the best gain for spectro-temporal features occurs in the clean case, in which spectro-temporal features for the best systems using the multi-modulation spectral division reduce 45% of the errors of the system (41% compared to MFCC), while that for the spectro-temporal MFCC division reduce over 53% of the errors (51% compared to MFCC). In noisy cases, the gains are much more modest, but still significant. The best system using multi-modulation spectral division reduced roughly 15% of errors when relative

Output	AM	GM	HM	IE	MLP	Log-MLP
Magnitude	15.04%	14.40%	15.59%	15.12%	14.82%	15.48%
Real	14.58%	14.91%	14.93%	14.87%	15.04%	14.92%
Imaginary	14.82%	21.26%	29.23%	16.77%	14.71%	16.09%
Real+Imag EF	13.85%	13.90%	14.26%	13.76%	14.33%	14.54%
Real+Imag+Mag EF	13.75%	14.65%	14.87%	***13.69%***	13.75%	14.63%
Real+Imag LF	14.36%	16.42%	29.22%	16.46%	14.40%	15.14%
Real+Imag+Mag LF	14.02%	15.38%	32.90%	16.43%	14.09%	14.58%

Table 5.3: *Performance of ASR systems based on multi-modulation spectral division of spectro-temporal features in noisy conditions. AM refers to arithmetic mean combination, GM to geometric mean combination, HM to harmonic mean combination, IE to inverse entropy weighting, MLP to MLP weighting of posteriors, and Log-MLP to MLP-weighting of log posteriors. EF refers to early fusion combination while LF refers to late fusion combination.*

System	Clean WER	Clean Rel. Improvement	Noisy WER	Noisy Rel. Improvement	Noisy Per-Cond. Avg. Imprv.
AFE (baseline)	1.63%	N/A	15.20%	N/A	N/A
Imaginary AM	0.88%	45.94%	14.82%	2.53%	4.67%
R+I+M EF IE	0.93%	43.08%	13.69%	9.95%	15.42%
R+I+M LF Log-MLP	0.88%	45.94%	14.58%	4.10%	8.38%

Table 5.4: *Improvement of ASR systems based on multi-modulation spectral division of spectro-temporal features relative to AFE baseline. Noisy per-condition average improvement averages the improvement of spectro-temporal systems relative to the AFE baseline from 20dB to 0dB SNR cases. EF refers to early fusion combination while LF refers to late fusion combination.*

improvement is averaged over different signal-to-noise ratios (22% compared to MFCC), while the best system using spectro-temporal MFCC division reduced 18% of the errors (24% compared to MFCC). Moreover, as we can see from Figure 5.3, the relative improvement of spectro-temporal systems decays roughly exponentially with lower signal-to-noise ratios. It is interesting to note, however, that the systems improve performance even in the poorest conditions.

In general, spectro-temporal MFCC features outperform their multi-modulation spectral feature counterparts. This is especially true in the clean case, in which a number of spectro-temporal MFCC systems outperforms the best multi-modulation spectral system, and the

Output	AM	GM	HM	IE	MLP	Log-MLP
Magnitude	0.81%	0.82%	1.07%	0.84%	0.80%	0.93%
Real	0.85%	0.97%	1.00%	0.90%	0.85%	0.83%
Imaginary	0.84%	0.85%	1.03%	**0.79%**	**0.80%**	0.92%
Real+Imag EF	0.87%	1.23%	1.07%	1.57%	0.95%	1.61%
Real+Imag+Mag EF	0.92%	1.08%	1.60%	0.93%	0.89%	1.13%
Real+Imag LF	0.87%	1.01%	1.04%	0.90%	**0.78%**	0.91%
Real+Imag+Mag LF	*0.77%*	0.90%	0.97%	**0.79%**	**0.78%**	0.88%

Table 5.5: *Performance of ASR systems based on spectro-temporal MFCCs division of features in clean conditions. AM refers to arithmetic mean combination, GM to geometric mean combination, HM to harmonic mean combination, IE to inverse entropy weighting, MLP to MLP weighting of posteriors, and Log-MLP to MLP-weighting of log posteriors. EF refers to early fusion combination while LF refers to late fusion combination.*

best spectro-temporal MFCC systems exhibits a roughly 10-12% improvement relative to the best multi-modulation spectral system. In the noisy case, spectro-temporal features generally outperform multi-modulation spectral features, but the relative improvement is quite a bit more modest, with only a 2% relative improvement. One thing to note, however, is that for multi-modulation spectral features, the systems that performed well under clean conditions were not the same as those that performed well in noise-added conditions. By contrast, a number of spectro-temporal MFCC systems perform well in both clean and noise-added conditions. Whether or not this is due to the non-standard structure of MLPs for multi-modulation spectral featuress versus the standard structure for spectro-temporal MFCCs, or something else, remains an open question.

As can be seen in the tables above, imaginary outputs alone performed better than its real and magnitude counterparts. Before making biological claims on why this may be true, one must realize that for real and magnitude outputs, a feature that has $+a$ Hz temporal modulation and a $-a$ Hz temporal modulation are indeed the same feature, while for imaginary outputs these two calculations generate different features. So imaginary outputs may perform better because more features are added. Combining different types of outputs leads to somewhat uneven conclusions. While one may expect that including all types of features should lead to better performance, this is only true in noise-added condition for multi-modulation spectral divisions and clean cases for spectro-temporal ones. While

Output	AM	GM	HM	IE	MLP	Log-MLP
Magnitude	15.12%	15.19%	15.49%	15.06%	14.82%	15.10%
Real	14.08%	13.90%	14.33%	14.08%	14.02%	13.85%
Imaginary	14.04%	14.23%	14.42%	**13.64%**	***13.37%***	13.63%
Real+Imag EF	14.46%	14.72%	16.05%	14.16%	13.77%	16.17%
Real+Imag+Mag EF	14.99%	14.29%	15.24%	14.57%	13.99%	14.37%
Real+Imag LF	14.55%	14.05%	14.25%	13.63%	**13.70%**	13.39%
Real+Imag+Mag LF	14.37%	13.55%	14.65%	**13.66%**	**13.76%**	14.10%

Table 5.6: *Performance of ASR systems based on spectro-temporal MFCCs division of features in noisy conditions. AM refers to arithmetic mean combination, GM to geometric mean combination, HM to harmonic mean combination, IE to inverse entropy weighting, MLP to MLP weighting of posteriors, and Log-MLP to MLP-weighting of log posteriors. EF refers to early fusion combination while LF refers to late fusion combination.*

the performance of systems that include real, imaginary, and magnitude outputs is not significantly worse than those with only imaginary outputs (which were the best performing systems), that the performance is worse suggests a suboptimality in the posterior combination method. Finally, combining multiple outputs via early fusion on average outperformed late fusion strategies for multi-modulation spectral divisions, but this conclusion is reversed for spectro-temporal ones.

Turning to combination methods, one notices that combination by weight-generating MLP generally outperforms all other combination methods in both the clean and noise-added conditions. In a few cases, such as the noise-added one for multi-modulation spectral division, simple inverse entropy combination outperformed MLP weight combination, but even in these cases, the systems combined with MLP weight generation are not significantly poorer. Moreover, dynamic combination methods outperformed static combination methods, and of the static combination methods, arithmetic outperformed geometric and harmonic mean combination.

5.3.1 Robustness

A major problem with the above discussion is that one cannot be sure if the results represent what would happen in general, or is a result of random chance, parameter tuning, or the recognition system. To test the robustness of the methods studied, I tested a subset

System	Clean WER	Clean Rel. Improvement	Noisy WER	Noisy Rel. Improvement	Noisy Per-Cond. Avg. Imprv.
AFE (baseline)	1.63%	N/A	15.20%	N/A	N/A
Imaginary IE	0.79%	51.38%	13.64%	10.26%	12.72%
Imaginary MLP	0.80%	50.77%	13.37%	12.08%	17.92%
R+I LF MLP	0.78%	52.00%	13.70%	9.87%	14.88%
R+I+M LF AM	0.77%	52.62%	14.37%	5.50%	8.49%
R+I+M LF IE	0.79%	51.38%	13.66%	10.14%	15.75%
R+I+M LF MLP	0.78%	52.00%	13.76%	9.47%	13.34%

Table 5.7: *Improvement of ASR systems based on spectro-temporal MFCC division of features relative to AFE baseline. Noisy per-condition average improvement averages the improvement of spectro-temporal systems relative to the AFE baseline from 20dB to 0dB SNR cases. EF refers to early fusion combination while LF refers to late fusion combination.*

of the techniques on the Number95 test set. The setup is the same as the exploratory work in Chapter 2.[1] In particular, there are two questions to be answered: are imaginary outputs always better than real or magnitude ones, and do spectro-temporal MFCC methods outperform multi-modulation spectral feature-based ones?

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	2.12%	2.23%	2.29%	2.31%
Real	2.82%	2.77%	2.88%	2.73%
Imaginary	2.61%	2.54%	2.44%	2.59%
Real+Imag LF	2.99%	2.84%	2.61%	2.84%

Table 5.8: *Performance of ASR systems based on multi-modulation spectral division of spectro-temporal features in clean conditions on the Numbers95 corpus. AM refers to arithmetic mean combination, GM to geometric mean combination, IE to inverse entropy weighting, and MLP to MLP weighting of posteriors. LF refers to late fusion combination.*

These results helps disabuse us from some conclusions presented had we looked at the Aurora2 test set alone. The first is that imaginary outputs are not necessarily the best output of the Gabor filter. For clean test conditions under multi-modulation spectral division, the magnitude outputs outperform all other outputs. The caveat here, however, is that the filters

[1] As with the Aurora2 test set, the per-comdition results are again relegated to the Appendices (Appendix C).

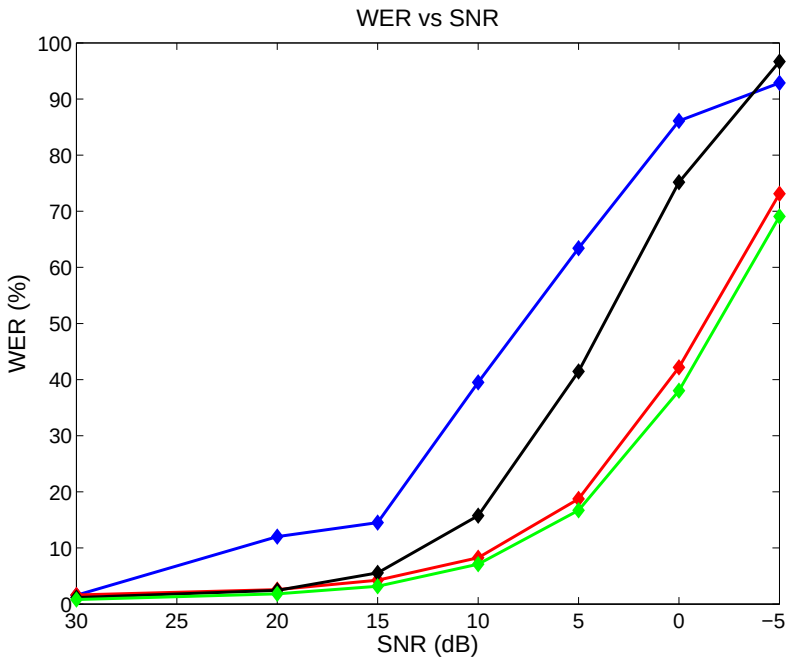

Figure 5.1: *WER vs. SNR for different ASR systems on Aurora2 corpus. The blue curve represents MFCC features; the black MFCC plus equally-weighted spectro-temporal features under multi-modulation spectral division; the red AFE features; and the green AFE+ imaginary spectro-temporal MFCCs with MLP weighting.*

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	27.89%	24.15%	22.11%	21.43%
Real	4.08%	5.78%	2.04%	7.14%
Imaginary	11.22%	13.61%	17.01%	11.90%
Real+Imag LF	-1.70%	3.40%	11.22%	3.40%

Table 5.9: *Improvement of ASR systems based on multi-modulation spectral division of spectro-temporal features relative to MFCC baseline in clean conditions on Numbers95 corpus. LF refers to late fusion combination.*

with magnitude output were tuned in [34] to perform well on this particular test set. For the noisy test case, real outputs perform all other types of outputs. For spectro-temporal MFCC features, real and imaginary parts outperform magnitude outputs. For noise-added test conditions, real outputs perform the best. This suggests that imaginary outputs are not the necessarily the best type of output for Gabor features. Moreover, combining real

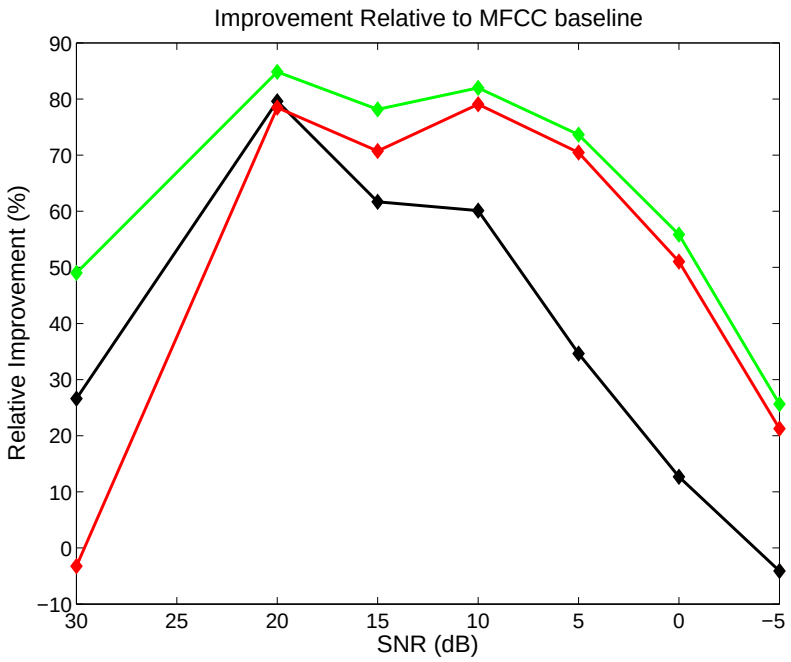

Figure 5.2: *Improvement relative to MFCC baseline vs. SNR for different ASR systems on Aurora2 corpus. The black MFCC plus equally-weighted spectro-temporal features under multi-modulation spectral division; the red AFE features; and the green AFE plus imaginary spectro-temporal MFCCs with MLP weighting.*

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	16.58%	15.72%	16.30%	16.65%
Real	15.44%	15.62%	15.97%	15.19%
Imaginary	16.05%	16.05%	16.65%	16.15%
Real+Imag LF	16.53%	16.20%	16.50%	16.25%

Table 5.10: *Performance of ASR systems based on multi-modulation spectral division of spectro-temporal features in noisy conditions on the Numbers95 corpus. AM refers to arithmetic mean combination, GM to geometric mean combination, IE to inverse entropy weighting, and MLP to MLP weighting of posteriors. LF refers to late fusion combination.*

and imaginary components under a late fusion scheme does not yield better results than using real or imaginary parts alone, suggesting that a better combination method needs to be researched.

As for the second question, the results again imply that the conclusions posited for only

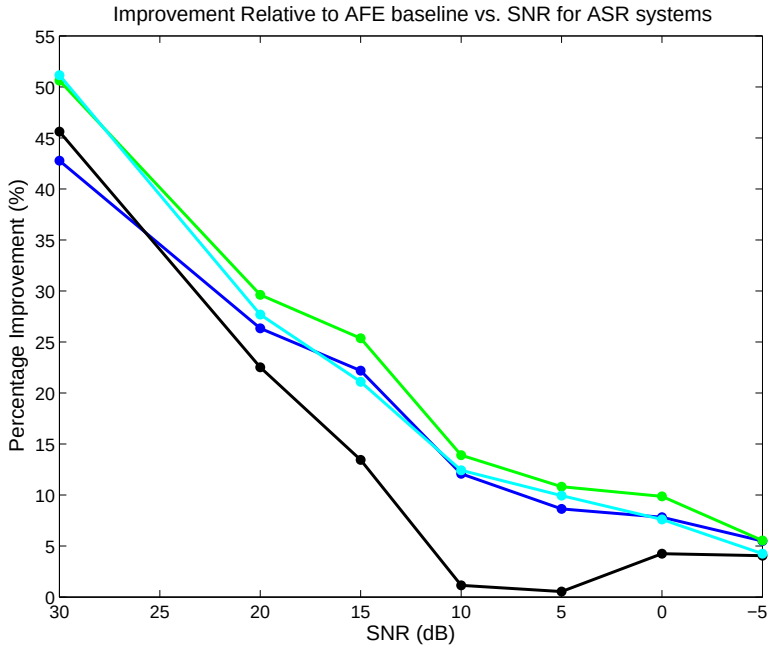

Figure 5.3: *Improvement relative to AFE baseline vs. SNR for different ASR systems on Aurora2 corpus. The green curve represents imaginary spectro-temporal MFCCs with MLP weighting; the cyan late fusion real, imaginary, and magnitude spectro-temporal MFCCs with inverse entropy weighting; the blue imaginary multi-modulation spectral features with inverse entropy weighting; and the black late fusion real, imaginary, and magnitude multi-modulation spectral features with MLP weighting on log-probabilities.*

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	6.12%	10.98%	7.69%	5.70%
Real	12.54%	11.55%	9.55%	13.97%
Imaginary	9.12%	9.12%	5.70%	8.55%
Real+Imag LF	6.41%	8.26%	6.55%	7.98%

Table 5.11: *Improvement of ASR systems based on multi-modulation spectral division of spectro-temporal features relative to MFCC baseline in noisy conditions on Numbers95 corpus. LF refers to late fusion combination.*

the Aurora2 corpus do not necessarily generalize. For the clean case, multi-modulation spectral division clearly outperforms spectro-temporal MFCC division. The same caveat that the multi-modulation spectral features are tuned to this test set also applies here. In

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	2.84%	3.01%	3.07%	2.90%
Real	2.86%	2.86%	2.82%	2.65%
Imaginary	2.65%	2.94%	2.56%	2.56%
Real+Imag LF	2.65%	2.86%	2.88%	2.92%

Table 5.12: *Performance of ASR systems based on spectro-temporal MFCC division of spectro-temporal features in clean conditions on the Numbers95 corpus. AM refers to arithmetic mean combination, GM to geometric mean combination, IE to inverse entropy weighting, and MLP to MLP weighting of posteriors. LF refers to late fusion combination.*

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	3.47%	-2.25%	-4.39%	1.33%
Real	2.76%	2.76%	4.19%	9.91%
Imaginary	9.91%	-0.10%	12.77%	12.77%
Real+Imag LF	9.91%	2.76%	2.04%	0.61%

Table 5.13: *Improvement of ASR systems based on spectro-temporal MFCC division of spectro-temporal features relative to MFCC baseline in clean conditions on Numbers95 corpus. LF refers to late fusion combination.*

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	15.80%	15.57%	15.37%	15.70%
Real	14.61%	15.24%	14.76%	15.14%
Imaginary	15.12%	15.27%	15.67%	15.95%
Real+Imag LF	15.14%	14.87%	15.12%	15.09%

Table 5.14: *Performance of ASR systems based on spectro-temporal MFCC division of spectro-temporal features in noisy conditions on the Numbers95 corpus. AM refers to arithmetic mean combination, GM to geometric mean combination, IE to inverse entropy weighting, and MLP to MLP weighting of posteriors. LF refers to late fusion combination.*

the noisy case, the spectro-temporal MFCC features outperform multi-modulation spectral features. Given that spectro-temporal MFCCs are not at all tuned to this test set implies that there exists some validity to the spectro-temporal MFCC approach.

Output	+ST AM	+ST GM	+ST IE	+ST MLP
Magnitude	10.53%	11.83%	12.97%	11.10%
Real	17.27%	13.70%	16.42%	14.27%
Imaginary	14.38%	13.53%	11.27%	9.68%
Real+Imag LF	14.27%	15.80%	14.38%	14.55%

Table 5.15: *Improvement of ASR systems based on spectro-temporal MFCC division of spectro-temporal features relative to MFCC baseline in noisy conditions on Numbers95 corpus. LF refers to late fusion combination.*

Chapter 6

Conclusion

Spectro-temporal features can be shown to improve performance in noise-added conditions, even with such a difficult baseline such as Advanced Front End features. In clean conditions, spectro-temporal features reduce roughly half of the errors, while in noise-added conditions, spectro-temporal features reduce roughly 18% of the errors. Compared to the MFCC baseline, spectro-temporal features reduce over 50% and 68% of the errors in clean and noisy cases, respectively. It is my view that specto-temporal processing is quite simply a novel filtering with a biological basis, but this filtering (which can be done at many different spectral and temporal modulations) can lead to a robust feature set for ASR.

Apart from the standard questions (such as, will these features work for conversational speech, etc.) there exist, however, a number of interesting areas on calculating spectro-temporal features for robust ASR. One is what spectro-temporal modulations are important for good speech recognition performance. Are the spectro-temporal modulations that humans respond to the same as what will give us best ASR performance?

The second question has more than just an academic importance. [34] picked spectral and temporal modulations to which ferrets (and therefore humans) respond in biological studies. In the implementation of such features, the temporal modulations above 0Hz, which should have ranged from ±2Hz to ±16Hz to be consistent with biological findings, instead captured modulations from ±6Hz to ±50Hz. As shown in Table 6.1, however, correcting the features with ones consistent with biological findings (i.e. including only temporal modulations from ±2Hz to ±16Hz yielded poorer performance. Even combining corrected and uncorrected

spectro-temporal features, so that the temporal modulations range from ±2Hz to ±50Hz, results in poorer performance than just using features with temporal modulations from ±6Hz to ±50Hz alone. This suggests that a more data-driven or careful approach to inclusion of certain spectro-temporal features is needed.

SNR	MFCC (baseline)	Uncorrected Spectro-Temporal	Corrected Spectro-Temporal	Corrected+Uncorrected Spectro-Temporal
clean	2.94%	2.23%	2.59%	2.33%
20dB	4.72%	4.59%	5.19%	5.36%
15dB	9.22%	7.72%	8.34%	5.60%
10dB	9.95%	8.31%	9.57%	9.57%
5dB	24.74%	22.07%	21.30%	21.81%
0dB	39.43%	35.70%	37.19%	37.93%
20-0dB Avg.	17.66%	15.72%	16.35%	16.10%

Table 6.1: *Comparison of recognition results of MFCC baseline to uncorrected and corrected spectro-temporal multi-modulation spectral features with magnitude outputs appended to MFCCs on the Numbers95 dataset. The spectro-temporal features in this experiment are geometrically weighted per frame.*

SNR	AFE (baseline)	Uncorrected Spectro-Temporal	Corrected Spectro-Temporal	Corrected+Uncorrected Spectro-Temporal
clean	1.63%	0.80%	0.89%	0.80%
20dB	2.59%	1.82%	1.90%	1.98%
15dB	4.25%	3.17%	3.38%	3.40%
10dB	8.27%	7.12%	6.94%	6.89%
5dB	18.74%	16.71%	15.47%	15.25%
0dB	42.18%	38.02%	35.60%	35.02%
-5dB	73.12%	68.07%	66.15%	66.09%
20-0dB Avg.	15.21%	13.37%	12.66%	12.51%

Table 6.2: *Comparison of recognition results of AFE baseline to uncorrected and corrected spectro-temporal MFCC features with imaginary outputs appended to AFE features on the Aurora2 dataset. The spectro-temporal features in this experiment are combined via MLP weight-generating network.*

Another interesting question is if spectro-temporal processing in the log critical band domain is the optimal one for spectro-temporal processing. It could be that the auditory spectrograms generated by PLP features, the spectro-temporal processing requires different

spacing of mel-channels, or spectrums optimized for ASR, as is used in [4], result in better performance. It could be a combination of the above features leads to better performance, and the avenues for tweaking such a system seems endless.

Perhaps a more important question than the above is how to combine this set of features for robust ASR. One possible line of attack is what is done in this thesis, discriminatively train via MLPs and combine different outputs to achieve low phone classification error. If one were to obtain perfect phone classification, [11] shows that one can obtain dramatic reductions in ASR errors. Unfortunately, it seems that these spectro-temporal features with the above discriminative training methods do not yet achieve perfect phone classification.

There exist many problems with the current method. For instance, in the spectro-temporal MFCC framework, some of the systems required combining over 250 posteriors. Even with such a small dataset (6 hours of training data and 40 hours of test), this requires 1.5TB of disk space to store all the features. While this amount of storage is feasible for a small amount of data, those requirements would become too cumbersome for most systems even with a few hundred hours of data. The static and inverse entropy combinations do not require all the intermediate probably estimates to be stored, but MLP weight-generation requires either all the intermediate data to be stored, or multiple calculations of the posteriors. With the success of this combination strategy, however, approximate methods should be explored.

There also exist more theoretical concerns. Consider, for example, three set of features, A,B, and C, such that feature sets A and B are equal and C is different. In all the combination methods described, the combined posterior may be significantly different than if only A and C were used, even though there exist the same information. An optimal combination strategy should be able to handle this problem, especially with a redundant feature set such as spectro-temporal features.

Moreover, the best combination method, the weight-generating MLP, does not necessarily have an easy interpretation. If we consider a simplified case in which are the inputs are only entropies of the different streams with one frame context, and "perfect" labels, we are trying to predict the best-performing stream from entropy information. There exist, however, two problems with this approach. First is that multi-layer perceptrons are typically confident, even when they are wrong, suggesting that measures better than entropies are needed. The

second problem, though, is more subtle and perhaps more insidious. One labels the "best stream" as the one which minimizes the cross-entropy between the label and the posterior of a given stream. This, however, does not imply that the posterior "best stream" is indeed the one for which WER is the lowest. Consider for instance, two streams. One stream has calculates the probability of the correct phone at 0.5 and another incorrect phone at 0.5. The other calculates the probability of a correct phone at 0.49, but calculates the probability of the 51 other phones at 0.01. The second stream seems like a better candidate for the "best stream", but it is the first stream that is labeled as correct. This is not to suggest that MLP weight generation is a poor strategy, as it outperformed other combination methods in this work, but it is difficult to determine what exactly is going on.

The problem with spectro-temporal combination stems from the fact that the current paradigm of speech recognition, which uses an HMM decoder, is typically tuned to a low-dimensional and orthogonal set of features. Spectro-temporal features do not exhibit these two characteristics. Essentially, the Tandem methods can be considered a way to transform features that do not fit the assumptions of the standard decoder into a form that can be used by the HMM. Spectro-temporal features, however, even test the limits of the normally robust Tandem method. It seems that if spectro-temporal features work for dramatic reductions in ASR errors in noise, it may need to be coupled with a different type of system than what is used in the current paradigm.

Appendix A

Spectro-MFCC Performance in Different Noises

The appendix represents the performance of spectro-temporal MFCCs at different spectral and temporal modulations. Blue indicates low WER while red indicates high WER.

A.1 Real

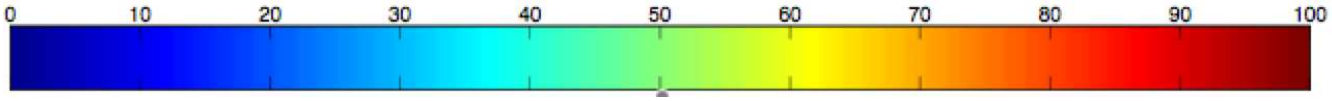

Figure A.1: *Colorbar for WER for different spectro-temporal MFCCs.*

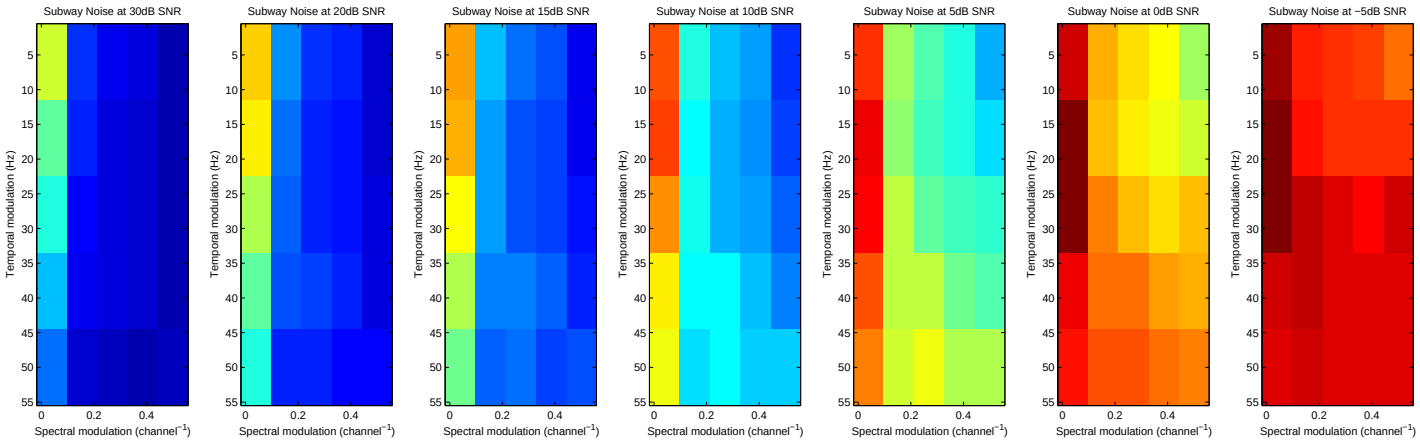

Figure A.2: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Subway noise.*

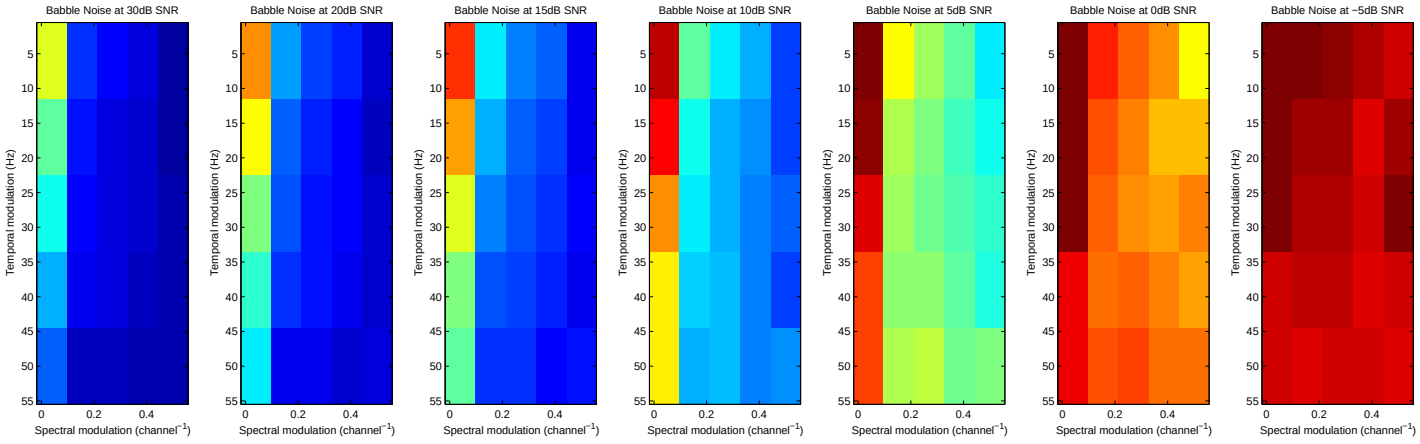

Figure A.3: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Babble noise.*

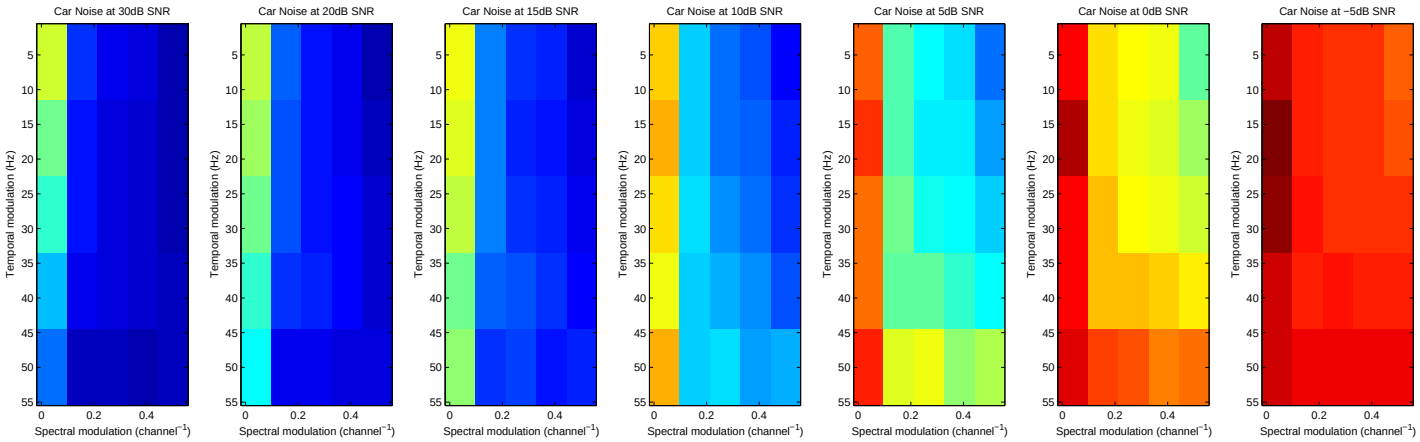

Figure A.4: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Car noise.*

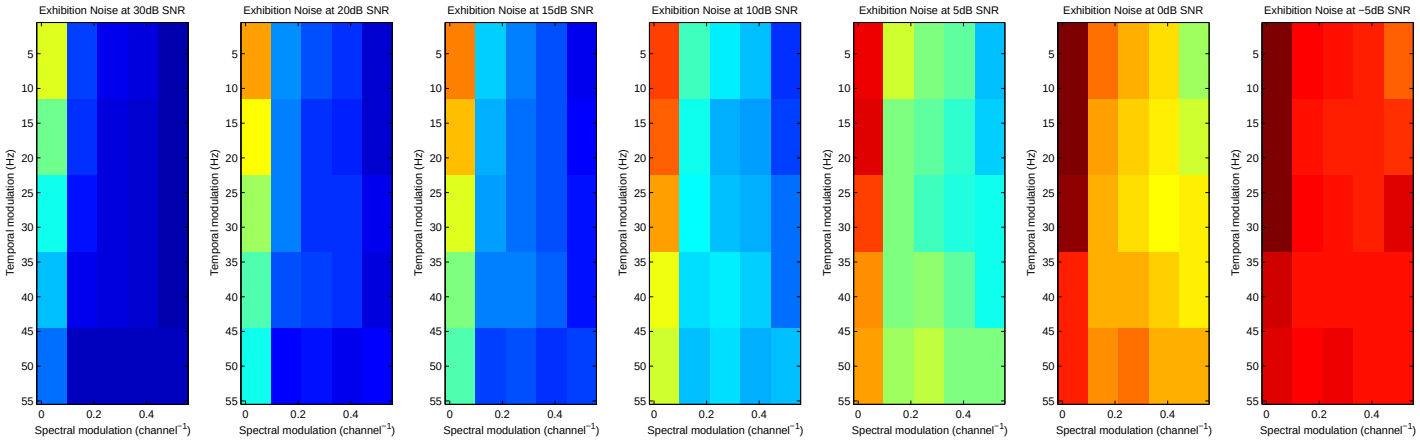

Figure A.5: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Exhibition noise.*

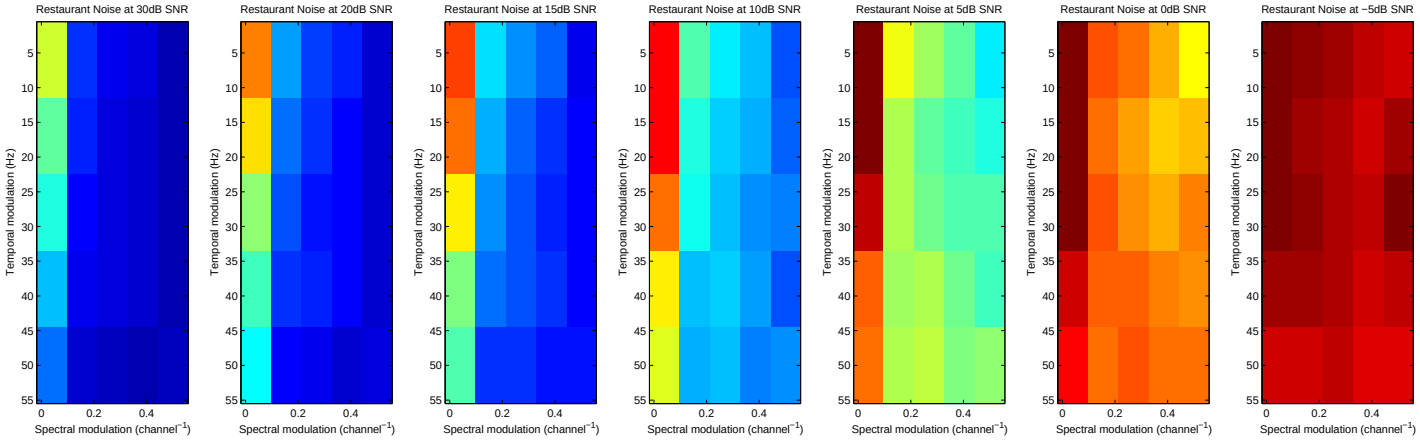

Figure A.6: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Restaurant noise.*

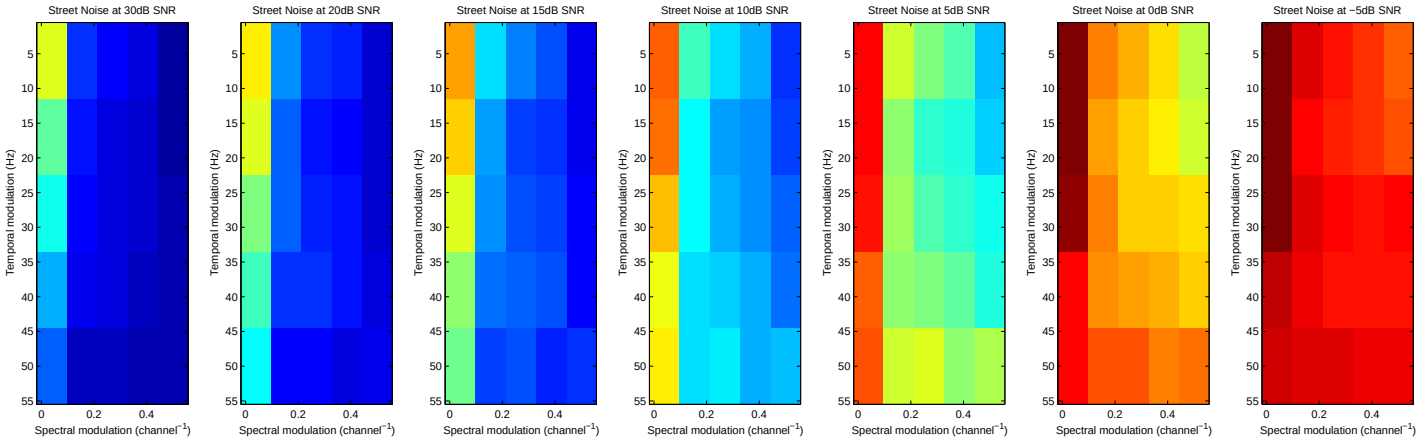

Figure A.7: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Street noise.*

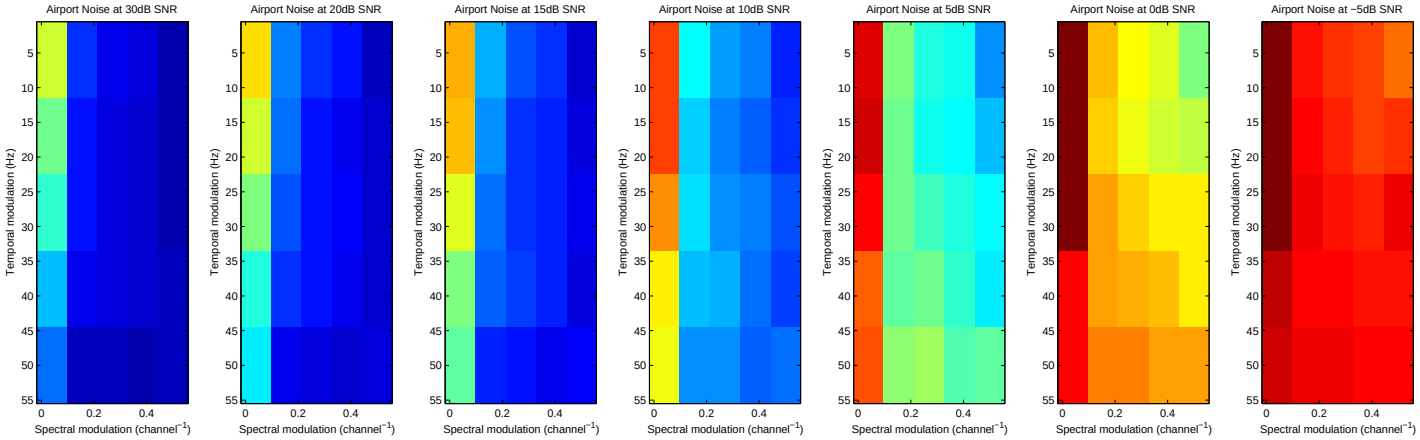

Figure A.8: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Airport noise.*

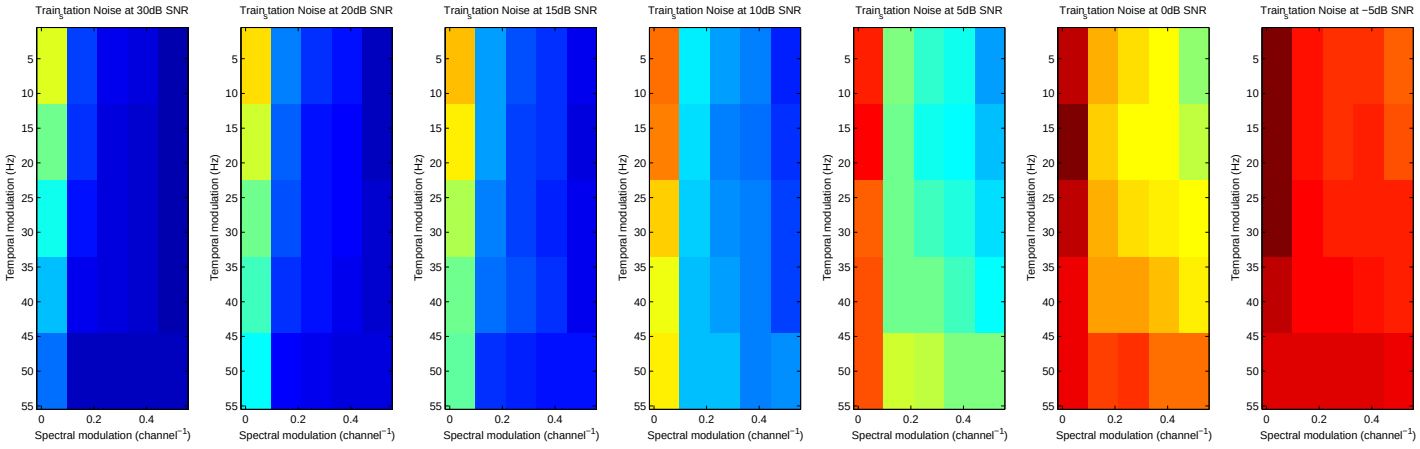

Figure A.9: *WER of real spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Train Station noise.*

A.2 Imaginary

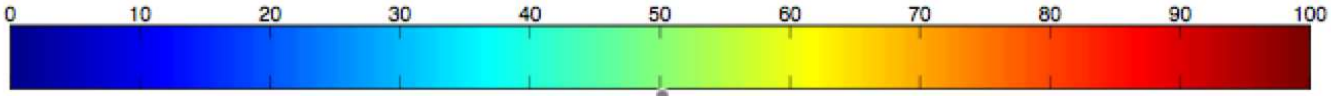

Figure A.10: *Colorbar for WER for different spectro-temporal MFCCs.*

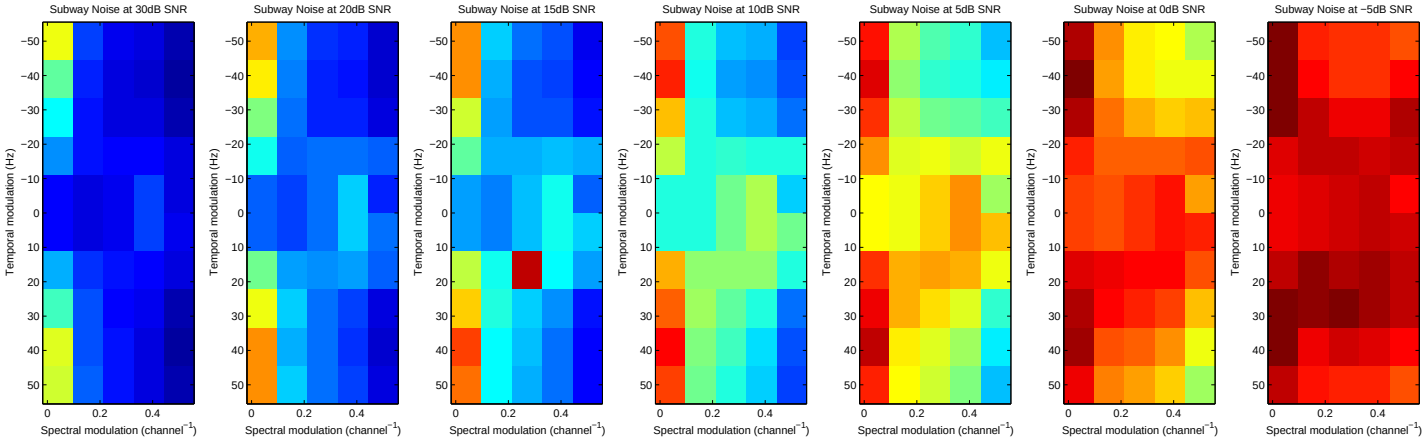

Figure A.11: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Subway noise.*

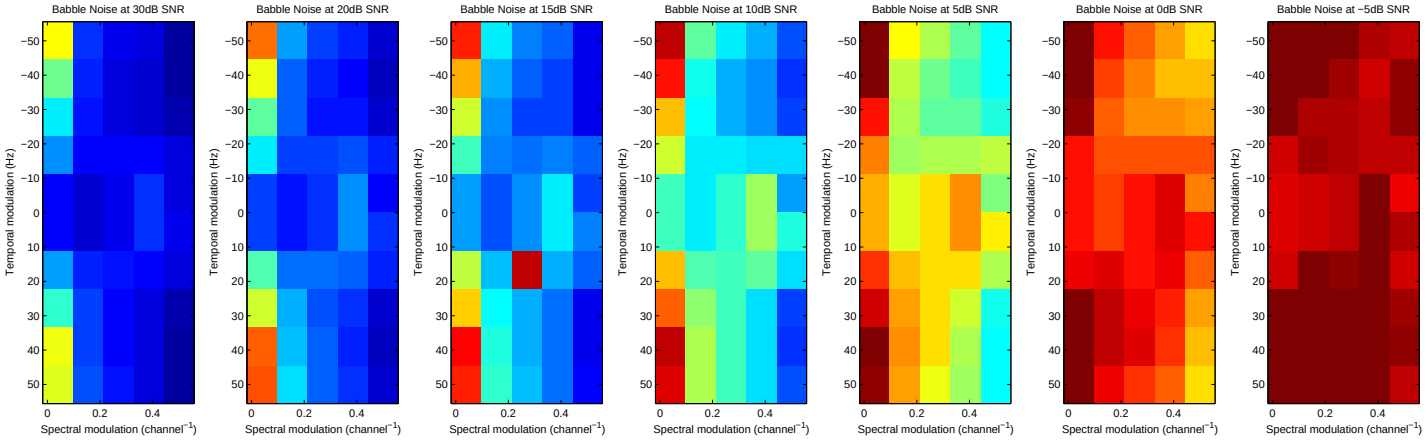

Figure A.12: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Babble noise.*

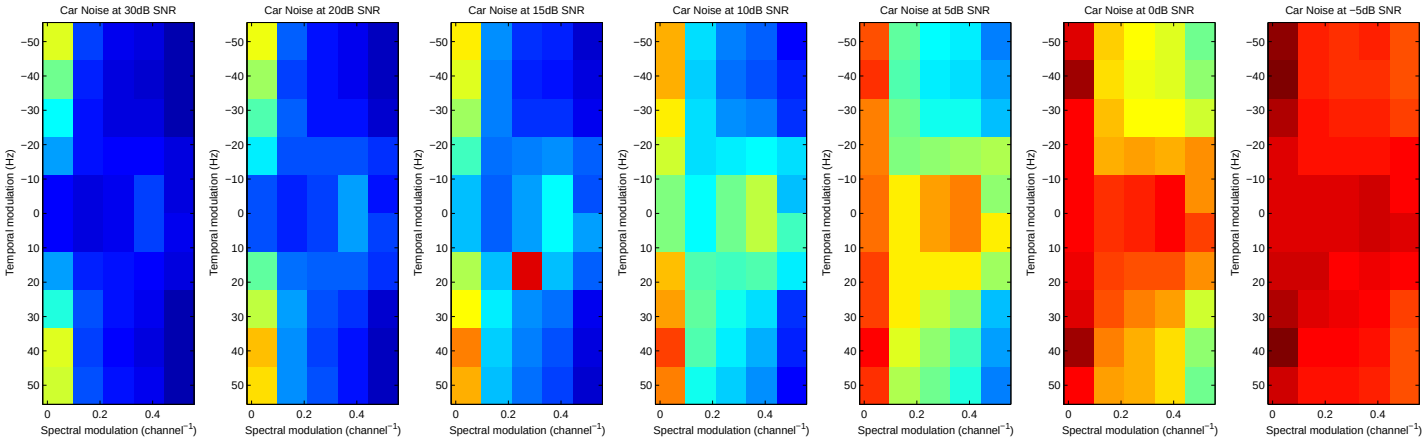

Figure A.13: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Car noise.*

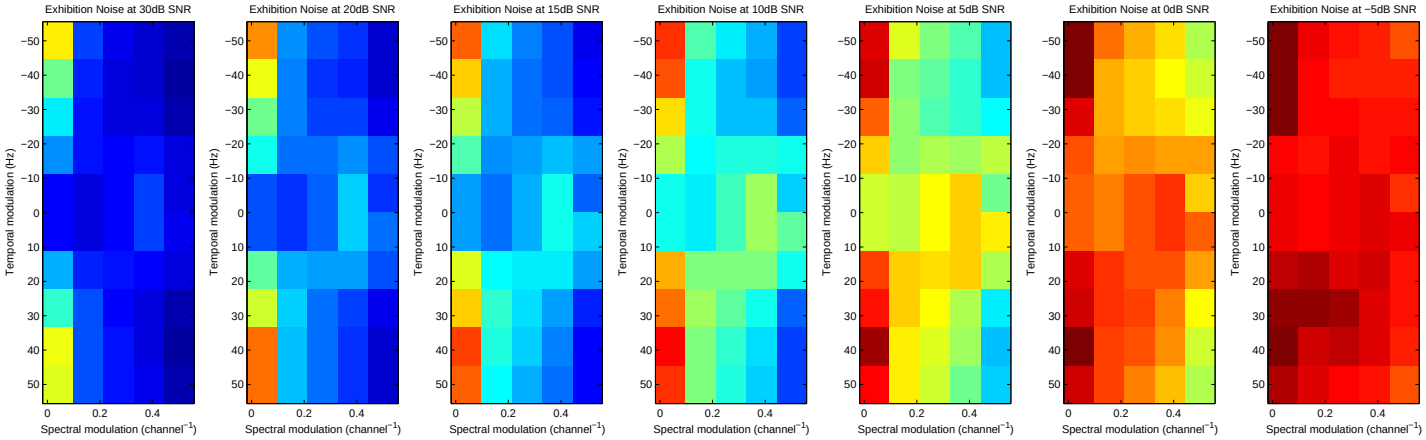

Figure A.14: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Exhibition noise.*

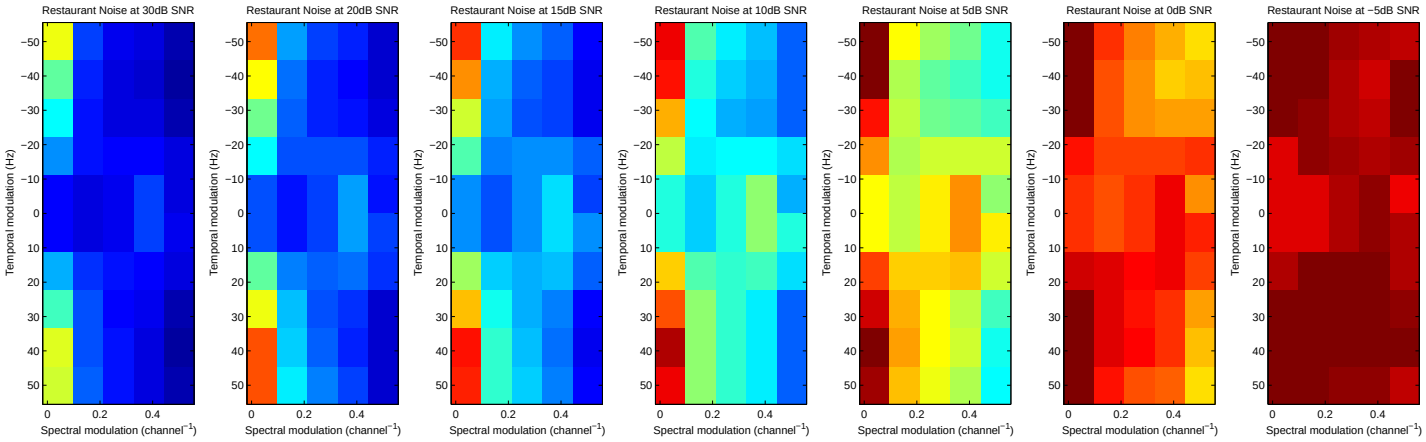

Figure A.15: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Restaurant noise.*

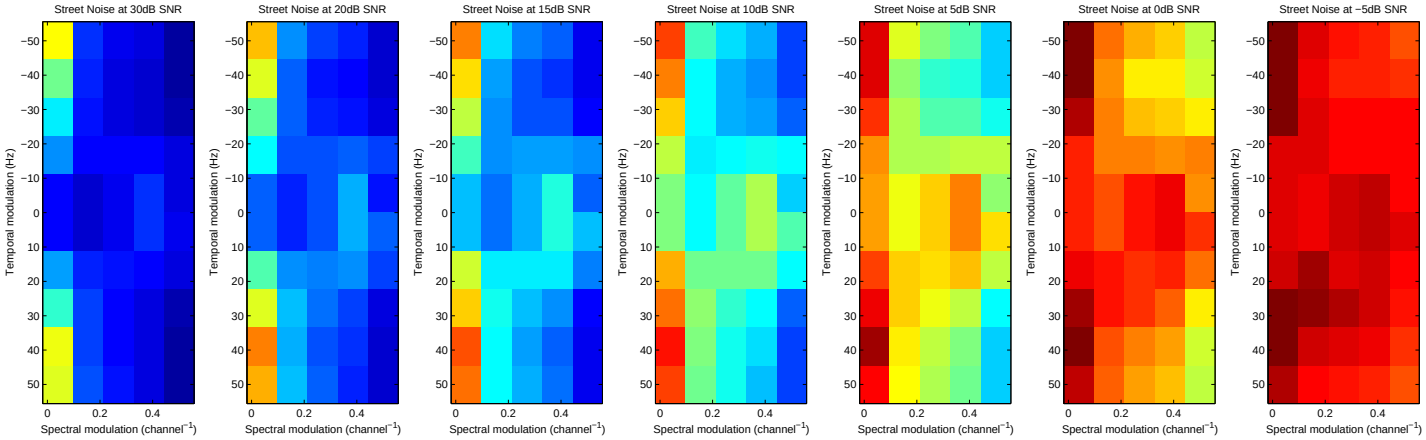

Figure A.16: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Street noise.*

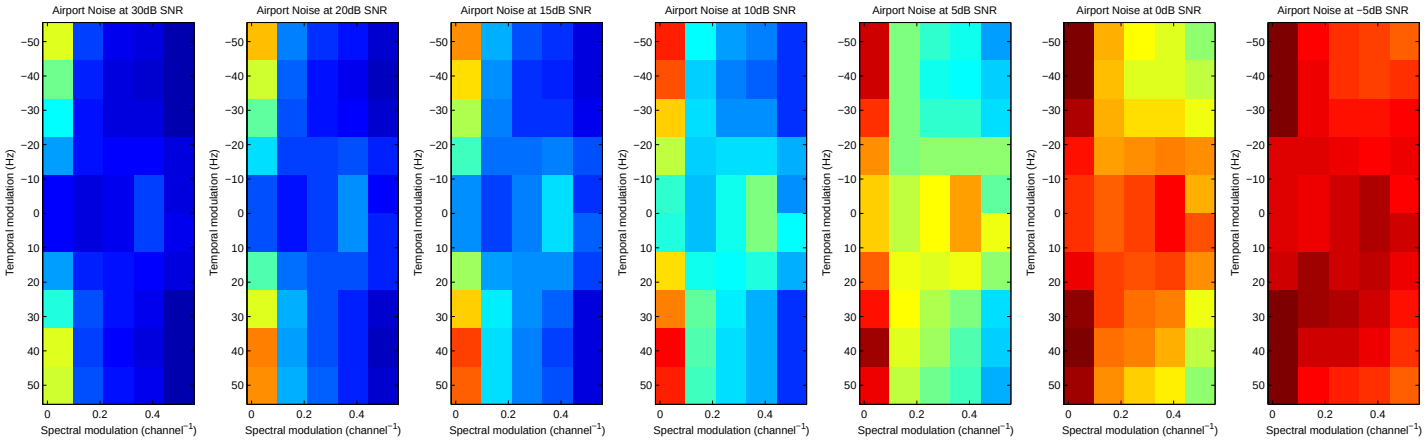

Figure A.17: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Airport noise.*

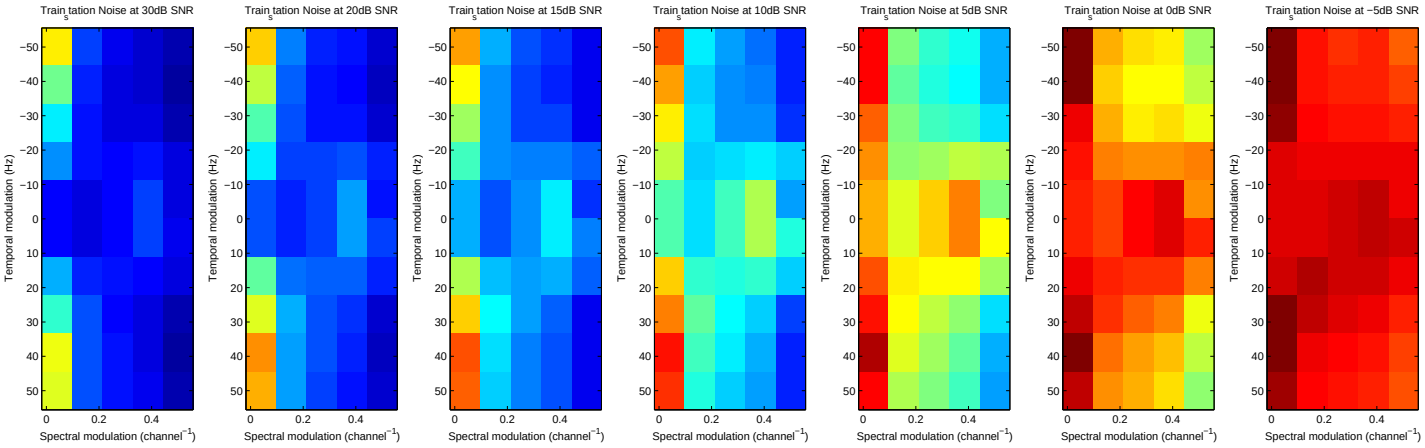

Figure A.18: *WER of imaginary spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Train Station noise.*

A.3 Magnitude

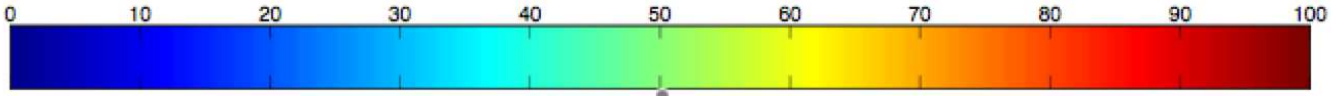

Figure A.19: *Colorbar for WER for different spectro-temporal MFCCs.*

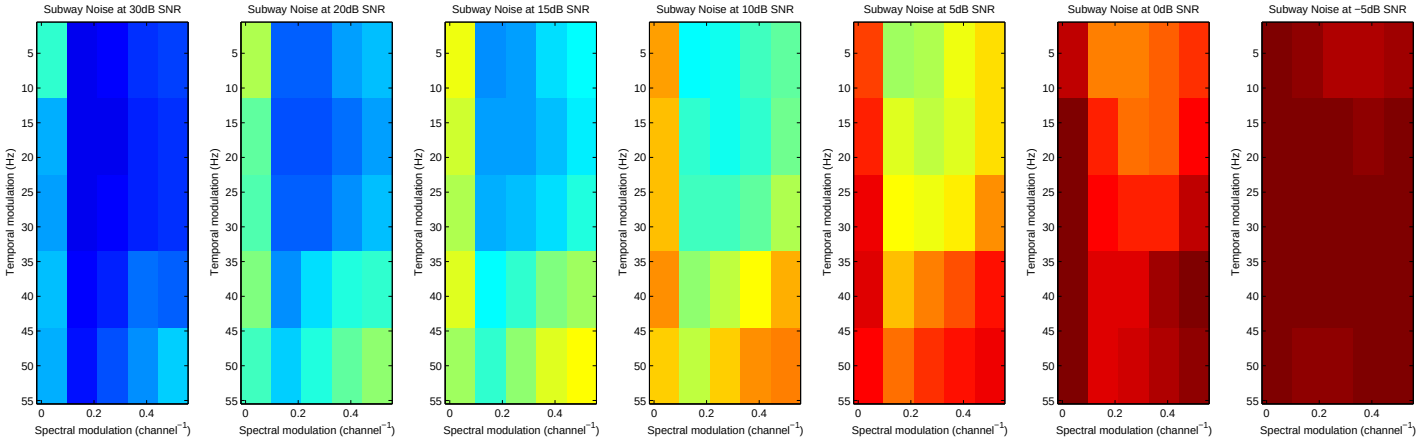

Figure A.20: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Subway noise.*

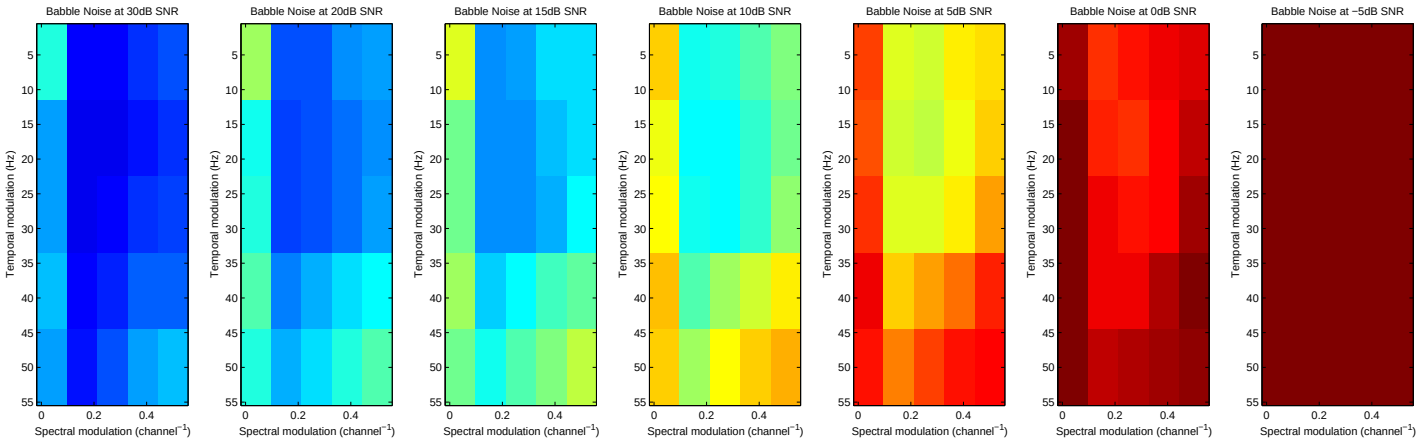

Figure A.21: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Babble noise.*

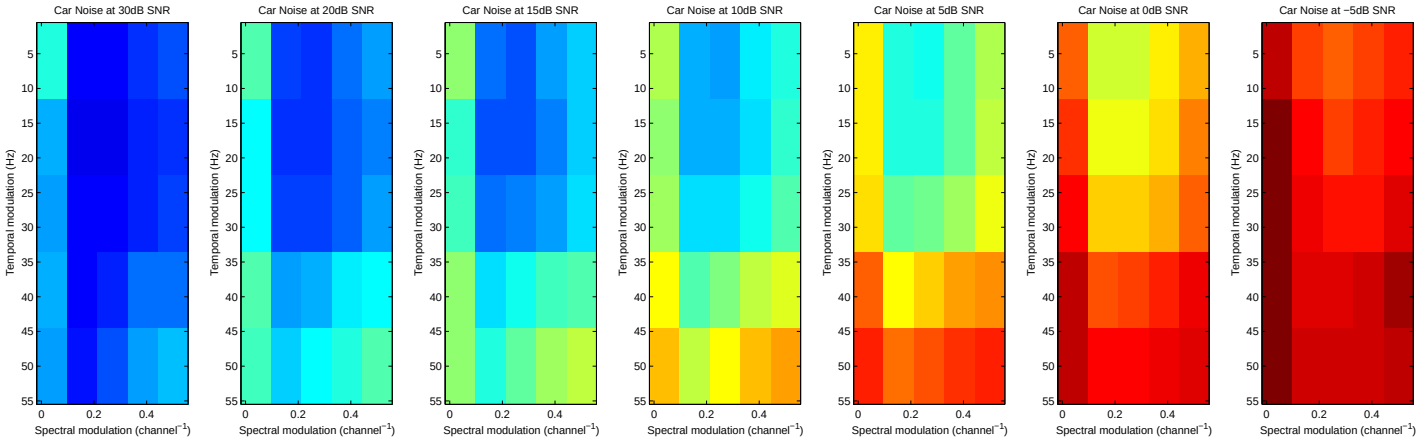

Figure A.22: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Car noise.*

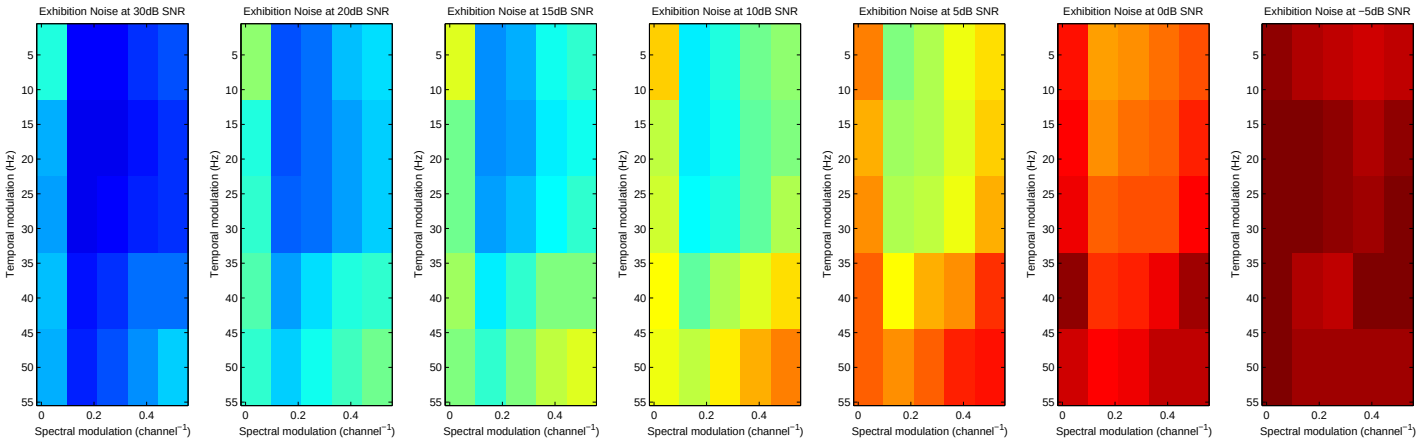

Figure A.23: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Exhibition noise.*

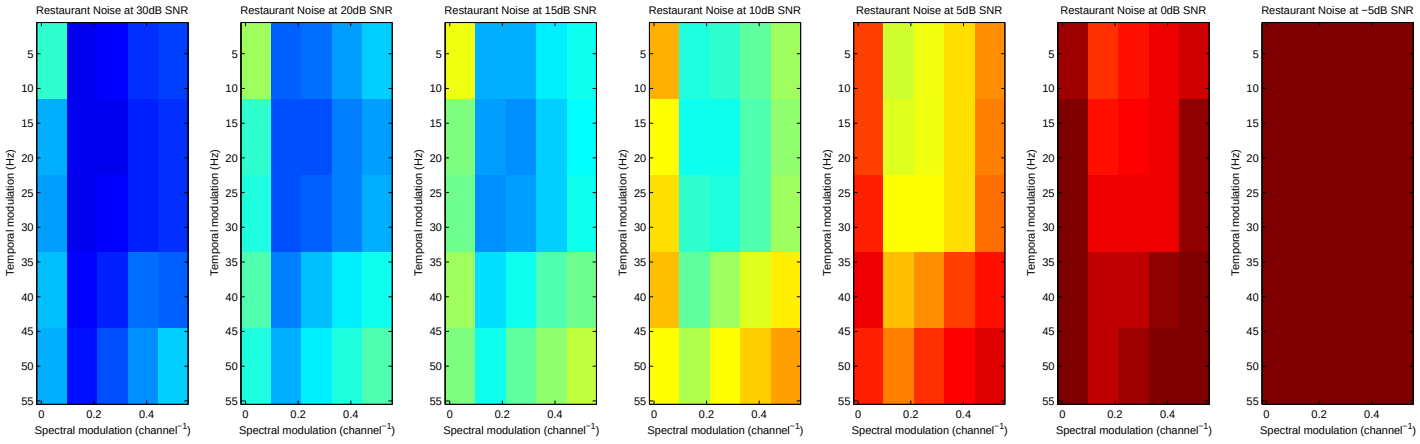

Figure A.24: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Restaurant noise.*

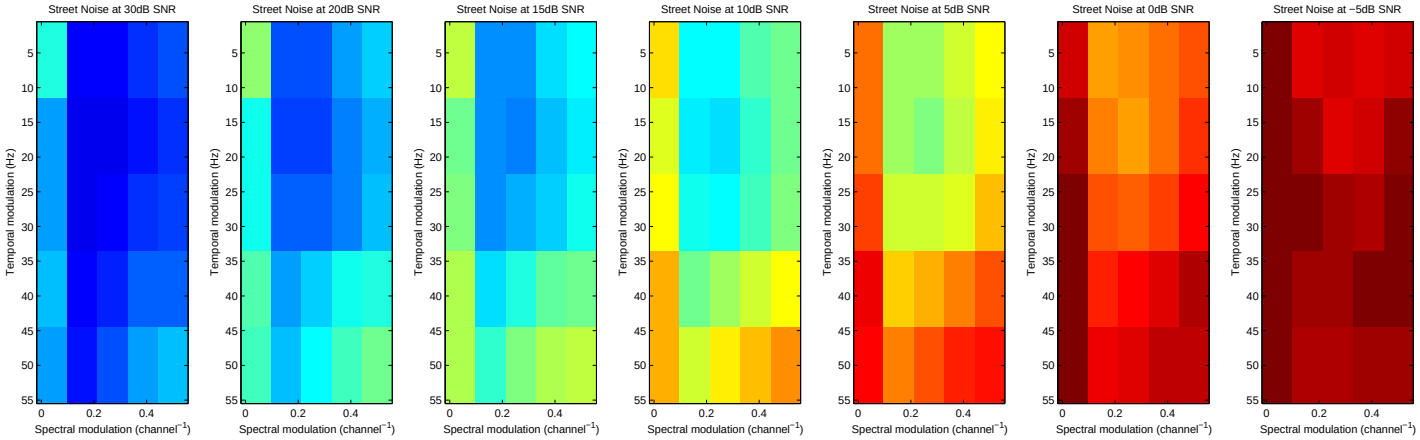

Figure A.25: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Street noise.*

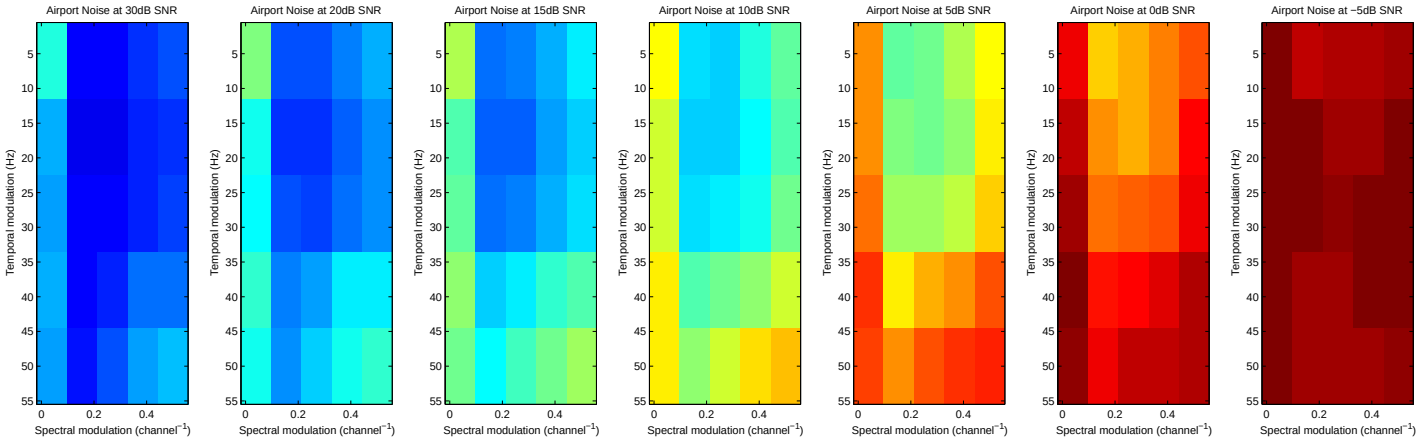

Figure A.26: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Airport noise.*

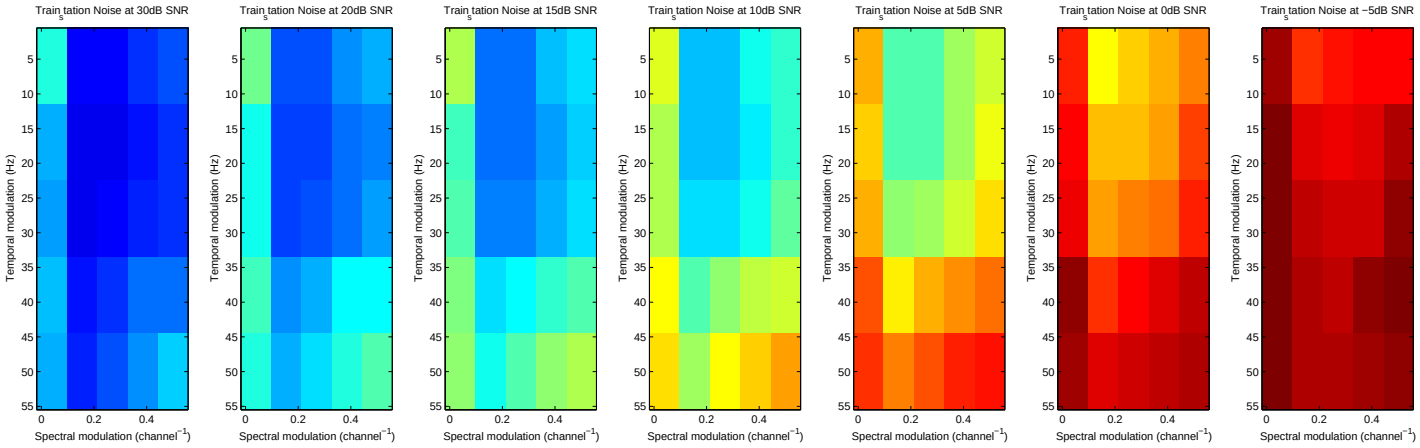

Figure A.27: *WER of magnitude spectro-temporal features for a given temporal and spectral modulation under under different signal to noise ratios for Train Station noise.*

Appendix B

Results per SNR for Aurora2 corpus

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	1.06%	1.06%	1.13%	1.05%	1.02%	1.14%
20dB	2.20%	2.10%	2.28%	2.19%	2.28%	2.33%
15dB	3.88%	4.14%	3.99%	4.05%	4.09%	4.14%
10dB	8.31%	8.85%	8.90%	8.74%	8.88%	8.75%
5dB	18.81%	19.18%	19.02%	19.04%	19.26%	19.05%
0dB	39.68%	40.31%	40.47%	40.32%	40.71%	40.30%
-5dB	69.93%	70.27%	70.40%	70.45%	69.86%	70.29%
20-0dB Avg.	14.58%	14.91%	14.93%	14.87%	15.04%	14.92%

Table B.1: *Word Error Rate per SNR for real outputs under multi-modulation spectral feature division.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	34.96%	34.96%	30.67%	35.58%	37.42%	30.06%
20dB	*2.59%*	15.05%	18.91%	11.96%	15.44%	11.96%	10.03%
15dB	*4.25%*	8.70%	2.58%	6.11%	4.70%	3.76%	2.58%
10dB	*8.27%*	-.48%	-7.01%	-7.61%	-5.68%	-7.37%	-5.80%
5dB	*18.74%*	-.37%	-2.34%	-1.49%	-1.60%	-2.77%	-1.65%
0dB	*42.18%*	5.92%	4.43%	4.05%	4.40%	3.48%	4.45%
-5dB	*73.12%*	4.36%	3.89%	3.71%	3.65%	4.45%	3.87%
20-0dB Avg.	*15.20%*	4.07%	1.90%	1.77%	2.17%	1.05%	1.84%

Table B.2: *Improvement relative to AFE baseline for real outputs under multi-modulation spectral feature division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.88%	1.50%	2.83%	1.02%	1.01%	1.02%
20dB	2.17%	3.77%	7.45%	2.60%	2.14%	2.40%
15dB	3.91%	7.00%	12.37%	4.85%	3.91%	4.23%
10dB	8.51%	14.23%	22.32%	10.28%	8.45%	9.50%
5dB	19.14%	28.42%	40.03%	21.66%	18.80%	21.01%
0dB	40.36%	52.90%	63.97%	44.44%	40.25%	43.33%
-5dB	70.19%	78.65%	84.14%	74.06%	70.54%	72.76%
20-0dB Avg.	14.82%	21.26%	29.23%	16.77%	14.71%	16.09%

Table B.3: *Word Error Rate per SNR for imaginary outputs under multi-modulation spectral feature division.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	46.01%	7.97%	-73.61%	37.42%	38.03%	37.42%
20dB	*2.59%*	16.21%	-45.55%	-187.64%	-.38%	17.37%	7.33%
15dB	*4.25%*	8.00%	-64.70%	-191.05%	-14.11%	8.00%	.47%
10dB	*8.27%*	-2.90%	-72.06%	-169.89%	-24.30%	-2.17%	-14.87%
5dB	*18.74%*	-2.13%	-51.65%	-113.60%	-15.58%	-.32%	-12.11%
0dB	*42.18%*	4.31%	-25.41%	-51.65%	-5.35%	4.57%	-2.72%
-5dB	*73.12%*	4.00%	-7.56%	-15.07%	-1.28%	3.52%	.49%
20-0dB Avg.	*15.20%*	2.50%	-39.86%	-92.30%	-10.32%	3.22%	-5.85%

Table B.4: *Improvement relative to AFE baseline for imaginary outputs under multi-modulation spectral feature division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	1.13%	1.00%	1.03%	0.97%	0.97%	1.19%
20dB	1.99%	1.69%	2.11%	1.93%	2.02%	2.20%
15dB	3.73%	3.39%	3.84%	3.88%	3.84%	4.05%
10dB	8.26%	7.67%	8.77%	8.34%	8.21%	8.78%
5dB	19.38%	18.19%	20.37%	19.27%	19.25%	20.23%
0dB	41.85%	41.04%	42.86%	42.20%	40.78%	42.12%
-5dB	71.62%	71.26%	72.29%	72.29%	69.82%	71.30%
20-0dB Avg.	15.04%	14.40%	15.59%	15.12%	14.82%	15.48%

Table B.5: *Word Error Rate per SNR for magnitude outputs under multi-modulation spectral feature division.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	30.67%	38.65%	36.80%	40.49%	40.49%	26.99%
20dB	*2.59%*	23.16%	34.74%	18.53%	25.48%	22.00%	15.05%
15dB	*4.25%*	12.23%	20.23%	9.64%	8.70%	9.64%	4.70%
10dB	*8.27%*	.12%	7.25%	-6.04%	-.84%	.72%	-6.16%
5dB	*18.74%*	-3.41%	2.93%	-8.69%	-2.82%	-2.72%	-7.95%
0dB	*42.18%*	.78%	2.70%	-1.61%	-.04%	3.31%	.14%
-5dB	*73.12%*	2.05%	2.54%	1.13%	1.13%	4.51%	2.48%
20-0dB Avg.	*15.20%*	1.05%	5.26%	-2.56%	.52%	2.50%	-1.84%

Table B.6: *Improvement relative to AFE baseline for magnitude outputs under multi-modulation spectral feature division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.94%	1.05%	0.99%	1.01%	1.05%	1.09%
20dB	1.88%	1.99%	2.03%	1.92%	2.03%	2.09%
15dB	3.44%	3.59%	3.75%	3.44%	3.60%	3.75%
10dB	7.70%	7.90%	8.15%	7.86%	8.09%	8.31%
5dB	17.68%	17.52%	18.35%	17.21%	18.33%	18.51%
0dB	38.55%	38.52%	39.04%	38.38%	39.62%	40.05%
-5dB	69.30%	68.92%	69.61%	69.15%	69.99%	70.41%
20-0dB Avg.	13.85%	13.90%	14.26%	13.76%	14.33%	14.54%

Table B.7: *Word Error Rate per SNR for real and imaginary outputs under multi-modulation spectral feature division. Real and imaginary components are combined via early fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	42.33%	35.58%	39.26%	38.03%	35.58%	33.12%
20dB	*2.59%*	27.41%	23.16%	21.62%	25.86%	21.62%	19.30%
15dB	*4.25%*	19.05%	15.52%	11.76%	19.05%	15.29%	11.76%
10dB	*8.27%*	6.89%	4.47%	1.45%	4.95%	2.17%	-.48%
5dB	*18.74%*	5.65%	6.51%	2.08%	8.16%	2.18%	1.22%
0dB	*42.18%*	8.60%	8.67%	7.44%	9.00%	6.06%	5.04%
-5dB	*73.12%*	5.22%	5.74%	4.80%	5.42%	4.28%	3.70%
20-0dB Avg.	*15.20%*	8.88%	8.55%	6.18%	9.47%	5.72%	4.34%

Table B.8: *Improvement relative to AFE baseline for real and imaginary outputs under multi-modulation spectral feature division. Real and imaginary components are combined via early fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	1.03%	1.10%	0.95%	0.93%	1.07%	1.17%
20dB	1.89%	1.98%	2.07%	1.91%	1.93%	2.12%
15dB	3.31%	3.58%	3.62%	3.31%	3.37%	3.71%
10dB	7.30%	7.90%	8.19%	7.27%	7.39%	8.18%
5dB	17.29%	18.66%	18.97%	17.12%	17.21%	18.37%
0dB	38.96%	41.14%	41.50%	38.88%	38.84%	40.79%
-5dB	68.97%	70.56%	70.36%	69.10%	68.62%	70.43%
20-0dB Avg.	13.75%	14.65%	14.87%	13.69%	13.75%	14.63%

Table B.9: *Word Error Rate per SNR for real, imaginary, and magnitude outputs under multi-modulation spectral feature division. Real, imaginary, and magnitude components are combined via early fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	36.80%	32.51%	41.71%	42.94%	34.35%	28.22%
20dB	*2.59%*	27.02%	23.55%	20.07%	26.25%	25.48%	18.14%
15dB	*4.25%*	22.11%	15.76%	14.82%	22.11%	20.70%	12.70%
10dB	*8.27%*	11.72%	4.47%	.96%	12.09%	10.64%	1.08%
5dB	*18.74%*	7.73%	.42%	-1.22%	8.64%	8.16%	1.97%
0dB	*42.18%*	7.63%	2.46%	1.61%	7.82%	7.91%	3.29%
-5dB	*73.12%*	5.67%	3.50%	3.77%	5.49%	6.15%	3.67%
20-0dB Avg.	*15.20%*	9.53%	3.61%	2.17%	9.93%	9.53%	3.75%

Table B.10: *Improvement relative to AFE baseline for real, imaginary, and magnitude outputs under multi-modulation spectral feature division. Real, imaginary, and magnitude components are combined via early fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	1.02%	1.04%	2.94%	1.02%	0.90%	0.94%
20dB	2.03%	2.45%	8.08%	2.45%	2.04%	2.27%
15dB	3.70%	4.61%	13.03%	4.62%	3.63%	4.02%
10dB	8.01%	9.97%	22.76%	9.83%	8.14%	8.89%
5dB	18.43%	21.52%	39.59%	21.47%	18.34%	19.47%
0dB	39.66%	43.53%	62.64%	43.91%	39.82%	41.06%
-5dB	69.55%	72.18%	83.00%	73.15%	69.57%	70.70%
20-0dB Avg.	14.36%	16.42%	29.22%	16.46%	14.40%	15.14%

Table B.11: *Word Error Rate per SNR for real and imaginary outputs under multi-modulation spectral feature division. Real and imaginary components are combined via late fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	37.42%	36.19%	-80.36%	37.42%	44.78%	42.33%
20dB	*2.59%*	21.62%	5.40%	-211.96%	5.40%	21.23%	12.35%
15dB	*4.25%*	12.94%	-8.47%	-206.58%	-8.70%	14.58%	5.41%
10dB	*8.27%*	3.14%	-20.55%	-175.21%	-18.86%	1.57%	-7.49%
5dB	*18.74%*	1.65%	-14.83%	-111.25%	-14.56%	2.13%	-3.89%
0dB	*42.18%*	5.97%	-3.20%	-48.50%	-4.10%	5.59%	2.65%
-5dB	*73.12%*	4.88%	1.28%	-13.51%	-.04%	4.85%	3.30%
20-0dB Avg.	*15.20%*	5.52%	-8.02%	-92.23%	-8.28%	5.26%	.39%

Table B.12: *Improvement relative to AFE baseline for real and imaginary outputs under multi-modulation spectral feature division. Real and imaginary components are combined via late fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.97%	0.95%	3.5%	0.96%	0.945%	0.88375%
20dB	1.87%	2.17%	10.35%	2.36%	1.93125%	2.00375%
15dB	3.49%	3.92%	16.36%	4.5%	3.4925%	3.67625%
10dB	7.59%	8.81%	27.05%	9.8%	7.49875%	8.17125%
5dB	17.62%	20.01%	44.35%	21.41%	17.895%	18.6363%
0dB	39.53%	41.97%	66.39%	44.07%	39.6262%	40.3875%
-5dB	70%	71.97%	84.81%	72.19%	69.4038%	70.155%
20-0dB Avg.	14.02%	15.38%	32.9%	16.43%	14.08874%	14.57501%

Table B.13: *Word Error Rate per SNR for real, imaginary, and magnitude outputs under multi-modulation spectral feature division. Real, imaginary, and magnitude components are combined via late fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	40.49%	41.71%	-114.72%	41.10%	42.02%	45.78%
20dB	*2.59%*	27.79%	16.21%	-299.61%	8.88%	25.43%	22.63%
15dB	*4.25%*	17.88%	7.76%	-284.94%	-5.88%	17.82%	13.50%
10dB	*8.27%*	8.22%	-6.52%	-227.08%	-18.50%	9.32%	1.19%
5dB	*18.74%*	5.97%	-6.77%	-136.65%	-14.24%	4.50%	.55%
0dB	*42.18%*	6.28%	.49%	-57.39%	-4.48%	6.05%	4.24%
-5dB	*73.12%*	4.26%	1.57%	-15.98%	1.27%	5.08%	4.05%
20-0dB Avg.	*15.20%*	7.76%	-1.18%	-116.44%	-8.09%	7.31%	4.11%

Table B.14: *Improvement relative to AFE baseline for real, imaginary, and magnitude outputs under multi-modulation spectral feature division. Real, imaginary, and magnitude components are combined via late fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.85%	0.97%	1.00%	0.90%	0.85%	0.83%
20dB	1.99%	1.90%	1.99%	2.09%	1.93%	1.82%
15dB	3.66%	3.41%	3.75%	3.79%	3.57%	3.36%
10dB	7.66%	7.38%	8.10%	7.61%	7.75%	7.49%
5dB	17.60%	17.30%	18.27%	17.60%	17.93%	17.58%
0dB	39.52%	39.50%	39.53%	39.33%	38.95%	38.98%
-5dB	70.46%	69.82%	69.36%	69.49%	68.61%	68.61%
20-0dB Avg.	14.08%	13.90%	14.33%	14.08%	14.02%	13.85%

Table B.15: *Word Error Rate per SNR for real outputs under spectro-temporal MFCC division.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	47.85%	40.49%	38.65%	44.78%	47.85%	49.07%
20dB	*2.59%*	23.16%	26.64%	23.16%	19.30%	25.48%	29.72%
15dB	*4.25%*	13.88%	19.76%	11.76%	10.82%	16.00%	20.94%
10dB	*8.27%*	7.37%	10.76%	2.05%	7.98%	6.28%	9.43%
5dB	*18.74%*	6.08%	7.68%	2.50%	6.08%	4.32%	6.18%
0dB	*42.18%*	6.30%	6.35%	6.28%	6.75%	7.65%	7.58%
-5dB	*73.12%*	3.63%	4.51%	5.14%	4.96%	6.16%	6.16%
20-0dB Avg.	*15.20%*	7.36%	8.55%	5.72%	7.36%	7.76%	8.88%

Table B.16: *Improvement relative to AFE baseline for real outputs under spectro-temporal MFCC division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.84%	0.85%	1.03%	0.79%	0.80%	0.92%
20dB	1.92%	1.92%	2.20%	2.11%	1.82%	1.84%
15dB	3.56%	3.45%	3.85%	3.66%	3.17%	3.30%
10dB	7.66%	7.73%	8.19%	7.33%	7.12%	7.49%
5dB	17.57%	17.96%	18.28%	16.70%	16.71%	17.19%
0dB	39.49%	40.10%	39.57%	38.41%	38.02%	38.34%
-5dB	69.93%	70.87%	69.66%	69.20%	68.07%	68.21%
20-0dB Avg.	14.04%	14.23%	14.42%	13.64%	13.37%	13.63%

Table B.17: *Word Error Rate per SNR for imaginary outputs under spectro-temporal MFCC division.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	48.46%	47.85%	36.80%	51.53%	50.92%	43.55%
20dB	*2.59%*	25.86%	25.86%	15.05%	18.53%	29.72%	28.95%
15dB	*4.25%*	16.23%	18.82%	9.41%	13.88%	25.41%	22.35%
10dB	*8.27%*	7.37%	6.52%	.96%	11.36%	13.90%	9.43%
5dB	*18.74%*	6.24%	4.16%	2.45%	10.88%	10.83%	8.27%
0dB	*42.18%*	6.37%	4.93%	6.18%	8.93%	9.86%	9.10%
-5dB	*73.12%*	4.36%	3.07%	4.73%	5.36%	6.90%	6.71%
20-0dB Avg.	*15.20%*	7.63%	6.38%	5.13%	10.26%	12.03%	10.32%

Table B.18: *Improvement relative to AFE baseline for imaginary outputs under spectro-temporal MFCC division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.81%	0.82%	1.07%	0.84%	0.80%	0.93%
20dB	2.52%	2.39%	2.26%	2.48%	2.16%	2.15%
15dB	4.56%	4.24%	4.17%	4.37%	3.83%	3.89%
10dB	8.54%	8.66%	8.70%	8.56%	8.34%	8.54%
5dB	18.66%	19.11%	19.69%	18.84%	18.86%	19.45%
0dB	41.32%	41.56%	42.65%	41.03%	40.89%	41.46%
-5dB	73.26%	72.00%	72.98%	71.67%	70.19%	70.53%
20-0dB Avg.	15.12%	15.19%	15.49%	15.06%	14.82%	15.10%

Table B.19: *Word Error Rate per SNR for magnitude outputs under spectro-temporal MFCC division.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	50.30%	49.69%	34.35%	48.46%	50.92%	42.94%
20dB	*2.59%*	2.70%	7.72%	12.74%	4.24%	16.60%	16.98%
15dB	*4.25%*	-7.29%	.23%	1.88%	-2.82%	9.88%	8.47%
10dB	*8.27%*	-3.26%	-4.71%	-5.19%	-3.50%	-.84%	-3.26%
5dB	*18.74%*	.42%	-1.97%	-5.06%	-.53%	-.64%	-3.78%
0dB	*42.18%*	2.03%	1.46%	-1.11%	2.72%	3.05%	1.70%
-5dB	*73.12%*	-.19%	1.53%	.19%	1.98%	4.00%	3.54%
20-0dB Avg.	*15.20%*	.52%	.06%	-1.90%	.92%	2.50%	.65%

Table B.20: *Improvement relative to AFE baseline for magnitude outputs under spectro-temporal MFCC division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.87%	1.23%	1.57%	0.96%	0.95%	1.61%
20dB	1.79%	2.25%	2.66%	1.84%	1.85%	3.10%
15dB	3.38%	3.87%	4.46%	3.35%	3.29%	5.20%
10dB	7.68%	7.93%	9.08%	7.32%	7.27%	9.53%
5dB	18.11%	18.30%	20.26%	17.70%	17.04%	20.29%
0dB	41.35%	41.24%	43.79%	40.61%	39.38%	42.74%
-5dB	73.29%	71.96%	74.15%	71.43%	69.91%	72.39%
20-0dB Avg.	14.46%	14.72%	16.05%	14.16%	13.77%	16.17%

Table B.21: *Word Error Rate per SNR for real and imaginary outputs under spectro-temporal MFCC division. Real and imaginary components are combined via early fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	46.62%	24.53%	3.68%	41.10%	41.71%	1.22%
20dB	*2.59%*	30.88%	13.12%	-2.70%	28.95%	28.57%	-19.69%
15dB	*4.25%*	20.47%	8.94%	-4.94%	21.17%	22.58%	-22.35%
10dB	*8.27%*	7.13%	4.11%	-9.79%	11.48%	12.09%	-15.23%
5dB	*18.74%*	3.36%	2.34%	-8.11%	5.54%	9.07%	-8.27%
0dB	*42.18%*	1.96%	2.22%	-3.81%	3.72%	6.63%	-1.32%
-5dB	*73.12%*	-.23%	1.58%	-1.40%	2.31%	4.39%	.99%
20-0dB Avg.	*15.20%*	4.86%	3.15%	-5.59%	6.84%	9.40%	-6.38%

Table B.22: *Improvement relative to AFE baseline for real and imaginary outputs under spectro-temporal MFCC division. Real and imaginary components are combined via early fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.92%	1.08%	1.60%	0.93%	0.89%	1.13%
20dB	1.90%	2.03%	2.59%	1.97%	1.79%	2.07%
15dB	3.47%	3.58%	4.10%	3.65%	3.23%	3.55%
10dB	7.95%	7.52%	8.16%	7.73%	7.11%	7.59%
5dB	18.97%	17.75%	18.92%	18.23%	17.47%	17.98%
0dB	42.69%	40.59%	42.44%	41.26%	40.36%	40.69%
-5dB	74.67%	71.61%	72.95%	72.06%	71.21%	70.81%
20-0dB Avg.	14.99%	14.29%	15.24%	14.57%	13.99%	14.37%

Table B.23: *Word Error Rate per SNR for real, imaginary, and magnitude outputs under spectro-temporal MFCC division. Real, imaginary, and magnitude components are combined via early fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	43.55%	33.74%	1.84%	42.94%	45.39%	30.67%
20dB	*2.59%*	26.64%	21.62%	0%	23.93%	30.88%	20.07%
15dB	*4.25%*	18.35%	15.76%	3.52%	14.11%	24.00%	16.47%
10dB	*8.27%*	3.86%	9.06%	1.33%	6.52%	14.02%	8.22%
5dB	*18.74%*	-1.22%	5.28%	-.96%	2.72%	6.77%	4.05%
0dB	*42.18%*	-1.20%	3.76%	-.61%	2.18%	4.31%	3.53%
-5dB	*73.12%*	-2.11%	2.06%	.23%	1.44%	2.61%	3.15%
20-0dB Avg.	*15.20%*	1.38%	5.98%	-.26%	4.14%	7.96%	5.46%

Table B.24: *Improvement relative to AFE baseline for real, imaginary, and magnitude outputs under spectro-temporal MFCC division. Real, imaginary, and magnitude components are combined via early fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.87%	1.01%	1.04%	0.90%	0.78%	0.91%
20dB	2.13%	2.02%	1.99%	2.01%	1.88%	1.79%
15dB	3.76%	3.60%	3.56%	3.55%	3.39%	3.11%
10dB	8.23%	7.66%	7.90%	7.25%	7.43%	7.15%
5dB	18.48%	17.68%	18.13%	16.82%	17.23%	16.82%
0dB	40.17%	39.29%	39.65%	38.51%	38.59%	38.07%
-5dB	70.28%	69.28%	70.19%	69.13%	68.29%	68.53%
20-0dB Avg.	14.55%	14.05%	14.25%	13.63%	13.70%	13.39%

Table B.25: *Word Error Rate per SNR for real and imaginary outputs under spectro-temporal MFCC division. Real and imaginary components are combined via late fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	46.62%	38.03%	36.19%	44.78%	52.14%	44.17%
20dB	*2.59%*	17.76%	22.00%	23.16%	22.39%	27.41%	30.88%
15dB	*4.25%*	11.52%	15.29%	16.23%	16.47%	20.23%	26.82%
10dB	*8.27%*	.48%	7.37%	4.47%	12.33%	10.15%	13.54%
5dB	*18.74%*	1.38%	5.65%	3.25%	10.24%	8.05%	10.24%
0dB	*42.18%*	4.76%	6.85%	5.99%	8.70%	8.51%	9.74%
-5dB	*73.12%*	3.88%	5.25%	4.00%	5.45%	6.60%	6.27%
20-0dB Avg.	*15.20%*	4.27%	7.56%	6.25%	10.32%	9.86%	11.90%

Table B.26: *Improvement relative to AFE baseline for real and imaginary outputs under spectro-temporal MFCC division. Real and imaginary components are combined via late fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log-MLP
clean	0.77%	0.90%	0.97%	0.79%	0.78%	0.88%
20dB	2.14%	1.98%	2.17%	1.87%	1.91%	1.92%
15dB	3.77%	3.41%	3.83%	3.35%	3.51%	3.53%
10dB	7.87%	7.14%	8.34%	7.24%	7.72%	7.86%
5dB	17.90%	16.76%	18.46%	16.88%	17.32%	18.02%
0dB	40.16%	38.48%	40.44%	38.97%	38.36%	39.18%
-5dB	70.59%	69.91%	70.63%	70.01%	67.52%	68.79%
20-0dB Avg.	14.37%	13.55%	14.65%	13.66%	13.76%	14.10%

Table B.27: *Word Error Rate per SNR for real, imaginary, and magnitude outputs under spectro-temporal MFCC division. Real, imaginary, and magnitude components are combined via late fusion.*

SNR	AFE (baseline)	+ST EQ Wt	+ST Geom. Wt	+ST Harm Wt	+ST IE Wt	+ST MLP Wt	+ST Log MLP
clean	*1.63%*	52.76%	44.78%	40.49%	51.53%	52.14%	46.01%
20dB	*2.59%*	17.37%	23.55%	16.21%	27.79%	26.25%	25.86%
15dB	*4.25%*	11.29%	19.76%	9.88%	21.17%	17.41%	16.94%
10dB	*8.27%*	4.83%	13.66%	-.84%	12.45%	6.65%	4.95%
5dB	*18.74%*	4.48%	10.56%	1.49%	9.92%	7.57%	3.84%
0dB	*42.18%*	4.78%	8.77%	4.12%	7.61%	9.05%	7.11%
-5dB	*73.12%*	3.46%	4.39%	3.40%	4.25%	7.65%	5.92%
20-0dB Avg.	*15.20%*	5.46%	10.85%	3.61%	10.13%	9.47%	7.23%

Table B.28: *Improvement relative to AFE baseline for real, imaginary, and magnitude outputs under spectro-temporal MFCC division. Real, imaginary, and magnitude components are combined via late fusion.*

Appendix C

Results per SNR for N95 corpus

One caveat with these results is that with the exception of the clean condition, the per SNR test conditions contain very few words (roughly 800). Numbers may seem artificially inflated or deflated, but is in fact a result of random error. For this reason, relative improvement per SNR results are not presented here.

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.82%	2.77%	2.88%	2.73%
20dB	4.72%	5.23%	5.10%	4.72%
15dB	7.60%	8.22%	8.22%	7.72%
10dB	8.06%	8.44%	9.07%	8.94%
5dB	21.30%	20.92%	21.43%	21.05%
0dB	35.32%	35.07%	35.82%	33.33%
20-0dB Avg.	15.44%	15.62%	15.97%	15.19%

Table C.1: *Word Error Rate per SNR for real outputs under multi-modulation spectral feature division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.61%	2.54%	2.44%	2.59%
20dB	4.46%	4.97%	4.72%	4.59%
15dB	8.47%	8.84%	9.34%	8.97%
10dB	9.95%	9.45%	10.08%	10.08%
5dB	19.77%	20.66%	20.79%	20.41%
0dB	37.31%	36.07%	38.06%	36.44%
20-0dB Avg.	16.05%	16.05%	16.65%	16.15%

Table C.2: *Word Error Rate per SNR for imaginary outputs under multi-modulation spectral feature division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.12%	2.23%	2.29%	2.31%
20dB	4.59%	4.59%	4.34%	4.85%
15dB	8.84%	7.72%	8.22%	7.97%
10dB	10.20%	8.31%	9.32%	10.33%
5dB	21.43%	22.07%	22.07%	21.81%
0dB	37.56%	35.70%	37.31%	38.06%
20-0dB Avg.	16.58%	15.72%	16.30%	16.65%

Table C.3: *Word Error Rate per SNR for magnitude outputs under multi-modulation spectral feature division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.99%	2.84%	2.61%	2.84%
20dB	5.74%	4.97%	5.10%	4.85%
15dB	8.22%	7.97%	8.34%	8.47%
10dB	8.82%	9.19%	9.45%	9.82%
5dB	22.45%	22.70%	22.07%	21.05%
0dB	37.19%	35.95%	37.31%	36.82%
20-0dB Avg.	16.53%	16.20%	16.50%	16.25%

Table C.4: *Word Error Rate per SNR for real and imaginary outputs under multi-modulation spectral feature division. Real and imaginary components are combined via late fusion.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.86%	2.86%	2.82%	2.65%
20dB	5.23%	4.97%	5.23%	5.87%
15dB	6.85%	6.97%	6.72%	6.23%
10dB	8.69%	9.57%	8.44%	9.19%
5dB	19.13%	19.39%	19.90%	20.41%
0dB	32.96%	35.07%	33.33%	33.83%
20-0dB Avg.	14.61%	15.24%	14.76%	15.14%

Table C.5: *Word Error Rate per SNR for real outputs under spectro-temporal MFCC division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.65%	2.94%	2.56%	2.56%
20dB	6.25%	5.74%	5.99%	5.74%
15dB	6.10%	6.85%	6.48%	7.60%
10dB	9.19%	8.82%	9.82%	9.82%
5dB	20.54%	20.28%	20.92%	21.56%
0dB	33.33%	34.45%	34.95%	34.83%
20-0dB Avg.	15.12%	15.27%	15.67%	15.95%

Table C.6: *Word Error Rate per SNR for imaginary outputs under spectro-temporal MFCC division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.84%	3.01%	3.07%	2.90%
20dB	5.61%	5.99%	5.61%	5.61%
15dB	7.72%	7.35%	7.10%	6.85%
10dB	8.82%	8.19%	8.94%	9.82%
5dB	21.05%	20.79%	20.54%	20.92%
0dB	35.57%	35.32%	34.45%	35.07%
20-0dB Avg.	15.80%	15.57%	15.37%	15.70%

Table C.7: *Word Error Rate per SNR for magnitude outputs under spectro-temporal MFCC division.*

SNR	+ST EQ Wt	+ST Geom. Wt	+ST IE Wt	+ST MLP Wt
clean	2.65%	2.86%	2.88%	2.92%
20dB	5.61%	5.48%	5.36%	5.74%
15dB	7.22%	6.35%	6.60%	7.10%
10dB	9.45%	9.19%	9.07%	9.19%
5dB	19.77%	19.52%	20.92%	21.81%
0dB	33.46%	33.58%	33.46%	31.47%
20-0dB Avg.	15.14%	14.87%	15.12%	15.09%

Table C.8: *Word Error Rate per SNR for real and imaginary outputs under spectro-temporal MFCC division. Real and imaginary components are combined via late fusion.*

Bibliography

[1] Agarwal, A., Cheng, Y.M., "Two-stage Mel-warped Wiener Filter for Robust Speech Recognition". The 1999 International Workshop on Automatic Speech Recognition and Understanding, pp. 67-70, 1999.

[2] Bourlard, H. and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands", In Proc. of Intl. Conf. on Spoken Language Processing, Philadelphia, PA, pp. 422-425, 1996.

[3] Chi, T., Gao, Y., Guyton, M.C., Ru, P., and Shamma, S.A., "Spectro-temporal modulation transfer functions and speech intelligibility", J. Acoust. Soc. Am., 106(5):2719-2732, 1999.

[4] Chiu, Y.-H. B. , Raj, B. , and Stern, R. M., "Learning-based auditory encoding for robust speech recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, March 2010, Dallas, Texas.

[5] Cole, R., Fanty, M., Noel, M. and Lander, T. "Telephone speech corpus development at CSLU", in Proc. Int. Conf. Spoken Lang. Proc., Yokohama, Japan, pp. 1815-1818, 1994.

[6] Davis, S.B., and Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pp. 357366. 1980.

[7] DeCharms R., Blake D., and Merzenich M. "Optimizing sound features for cortical neurons," Science 280: 14391443, 1998.

[8] Depireux, D.A., Simon, J.Z., Klein, D.J., and Shamma, S.A., "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex, J. Neurophysiology, 85:1220-134, 2001.

[9] Domont, X., Heckmann, M., Joublin, F., Goerick, C., "Hierarchical spectro-temporal features for robust speech recognition", In Proc. ICASSP, Las Vegas, USA, pp. 4417-4420, 2008.

[10] Ellis, D., and Morgan, N. "Size Matters: An Empirical Study of Neural Network Training for Large Vocabulary Continuous Speech Recognition". Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999), Phoenix, Arizona, pp. II-1013-1016.

[11] Faria, A., and Morgan, N. "Corrected Tandem Features for Acoustic Model Training". Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, pp. 4737-4740.

[12] ETSI standard doc. "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Feature Extraction Algorithm", ETSI ES 202 050 Ver.1.1.1 (2002-10).

[13] Gelbart, D., "Noisy numbers data and numbers testbeds", International Computer Science Institute, Berkeley, CA. http://www.icsi.berkeley.edu/speech/papers/gelbart-ms/.

[14] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am., 87(4), pp. 1738-1752. (1990)

[15] Hermansky, H., Tibrewala, S., and Pavel, M., "Towards ASR on Partially Corrupted Speech". In the Proceedings of International Conference on Spoken Language Processing, Pittsburgh, Pennsylvania, 1996.

[16] Hermansky, H., Ellis, D., Sharma, S., "Tandem connectionist feature extraction for conventional HMM systems", in Proc. ICASSP, Istanbul, Turkey, pp. 1635-1638, 2000.

[17] Hermansky, H., Fousek, P., "Multi-resolution rasta filtering for tandem-based asr", In Proceedings of Interspeech, Lisbon, Portugal, pp. 361-364, 2005.

[18] Hirsch, H.G., and Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in ISCA ITRW ASR: Challenges for the Next Millennium, Paris, France, pp. 18-20, 2000.

[19] Kanedera, N., Arai, T., Hermansky, H., Pavel, M., "On the relative importance of various components of the modulation spectrum for automatic speech recognition", Speech Communication, 28:43-55, 1999.

[20] Kingsbury, B.E.D., and Morgan, N., "The modulation spectrogram: In pursuit of an invariant representation of speech". Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1259-62, 1997

[21] Klein, D.J., Depireux, D.A., Simon, J.Z., Shamma, S.A., "Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design, J. Comp. Neuroscience, 9:85-111, 2000.

[22] Kleinschmidt, M., "Localized spectro-temporal features for automatic speech recognition", in Proceedings of Eurospeech, pp. 2573-2576, 2003.

[23] Kowalski N., Depireux D., and Shamma S. "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single unit responses to moving ripple spectra." J Neurophysiol 76: 35033523, 1996a.

[24] Kowalski N., Depireux D., and Shamma S. "Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra." J Neurophysiol 76: 35243534, 1996b.

[25] Lei, X., Siu, M., Hwang, M.Y., Ostendorf, M., and Lee, T. "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition", in Proc. of Intl. Conf. of Spoken Language Processing, Pittsburgh, PA, pp. 1237-1240, 2006.

[26] Mesgarani, N., Slaney, M., and Shamma, S., "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", IEEE Trans. Audio, Speech, and Language Proc., 14(3):920-929, 2006.

[27] Mesgarani, N., Thomas, S., and Hermansky, H., "A Multistream Multiresolution Framework for Phoneme Recognition", In proceedings of Interspeech, Murakami, Japan, 2010.

[28] Misra, H., Bourlard, H., Tyagi, V., "New entropy based combination rules in HMM/ANN multi-stream ASR, in Proc. ICASSP, pp. II-741-4 vol.2, Hong Kong, 2003.

[29] Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., Bourlard, H., and Athineos, M., "Pushing the envelope - aside", IEEE Signal Processing Magazine, 22(5):81-88, 2005.

[30] Ravuri, S., and Morgan, N., "Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR", In proceedings of Interspeech, Murakami, Japan, 2010.

[31] Schreiner C., and Cahoun B. "Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions," Aud Neurosci 1:3961, 1994.

[32] Shamma S., Versnel H., and Kowalski N. "Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra," Aud Neurosci 1: 233254, 1995.

[33] Valente, H. and Hermansky, H., "On the combination of auditory and modulation frequency channels for ASR applications", In Proceedings of Interspeech, Brisbane, Australia, pp. 2242-2245, 2008.

[34] Zhao, S.Y., Morgan, N. "Multi-stream spectro-temporal features for robust speech recognition", In Proceedings of Interspeech, Brisbane, Australia, pp. 898-901, 2008.

[35] Zhao, S., Ravuri, S., and Morgan, N. "Multi-Stream to Many-Stream: Using Spectro-temporal Features for ASR", In Proceedings of Interspeech, Brighton, UK, pp. 2951-2954, 2009.