Speech Recognition System Robustness to Microphone Variations

by

Jane W. Chang

B.S., Electrical Engineering, Stanford University, 1992

Submitted to the Department of Electrical Engineering
and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1995

Author .....
Department of Electrical Engineering
and Computer Science
January 20, 1995

Certified by .....
Victor W. Zue
Senior Research Scientist
Thesis Supervisor

Accepted by .....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Speech Recognition System Robustness to Microphone Variations

by

Jane W. Chang

Submitted to the Department of Electrical Engineering

and Computer Science

on January 20, 1995, in partial fulfillment of the

requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

Abstract

This thesis seeks to improve our understanding of the effects of microphone variations and compensation techniques on a speech recognition system, focusing on mismatched conditions when the testing microphone is of lower quality than the training microphone. A methodology is designed to enable the isolation of microphone effects and the benchmarking and comparison of techniques. The tasks are phonetic classification and recognition. The corpora and systems are respectively configured from TIMIT [10] and SUMMIT [36].

TIMIT provides three micropone recordings, the close-talking Sennheiser, far-field B&K and Telephone, that are particularly useful for experiments on microphone variations. Baseline analyses show that the Sennheiser and B&K differ mainly at low frequencies and result in moderate performance degradations under mismatched conditions. In comparison, the Telephone shows larger deviations, and even after downsampling to remove differences at high frequencies, still suffers severe performance degradations.

In reducing the errors due to mismatched testing, the thesis focuses on preprocessing techniques that compensate for microphone effects prior to recognition. Several preprocessing techniques are compared and analyzed. The most effective techniques significantly compensate for the relatively small differences and reduce error rates for the B&K and slightly reduce the larger mismatch and error rates for the Telephone. The thesis also explores further increases in microphone robustness that can be achieved by techniques that incorporate microphone-specific data in training.

Thesis Supervisor: Victor W. Zue

Title: Senior Research Scientist

Speech Recognition System Robustness to Microphone Variations

by

Jane W. Chang

Submitted to the Department of Electrical Engineering

and Computer Science

on January 20, 1995, in partial fulfillment of the

requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

Abstract

This thesis seeks to improve our understanding of the effects of microphone variations and compensation techniques on a speech recognition system, focusing on mismatched conditions when the testing microphone is of lower quality than the training microphone. A methodology is designed to enable the isolation of microphone effects and the benchmarking and comparison of techniques. The tasks are phonetic classification and recognition. The corpora and systems are respectively configured from TIMIT [10] and SUMMIT [36].

TIMIT provides three micropone recordings, the close-talking Sennheiser, far-field B&K and Telephone, that are particularly useful for experiments on microphone variations. Baseline analyses show that the Sennheiser and B&K differ mainly at low frequencies and result in moderate performance degradations under mismatched conditions. In comparison, the Telephone shows larger deviations, and even after downsampling to remove differences at high frequencies, still suffers severe performance degradations.

In reducing the errors due to mismatched testing, the thesis focuses on preprocessing techniques that compensate for microphone effects prior to recognition. Several preprocessing techniques are compared and analyzed. The most effective techniques significantly compensate for the relatively small differences and reduce error rates for the B&K and slightly reduce the larger mismatch and error rates for the Telephone. The thesis also explores further increases in microphone robustness that can be achieved by techniques that incorporate microphone-specific data in training.

Thesis Supervisor: Victor W. Zue

Title: Senior Research Scientist

Acknowledgments

I deeply thank Victor Zue, my thesis supervisor, for his advice, support and encouragement. His valuable teachings have and continue to guide me through work and life.

I also thank all of the Spoken Language Systems group: Jim Glass for providing answers and encouragement; Mike Phillips for instructing me on the recognizer; Mike McCandless for improving the recognizer; Stephanie Seneff for helping me with my writing; Joe Polifroni for reading and encouraging; Christine Pao for maintaining the systems; Vicky Palay and Sally Lee for keeping things in order; my first officemates, Dave Goddeau and Bill Goldenthal, for showing me that students do graduate; and everyone else for providing an excellent research environment.

I thank Rich Cox and others at AT&T Bell Laboratories for their support.

Finally, I thank my family, Mom, Dad, Kay and Fay, for their enduring love that has sustained me through all of my endeavors.

This research was supported by an AT&T Bell Laboratories GRPW Fellowship and Department of Defense Contract MDA904-93-C-4180.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Historically, many factors have impeded the deployment of speech recognition technology. One factor is accuracy, in that speech recognition systems must be able to achieve low error rates in order to perform their intended tasks in deployment. Another factor is robustness, in that systems must also be able to maintain low error rates under conditions that may vary in deployment. With the progress of speech recognition technology, systems have attained high accuracy under testing conditions that are well matched to the conditions used in training. Yet, systems still cannot maintain such accuracy under mismatched training and testing conditions.

Lack of robustness to variations in testing continues to impede the deployment of speech recognition technology. For example, the speaker, environment and microphone can all contribute to variations in the input signal to the speech recognition system. The speaker may vary what or how he speaks. The environment may vary in reverberation or noise level. The microphone may vary in transductional or positional characteristics. Under such deployment conditions, current speech recognition systems cannot maintain low error rates to perform their intended tasks.

1.1 Background

This thesis studies speech recognition system robustness to microphone variations. The microphone can have large effects on the speech recognition system. Even with the same speaker and environment, different microphones record different signals for input to the system. For example, microphones use different transduction principles, such as pressure or pressure gradience, which alter the signal in different ways. The positioning of the microphone also causes distortion. Generally, as the distance relative to the speaker's mouth increases, the microphone records less oral resonance and more non-oral sounds, such as nasal and glottal resonances and environmental noise.

In order to achieve the lowest error rates, most speech recognition systems are trained and tested using a high quality, head-mounted, close-talking, noise-canceling microphone. While such microphones may be suitable for some applications, other transducers, such as hand-, lapel-, table- or boom-mounted microphones and telephones, may be more suitable for other applications. Nevertheless, current speech recognition systems lack robustness to microphone variations and cannot be used with microphones that are mismatched to the one used in training. For example, preliminary experiments show a 150% increase in word error rate, from 28% under the matched condition when training on a Sennheiser to 69% under the mismatched condition when testing on a table-mounted Crown microphone, on the Air Travel Information Service (ATIS) corpus using the SUMMIT [36, 37] system developed at MIT. Table 1.1 shows similar increases in word error rate in percent from matched to mismatched training and testing conditions for various microphones, corpora and systems. Regardless of the microphones or corpora used, all systems suffer at least a 150% increase in error rate from matched to mismatched conditions.

[1] ATIS is the Air Travel Information Service corpus.
[2] Sennheiser refers to a head-mounted microphone.
[3] Crown refers to a table-mounted microphone.
[4] WSJ is the Wall Street Journal corpus.
[5] Assorted refers to the secondary microphones, including lapel-, table-, and boom-mounted microphones, telephones and speaker phones, used to collect the Wall Street Journal (WSJ) corpus.
[6] Shure refers to a unidirectional microphone.
[7] Realistic refers to a unidirectional microphone.

System	Corpus	Train	Test	Match	Mismatch	Increase
SUMMIT (MIT)	ATIS[1]	Sennheiser[2]	Crown[3]	28	69	150
DECIPHER (SRI) [27]	ATIS	Sennheiser	Crown	23	91	300
SPHINX (CMU) [23]	WSJ[4]	Sennheiser	Assorted[5]	8	39	390
BYBLOS (BBN) [3]	WSJ	Sennheiser	Assorted	12	40	230
TANGORA (IBM) [6]	private	Shure[6]	Realistic[7]	2	8	300

Table 1.1: Increase in word error rate in percent from matched to mismatched training and testing conditions for various microphones, corpora and systems

1.2 Previous Work

As the severe performance degradations incurred by microphone variations become apparent, researchers have begun to address the issue of microphone robustness. Many techniques have been developed to reduce the performance degradations incurred by mismatched training and testing conditions. The most common are preprocessing techniques that apply signal processing algorithms to the recorded signal in order to compensate for microphone variations before input to the speech recognition system. Other techniques compensate for microphone variations as part of the recognition process.

1.2.1 Preprocessing Techniques

Preprocessing techniques apply speech enhancement algorithms [21] to compensate for the effects of microphone variations on the recorded signal. Most microphone effects are mathematically modeled by convolution and addition in the time domain. For example, variations in the speaker vocal tract, environmental acoustics and microphone transfer function have convolutional effects. Variations in environmental noise have additive effects.

Convolutional Effects

Some techniques focus on compensating for the convolutional effects of microphone variations on the recorded signal. These techniques often take advantage of the cor-

respondence of convolution in the time domain to addition in the cepstral domain and estimate cepstral compensation vectors to subtract the microphone effects. For example, Mean Normalization (MN) [3, 23] uses the cepstral mean of each utterance as its compensation vector. Relative Spectral Processing (RASTA) [14] applies an exponentially decaying highpass filter to the cepstral vectors of each utterance. Other filters, such as bandpass filters [19], have also been applied.

Additive Effects

Other techniques focus on compensating for the additive effects of microphone variations on the recorded signal. These techniques often estimate spectral or log spectral compensation vectors to subtract the microphone effects. For example, RASTA [16] filtering has been applied to the log spectral vectors of each utterance. Some techniques discriminate between speech and noise to estimate compensation vectors. For example, Spectral Subtraction [33] and Log-Spectral Subtraction (SUB) [34] respectively use histograms in the spectral and log spectral domains to determine speech and noise thresholds. Other techniques [7] apply optimal algorithms such as Minimum Mean Square Error (MMSE) estimation to determine compensation vectors. For example, Minimum Mean Log-Spectral Distance [8] applies MMSE estimation to minimize log spectral distance.

Combined Effects

Combined techniques compensate for both the convolutional and additive effects of microphone variations on the recorded signal. Some techniques combine independently estimated compensation vectors. For example, MN and SUB have been combined in cascade [1]. Other techniques estimate joint compensation vectors. For example, Linear-Logarithmic (LIN-LOG) RASTA [15] uses logarithmic transforms to estimate non-linear compensation vectors. Many jointly combined techniques [1] apply Vector Quantization (VQ) algorithms. For example, Adaptive Labeling (AL) [28] and Tied Mixture Normalization (TMN) [3] apply VQ codebook transformations to

adapt different microphone training and testing conditions.

Researchers at CMU have developed the largest number of techniques for increasing microphone robustness. SNR-Dependent Cepstral Normalization (SDCN) [1] and Phone-Dependent Cepstral Normalization (PDCN) [23] respectively use instantaneous Signal-to-Noise Ratios (SNRs) and preliminary phone hypotheses to estimate compensation vectors. Codebook-Dependent Cepstral Normalization (CDCN) [1] applies Maximum Likelihood (ML) estimation to determine convolutional and additive parameters and MMSE esimation to minimize VQ codeword distances. Fixed CDCN [2] (FCDCN) combines SDCN SNR measurements with CDCN VQ codewords. Multiple FCDCN (MFCDCN) combines multiple FCDCN techniques for different microphones. Interpolated MFCDCN (IMFCDCN) interpolates between MFCDCN compensation vectors.

Results

Preprocessing techniques reduce the performance degradations incurred by microphone mismatches. For example, preliminary experiments show a 36% decrease in word error rate, from 69% to 44%, when using either MN or CDCN under the first mismatched condition described in Table 1.1. Yet, despite such error reductions, preprocessing techniques cannot fully recover the increased error rates caused by mismatched conditions. For example, given that mismatched testing incurs a 150% increase in error, a 36% decrease after preprocessing still results in a net 60% increase in error from the matched condition. Table 1.2 shows similar decreases in word error rate in percent under mismatched conditions and net increases in word error rate in percent from matched to mismatched conditions when using various preprocessing techniques for the microphones, corpora and systems described in Table 1.1. Regardless of the techniques, microphones or corpora used, all systems suffer at least a net 60% increase in error rate from matched to mismatched conditions even after preprocessing.

System	Corpus	Train	Test	Technique	Without Technique	With Technique	Decrease	Net Increase
SUMMIT	ATIS	Sennheiser	Crown	MN	69	44	36	60
				CDCN	69	44	36	60
DECIPHER	ATIS	Sennheiser	Crown	RASTA	91	62	32	170
SPHINX	WSJ	Sennheiser	Assorted	RASTA	39	28	28	250
				MN	39	21	46	160
				MFCDCN	39	15	62	90
				IMFCDCN	39	15	62	90
				PDCN	39	16	59	100
BYBLOS	WSJ	Sennheiser	Assorted	MN	40	32	20	160
				TMN	40	21	47	70
TANGORA	private	Shure	Realistic	AL	8	4	50	100

Table 1.2: Decrease in word error rate in percent under mismatched conditions and net increase in word error rate in percent from matched to mismatched conditions for various techniques, microphones, corpora and systems.

1.2.2 Other Techniques

Other techniques compensate for microphone variations as part of the speech recognition process. For example, feature extraction techniques [25], such as those based on auditory models, can be applied to extract more robust features and produce more robust models. Auditory models [32, 12, 17] approximate the characteristics of the human auditory system and may capture some of the robustness of the human recognition system. Training techniques can also be applied to reduce mismatch and produce more robust models. For example, multi-style training [22] trains the speech recognition system on multiple speaking styles in order to increase robustness to mismatched conditions when testing on abnormal speaking styles, such as stressed speech. Multi-style training on different microphones may also increase microphone robustness.

1.2.3 Discussion

Although researchers have begun to address the issue of microphone robustness, understanding of the effects of microphone variations and techniques on the speech recognition system is still lacking. This need for improved understanding is related to the lack of comparative study, despite numerous efforts towards increasing microphone robustness.

As shown in Table 1.2, researchers use many different techniques, microphones, corpora, and systems. For example, techniques vary in their data requirements. While some techniques do not require microphone data and can be applies to many microphones, other microphone-specific techniques require simultaneously recorded microphone training data and apply only to the trained microphones.

Tasks also vary, from phonetic classification and recognition, to isolated and continuous speech word recognition. Many experiments are performed in word recognition. These experiments are particularly difficult to compare, due to confounding factors in word recognition. For example, corpora use different vocabularies, and systems use different language models. Word recognition experiments are also particularly difficult to perform due to computational requirements. For example, CMU [1] developed many techniques on their private Alphanumeric corpus instead of the Wall Street Journal (WSJ) corpus because the computation for the larger WSJ corpus would have been prohibitive.

These differences in tasks, corpora and systems confound understanding and comparison. With so many other variations between experiments, the particular effects of the microphone variations are obscured, making it difficult, if not impossible, to compare and understand the effects of different techniques.

1.3 Objective

The objective of this thesis is to improve our understanding of the effects of microphone variations and compensation techniques on the speech recognition system. To this end, the thesis performs a comparative study, with a focus on realistic mismatched conditions in deployment, where the testing microphone is of lower quality than the training microphone.

First, the thesis designs a methodology in order to enable the isolation, analysis and comparison of microphone variations and techniques. The TIMIT [9, 10, 20] corpus and SUMMIT [36, 37] system are configured for experiments in phonetic classification and recognition under different microphone training and testing con-

ditions. These experiments focus on fundamental effects of microphone variations and techniques at the phonetic level and reduce confounding effects of corpus and system dependent variables at the word level. They require shorter training and testing cycles and allow generalization from classification to recognition. They also use a commonly accepted corpus designed for acoustic-phonetic experiments to provide baseline comparison and analysis.

Using this methodology, the thesis benchmarks and compares a wide range of techniques in order to understand their effects on the speech recognition system. The techniques are implemented and developed with attention to data requirements. The thesis focuses on preprocessing techniques that do not require microphone-specific data. Analysis of these techniques reveals their ability to compensate for microphone effects on the recorded signal and reduce the performance degradations incurred by mismatched training and testing conditions. The thesis also considers training techniques that require microphone-specific data to understand further increases in microphone robustness that can be achieved.

Overall, the thesis is directed towards fundamental improvements in understanding and performance, with the expectation that increased microphone robustness at the phonetic level will generalize to other tasks and domains.

1.4 Outline

The remainder of the thesis contains six chapters and an appendix. Chapter 2 covers the experimental methodology. A methodology is designed in order to perform a comparative study. The tasks are phonetic classification and recognition. The corpora and systems are respectively configured from TIMIT [9, 10, 20] and SUMMIT [36, 37]. The evaluation is based on error rate and statistical significance.

Chapter 3 covers the preliminary data analysis. An analysis of the microphones and data serves as the basis for understanding the effects of microphone variations and techniques. The microphones are described. The effects of the microphones on the recorded signal are modeled. The signal to noise, spectrographic, spectral and

cepstral characteristics of the microphone data are examined.

Chapter 4 presents the baseline experiments in phonetic classification and recognition for different training and testing conditions. These experiments are analyzed to understand the effects of microphone variations and provide a baseline for experiments with compensation techniques. General results are discussed, and mismatched conditions are analyzed.

Chapter 5 presents the experiments with preprocessing techniques. These experiments are analyzed and compared, in order to understand the effects of preprocessing without microphone-specific data on the speech recognition system. Techniques are described, general results are compared and the most effective techniques are analyzed.

Chapter 6 presents the experiments with training techniques. These experiments are analyzed and compared, in order to understand the effects of training with microphone-specific data on the speech recognition system. Techniques are described, and results are compared.

In Chapter 7, the thesis is summarized, and future work is discussed. Appendix A provides more experimental results for various preprocessing techniques.

Chapter 2

Methodology

Despite efforts towards microphone robustness, differences in methodology confound understanding of the effects of microphone variations and compensation techniques on the speech recognition system. For example, tasks vary from phonetic classification to word recognition. Corpora vary from small private corpora to large standard corpora. These differences obscure the effects of microphone variations and techniques and confound comparison.

This thesis designs a consistent experimental methodology for a comparative study. This methodology enables the isolation of microphone effects and the benchmarking of techniques. The tasks are phonetic classification and recognition. The microphone corpora and phonetic classification and recognition systems are respectively configured from the TIMIT [9, 10, 20] corpus and SUMMIT [36, 37] system. The evaluation is based on error rate and statistical significance.

2.1 Task

The tasks are phonetic classification and recognition. Phonetic classification requires the determination of the phonetic identity of a segment given the signal and its endpoints. Classification involves signal representation, feature extraction and acoustic modeling. Phonetic recognition requires the determination of a phonetic string given

the signal only. Recognition combines classification with segmentation and search.

Experiments on the phonetic level have many advantages. They focus on the fundamental units of sound in speech. They reduce the confounding effects of corpus and system dependent variables, such as vocabulary and language models. They require less computation and shorter training and testing cycles. They also allow generalization in analysis from classification to recognition.

2.2 Corpus

The corpora are configured from TIMIT [9, 10, 20], a collection of read speech with time-aligned phonetic and orthographic transcriptions. Experiments on TIMIT have many advantages. TIMIT is specifically designed for the acquisition of acoustic-phonetic knowledge and the development of phonetic recognition systems. TIMIT is also commonly accepted for benchmarking and comparison. Most importantly, TIMIT is recorded on three different microphones. These recordings of identical utterances spoken by identical speakers allow comparison of differences incurred by variations in microphone only. With these recordings, TIMIT provides three microphone corpora that are particularly useful for phonetic experiments on microphone variations.

2.2.1 Data

TIMIT was collected from 630 speakers, 70% male and 30% female, covering 8 major dialects of American English. Each speaker read 10 utterances, 2 "sa" dialect utterances that were read by all 630 speakers, 5 of the 450 "sx" phonemically compact utterances that were each read by 7 speakers, and 3 of the 1890 "si" phonetically diverse utterances that were each read by only 1 speaker, for a total of 6300 utterances.

2.2.2 Transcriptions

TIMIT provides time-aligned acoustic-phonetic transcriptions for all utterances. Table 2.1 shows the 61 TIMIT acoustic-phonetic symbols with their International Phonetic Alphabet (IPA) symbols and example occurrences.

2.2.3 Microphones

The TIMIT microphone corpora are referred to as the Sennheiser, B&K and Telephone. The first two corpora were recorded in stereo with a head-mounted Sennheiser model HMD414 on one channel and a boom-mounted Bruel and Kjaer (B&K) model 4165 on the other. A quiet environment was maintained using a noise-isolated sound booth. Nevertheless, the B&K recorded a low frequency acoustic rumble that was later removed with a 70 Hz cutoff high pass filter. The third corpus [18] was recorded after transmitting the Sennheiser data over a telephone network. A telephone environment was simulated using an artificial mouth, a telephone handset and local and long distance telephone lines.

The Sennheiser and Telephone corpora were respectively released as TIMIT [9, 10, 20] and Network TIMIT (NTIMIT) [18], but the B&K corpus was never released. In order to configure TIMIT for experiments on microphone variations, the B&K data were acquired from archived tapes. Data from 97.3%, 613 out of 630, speakers were read from tape, but data from the remaining 2.7%, 17 out of 630, speakers could not be recovered[1].

2.2.4 Subsets

TIMIT training and testing subsets have been determined by the National Institute of Standards and Technology (NIST) [10] based on several criteria. The sets do not include the "sa" utterances that were read by all speakers, nor do they share speakers or "sx" and "si" utterances. Each set covers all dialects and phonemes in different

[1]The 17 unread speakers were fajw0, fbmh0, fjem0, fjwb0, flet0, mcal0, mcmb0, mdac0, mdas0, mdwd0, mgjf0, mjjb0, mrkm0, mrvg0, msfh1, msfv0 and msjs1.

IPA	TIMIT	Example	IPA	TIMIT	Example
a	aa	b*o*t	|	ix	deb*i*t
@	ae	b*a*t	i	iy	b*ee*t
^	ah	b*u*t	J	jh	*j*oke
O	ao	b*ough*t	k	k	*k*ey
a˚	aw	b*ou*t	k	kcl	k closure
{	ax	*a*bout	l	l	*l*ay
{˚	ax-h	s*u*spect	m	m	*mom*
}	axr	butt*er*	n	n	*noon*
a˚	ay	b*i*te	4	ng	si*ng*
b	b	*b*ee	F	nx	wi*nn*er
b	bcl	b closure	o	ow	b*oa*t
C	ch	*ch*oke	O˚	oy	b*oy*
d		*d*ay	p	p	*p*ea
d	dcl	d closure	ˆ	pau	pause
D	dh	*th*en	p	pcl	p closure
F	dx	mu*dd*y	?	q	ba*t*
E	eh	b*e*t	r	r	*r*ay
l̃E	el	bott*le*	s	s	*s*ea
m̃E	em	botto*m*	S	sh	*sh*e
ñE	en	butto*n*	t	t	*t*ea
4̃E	eng	Wash*ing*ton	t	tcl	t closure
•	epi	epenthetic silence	T	th	*th*in
5	er	b*ir*d	U	uh	b*oo*k
e	ey	b*a*it	u	uw	b*oo*t
f	f	*f*in	u	ux	t*oo*t
g	g	*g*ay	v	v	*v*an
g	gcl	g closure	w	w	*w*ay
h	hh	*h*ay	y	y	*y*acht
H	hv	a*h*ead	z	z	*z*one
I	ih	b*i*t	Z	zh	a*z*ure
-	h#	utterance initial and final silence			

Table 2.1: TIMIT acoustic-phonetic symbols with their IPA symbols and example occurrences

contexts. The "core" testing set contains data from 24 speakers, 2 male and 1 female of each dialect, each of whom spoke 8 different utterances, for a total of 192 testing utterances. The training set contains data from 462 speakers, each of whom spoke 8 utterances not included in the testing set, for a total of 3696 utterances.

In determining subsets for experiments on microphone variations, the criterion is to maintain consistency across microphones. Because the B&K data are incomplete, the subsets only include NIST utterances that could be acquired for the B&K. All NIST testing utterances were acquired. However, utterances for 11 NIST training speakers could not be read, and an additional 5 NIST training utterances were found to be corrupted[2]. The resulting training and testing sets are respectively 97.5% and 100% of the NIST subsets. A development set is also determined from the speakers and utterances not included in the training and testing sets. Table 2.2 describes the training, testing and development sets.

	# phonemes	# utterances	# speakers
Training set	139,257	3,603	451
Testing set	7330	192	24
Development set	12,978	383	48

Table 2.2: Training, testing and development sets

2.3 System

The classification and recognition systems are configured from the Speech Understanding by Machine at MIT (SUMMIT) [36, 37] system. SUMMIT is a segment-based system that explicitly detects phonetic segment boundaries in order to extract features in relation to specific acoustic events. The classification system involves signal representation, feature extraction and acoustic modeling. The recognition system combines these components with segmentation and search. Consistency is maintained

[2]The corrupted utterances were si1368 by mjpm0, si1412 by mppc0, si2151 by mtdp0, sx90 by mrmg0 and sx107 by mtkd0.

across systems in order to allow generalization in analysis. In addition, simple parameters are used in order to expedite training.

2.3.1 Signal Representation

The classification and recognition systems use a Mel-Frequency Cepstral Coefficient (MFCC) [24, 25] signal representation. This representation is both effective and efficient.

Short Time Fourier Analysis

Given a signal, amplitude normalization is applied to remove differences in recording amplitudes. The signal is multiplicatively scaled such that the maximum sample is 16 bits. Then, preemphasis is applied to enhance higher frequency components and attenuate lower frequency components. Equation 2.1 shows the first difference preemphasis.

$$y[m] = x[m] - \alpha x[m-1] \tag{2.1}$$

where

$x[m]$: original signal

$y[m]$: preemphasized signal

$\alpha = 0.97$

A Short Time Fourier Transform (STFT) is applied to produce a time dependent spectral representation. At an analysis rate of 200 Hz, the normalized and preemphasized signal is windowed using a 25.6 ms Hamming window, and the windowed signal is transformed using a 512 point FFT, to produce 1 frame of spectral coefficients every 5 ms.

Spectral Representation

Given a frame of spectral coefficients, an auditory filter bank is applied to produce the Mel-Frequency Spectral Coefficient (MFSC) [25] representation. Figure 2-1 shows

the MFSC filter bank.

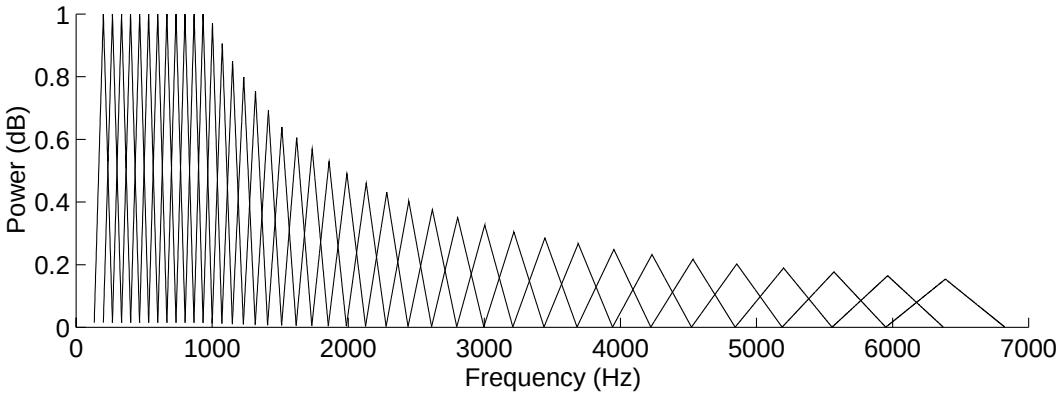

Figure 2-1: MFSC filter bank

The MFSC filter bank contains 40 filters that roughly approximate the frequency response of the basilar membrane in the cochlea of the inner ear. The filters span 156–6844 Hz and are spaced on a Mel-frequency scale, which is respectively linear and logarithmic below and above 1 kHz. The filters are triangular and multiplicatively scaled by area.

The MFSC representation consists of the 40 coefficients that correspond to the logarithm of the signal energy in the 40 MFSC filters. This log spectral representation is useful in analysis and the development of compensation techniques.

Cepstral Representation

Given a frame of MFSCs, a cosine transformation is applied to produce the MFCC signal representation. Equation 2.2 shows the cosine transformation.

$$Y[i] = \sum_{j=1}^{N} X[j]\cos[i(j - \frac{1}{2})\frac{\pi}{N}] \tag{2.2}$$

where

$$X[j] : \text{MFSC coefficient } j$$

$$Y[i] : \text{MFCC coefficient } i$$

$$N : \text{number of MFSC coefficients}$$

The MFCC representation consists of the 14 lower-order coefficients that correspond to the cosine transformation of N MFSC coefficients. The MFCC indices, i, range from 0 to 13. The MFSC indices, j, range from 1 to N. The number of MFSCs used in the cosine transformation, N, is either 40 or 30. When N is 40, all 40 MFSCs are used, and the cepstral representation spans 156–6844 Hz. This case corresponds to the original 16 kHz sampling rate. When N is 30, only the 30 lower-order MFSCs are used, and the cepstral representation is effectively bandlimited to span only 156–3469 Hz, the lower half of the original bandwidth. This case corresponds approximately to downsampling by a factor of 2 to an 8 kHz sampling rate.

The MFCC representation is widely used for benchmarking and comparison. It also uses fewer coefficients and is quite efficient, and these coefficients are less correlated through cosine transformation and more effectively modeled by independent densities [24].

2.3.2 Segmentation

In classification, the TIMIT time-aligned phonetic transcription is used to provide segment boundaries. In recognition, a segmentation algorithm [13] is used to provide segment hypotheses and scores. The algorithm associates each frame with one of its neighbors and marks a boundary when the direction of association changes from past to future. Using the resulting regions, the algorithm iterates until the entire utterance is described by one multi-level representation, called a dendrogram.

2.3.3 Feature Extraction

The classification and recognition systems extract 36 features for each segment. Of the 36 features, 1 is the duration of the segment, and the remaining 35 are averages over

varying intervals inside and outside the segment, respectively representing intra- and inter-segmental information. The features were determined by applying an automatic feature selection algorithm [31]. Acoustic-phonetic knowledge was used to propose property detectors with free parameters. For example, one property detector was time averaging over an interval, and the corresponding parameters were the initial and final times of the interval. A measure of phonetic discrimination was maximized over the space spanned by the parameters in order to determine the maximally discriminative features, called Generalized Measurements (GMs).

2.3.4 Acoustic Modeling

The classification and recognition systems use a maximum of 16 mixtures of diagonal Gaussians to model the distribution of acoustic features for each of the 61 TIMIT phones. The acoustic models are applied to produce phone hypotheses and scores. In classification, a unigram language model is also applied to produce scores that incorporate the probability of each phone occurrence in English.

2.3.5 Search

A Viterbi search determines the best path through the network of segment and phone hypotheses and scores. In recognition, a bigram language model is also applied to produce scores that reflect the probability of each phone pair occurring in English.

2.4 Evaluation

The evaluation of classification and recognition experiments is based on phonetic error rate. The comparison of classification experiments is also based on statistical significance.

2.4.1 Error Rate

The overall performance metric is the error rate among 56 phonetic classes. The 6 closures, /b/, /d/, /g/, /p/, /t/ and /k/, are grouped into one class, /cl/. Each of the other 55 TIMIT phonemes constitutes its own class. In classification, the error rate consists of only substitutions of one class for another. In recognition, the error rate includes substitutions, deletions and insertions, and varies depending on the evaluation technique used. The NIST alignment and scoring algorithm [30] minimizes the total error rate, as measured by the sum of the substitution, deletion and insertion error rates. In this thesis, the NIST algorithm is used to evaluate the recognition experiments.

2.4.2 Statistical Significance

Statistical significance is also used as a comparative metric. The availability of different microphone recordings of identical TIMIT data allows the application of the McNemar significance test [11]. In measuring the significance between two experiments, the McNemar test considers only those tokens which are correct in one experiment and incorrect in the other, since tokens which are correct or incorrect in both experiments do not contribute to information about relative performance. Equation 2.3 shows the McNemar significance between two experiments.

$$S = \begin{cases} 0.5^{k-1} \sum_{m=i}^{k} \binom{k}{m} & \text{if } i > k/2 \\ 0.5^{k-1} \sum_{m=0}^{j} \binom{k}{m} & \text{if } j < k/2 \\ 1.0 & \text{if } i = k/2 \end{cases} \tag{2.3}$$

where

i : number of tokens correct in experiment one and incorrect in experiment two

j : number of tokens correct in experiment two and incorrect in experiment one

$k = i + j$

S : significance

When S is lower than the significance level, the two experiments are significantly different. When S is higher than the significance level, there is not enough evidence to conclude on a difference. In this thesis, the McNemar test is used to compare all classification experiments with a significance level of 0.01. Unless stated otherwise, comparative results in classification are statistically significant.

Chapter 3

Data Analysis

TIMIT provides three microphone corpora that are useful for experiments on microphone variations. The recordings of identical utterances spoken by identical speakers allow comparison of differences incurred by variations in microphone only. A preliminary analysis of the microphones and data serves as the basis for understanding these differences and the subsequent experimental results. The TIMIT microphones are described, the effects of the microphones on the recorded signal are modeled, and the signal to noise, spectrographic, spectral and cepstral characteristics of the microphone data are examined.

3.1 Microphones

The TIMIT microphones have different properties that result in different recordings of the same input. The three sets of microphone data are referred to as the Sennheiser, B&K and Telephone.

3.1.1 Sennheiser

The Sennheiser [9, 35] is a pressure-gradient microphone with a flat frequency response, plus or minus 2 dB, that extends well beyond the frequency range used by the speech recognition system, approximately 100–7000 Hz. Pressure-gradient micro-

phones [4] record the pressure difference between two closely spaced transducers and are highly dependent on recording distance and direction relative to the speaker's mouth. The Sennheiser is a head-mounted, close-talking microphone that maintains a constant distance and direction near and in line with the mouth. It also has noise-canceling characteristics that reduce its sensitivity to non-oral sounds from other distances and directions, such as nasal and glottal resonances and environmental noise. The Sennheiser is the highest quality of the TIMIT microphones and is the standard recording microphone for many corpora and systems.

3.1.2 B&K

The B&K [9] is a pressure microphone also with a flat response extending beyond the frequency range of interest. Pressure microphones [4] record pressure directly. Although the B&K is an omnidirectional microphone that is not as dependent on direction, it is also a boom-mounted, far-field microphone that records from a more variable distance farther from the speaker's mouth. As a result, the B&K is more sensitive to non-oral sounds and of lower quality than the Sennheiser.

3.1.3 Telephone

The Telephone [18] is a combination of a Sennheiser microphone, a telephone microphone and a telephone channel. The Sennheiser recording was played back through an artificial mouth that simulated the acoustic characteristics between a speaker's mouth and a telephone handset. The playback was recorded using a telephone handset-mounted pressure microphone and transmitted over local and long distance telephone lines. In addition to microphone effects, the Telephone also suffers from transmission effects, which include bandlimiting to 300–3400 Hz and other distortions. As a result, the Telephone is the lowest quality of the TIMIT microphones.

3.2 Effects

The effects of the microphones on the recorded signal can be mathematically modeled by convolution and addition in the time domain. These models are used in the development of preprocessing techniques, and we maintain their dichotomy for purposes of discussion, recognizing that the distinction is rather arbitrary. For example, a convolutional effect, viewed in the appropriate domain, is additive, so a spectral subtraction compensates for either additive or convolution effects, depending on whether a logarithm has been taken or not.

3.2.1 Convolutional Effects

Some effects are modeled as convolutional distortions caused by Linear Time Invariant (LTI) filters. For example, variations in the speaker vocal tract, environmental acoustics and microphone transfer function have convolutional effects. In addition, bandlimiting and other linear distortions in the Telephone are also modeled by LTI filters. Equation 3.1 shows the convolutional effect in the time domain.

$$y[m] = x[m] * h[m] \tag{3.1}$$

where

$x[m]$: original signal

$y[m]$: distorted signal

$h[m]$: impulse response of the distortion

This effect is multiplicative in the spectral domain,

$$Y(e^{jw}) = X(e^{jw})H(e^{jw}), \tag{3.2}$$

additive in the log spectral domain,

$$\hat{Y}(e^{jw}) = \hat{X}(e^{jw}) + \hat{H}(e^{jw}), \tag{3.3}$$

and additive in the cepstral domain,

$$\hat{y}[m] = \hat{x}[m] + \hat{h}[m]. \tag{3.4}$$

Convolutional effects are assumed to vary slowly with respect to speech in that the characteristics of the speaker, environment and microphone remain relatively constant over the duration of each utterance. Using this assumption, additive log spectral or cepstral compensation vectors can be estimated.

3.2.2 Additive Effects

Other effects are modeled as additive distortions that are uncorrelated with speech. For example, variations in environmental noise level and interference in telephone transmission have additive effects. Equation 3.5 shows the additive effect in the time domain.

$$y[m] = x[m] + n[m] \tag{3.5}$$

where

$x[m]$: original signal

$y[m]$: noisy signal

$n[m]$: noise

This effect is also additive in the spectral domain,

$$Y(e^{jw}) = X(e^{jw}) + N(e^{jw}). \tag{3.6}$$

Additive effects are also assumed to vary slowly with respect to speech and are modeled as stationary white Gaussian random processes. Using these assumptions, additive spectral compensation vectors can be estimated.

3.3 Characteristics

The signal to noise, spectrographic, spectral and cepstral characteristics of the microphone data are examined. Given identical speakers and utterances, these characteristics reflect differences between the microphones.

3.3.1 Signal to Noise Characteristics

Signal to Noise Ratio (SNR) measures the ratio of signal power to noise power. SNR is a commonly used measurement to compare the quality of different microphone recordings. In general, the higher the SNR, the better the quality.

The average SNR for each microphone recording is computed as the mean signal power averaged over the training set utterances divided by the mean noise power averaged over the training set utterances. For each utterance, the mean signal power is the power averaged over all the frames, except those within the beginning and ending silence segments labeled as /h#/, which are averaged to compute the mean noise power. In this thesis, the power in each frame is taken to be the first MFCC coefficient, MFCC[0], which is the sum of the MFSC coefficients, that correspond to the power in dB in the MFSC filters. The original 16 kHz SNR is computed over all 40 MFSCs. The downsampled 8 kHz SNR is computed over only the lower 30 MFSCs, Table 3.1 shows the average SNRs in dB for each microphone before and after downsampling.

	Sennheiser	B&K	Telephone
16 kHz SNR	23.8	21.2	11.4
8 kHz SNR	26.1	23.3	15.4

Table 3.1: Average SNRs in dB for each microphone before and after downsampling

The high quality noise-canceling Sennheiser has the highest SNR. The lower quality non-noise-canceling B&K and Telephone have lower SNRs, suggesting higher error rates in classification and recognition. The particularly low Telephone SNR is presumably due to transmission effects. For example, transmission bandlimiting causes

a lack of high frequency energy that correponds to low SNRs in the upper MFSC filters. Downsampling increases the Telephone SNR by about 2 dB relative to the Sennheiser and B&K, but even after downsampling, the Telephone still has the lowest SNR by many dB.

3.3.2 Spectrographic Characteristics

Spectrograms show temporal and spectral characteristics of the STFT. Figure 3-1 shows spectrograms of the word "discipline" extracted from the same utterance spoken by the same speaker for each microphone[1]. The x axis shows time, the y axis shows frequency and the gray scale shows spectral magnitude. Plots of zero crossing rate, total energy, and low frequency energy are also included. The word "discipline" is transcribed as [d d Is{ p p l|n (@)].

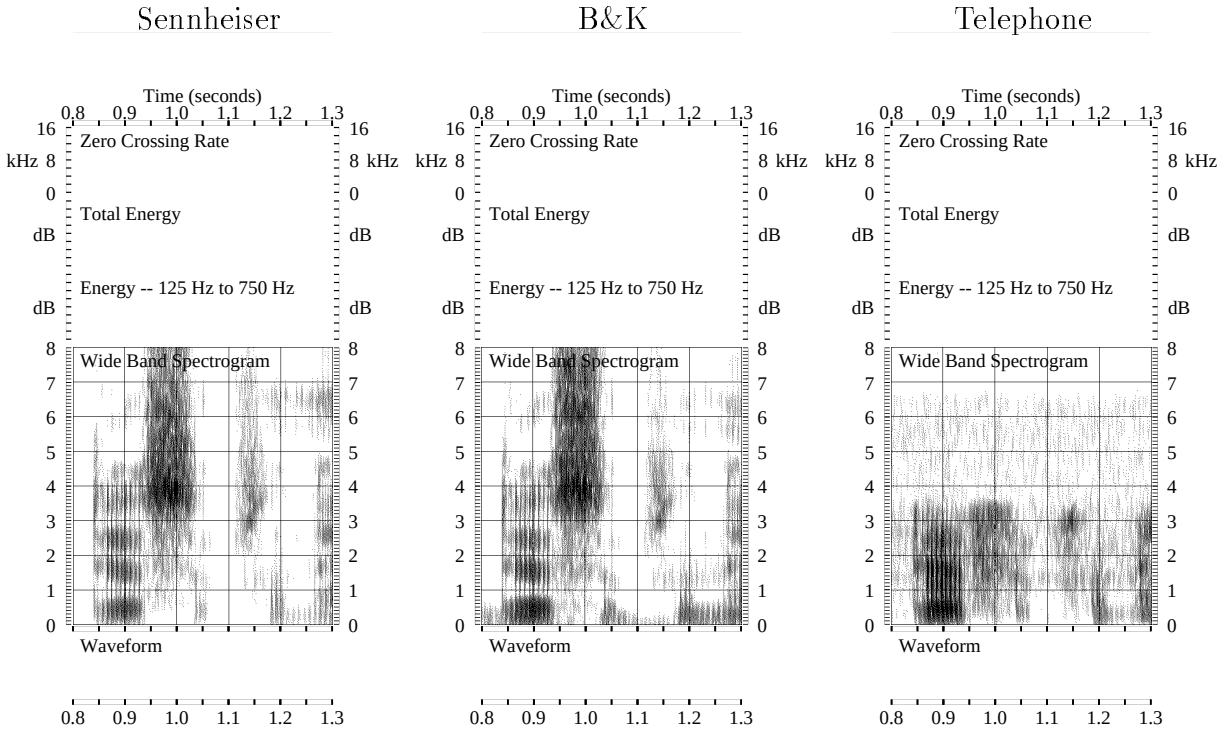

Figure 3-1: Spectrograms of the word "discipline" for each microphone

In comparison to the Sennheiser and Telephone, the B&K shows an increase in

[1]The utterance, "si" 2119, is "the instinct to discipline has been lost". The speaker, rjm, is male.

low frequency energy. The non-gradient B&K does not have the highpass characteristic that results from taking the difference between two transducers in the gradient Sennheiser. In addition, the boom-mounted far-field B&K records from a farther distance relative to the speaker's mouth and is more sensitive to non-oral sounds than the close-talking noise-canceling Sennheiser. Some of these non-oral sounds, such as nasal and glottal resonances and environmental noise, can occur at low frequencies. For example, the B&K shows more low frequency energy in the nasal, /n/. The B&K also shows more low frequency energy in the closures, /d / and /p /, stops, /d/ and /p/, and fricative, /s/. This increase in low frequency energy can obscure the voicing feature and suggests difficulty in discriminating phonemes along this dimension.

In comparison to the Sennheiser and B&K, the Telephone shows a lack of high frequency energy due to transmission bandlimiting and an increase in background energy due to signal normalization and other transmission effects. For example, the Telephone shows no high frequency energy for the stops, /d/ and /p/, and the fricative, /s/. This lack of high frequency energy can obscure stop and fricative features and suggests difficulty in discriminating phonemes along the manner dimension.

3.3.3 Spectral Characteristics

Figure 3-2 shows mean MFSCs averaged over the training set for each microphone. For notational consistency, solid, dashed and dash-dotted lines respectively represent the Sennheiser, B&K and Telephone.

As seen in the spectrograms, the B&K shows an increase in low frequency energy, presumably due to effects recorded by the non-gradient, boom-mounted far-field B&K but not by the gradient, close-talking, noise-canceling Sennheiser. The Telephone shows a lack of high frequency energy and an increase in background energy due to transmission and normalization effects.

These differences between microphones can be quantified by applying a distance measure to their mean and variance vectors. Given vectors for microphone x and y, the normalized distance from microphone x to y is computed as the square root of the sum of the squared differences between the mean coefficients normalized by the

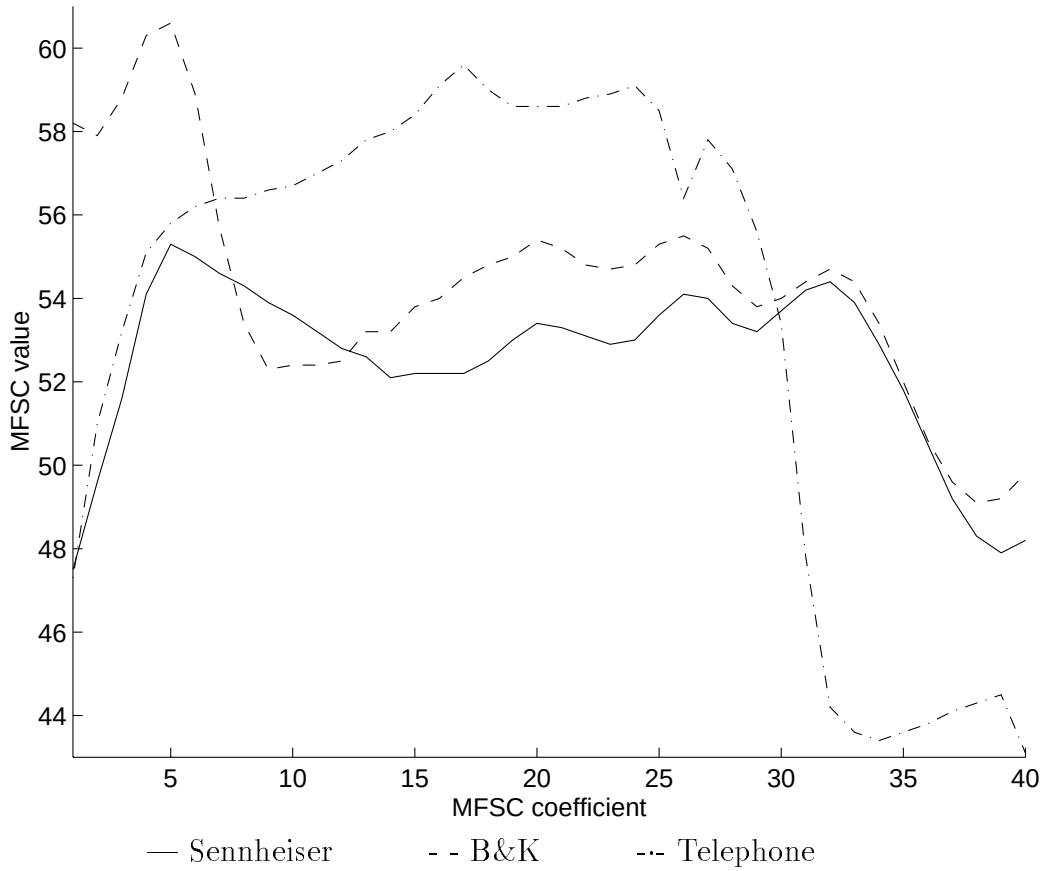

Figure 3-2: Mean MFSCs over the training set for each microphone

variance coefficients of microphone y, i.e.

$$D = \sqrt{\sum_{i=1}^{N} \frac{(m_x[i] - m_y[i])^2}{\sigma_y^2[i]}} \tag{3.7}$$

where

$m_x[i]$: mean vector for microphone x

$m_y[i]$: mean vector for microphone y

$\sigma_y^2[i]$: variance vector for microphone y

N : number of coefficients

As measured by normalized distance over all 40 MFSCs, the distance between Telephone and Sennheiser, 3.0, is more than twice as large as the distance between the B&K and Sennheiser, 1.4, suggesting higher error rates under mismatched condi-

tions involving the Telephone. Downsampling, as measured by normalized distance over only the 30 lower MFSCs, decreases the distance between the Telephone and Sennheiser to 2.0, suggesting that downsampling can reduce Telephone errors.

Deviations between microphones have different effects on different phonemes depending on their spectral characteristics. Phonemes that share similar spectral characteristics are generally produced by the same manner of articulation and can be grouped into broad manner classes. Under mismatched conditions, when deviations are small, phonemes may be classified in the correct broad class but may be misclassified within the class. For example, the small deviations at low frequencies in the B&K can obscure the voicing feature that discriminates phonemes within the obstruent classes. When deviations are large, phonemes are more likely to be misclassified in the incorrect broad class. For example, the large deviations at high frequencies in the Telephone can obscure obstruent features that discriminate fricatives and stops. In order to examine these effects, the TIMIT phonemes are grouped into six broad classes generally based on manner of articulation. Table 3.2 shows these six broad classes with example phonemes.

Broad class	Example phonemes
Vowel	a @ ^ O a ˚ { a ˜ E H 5 e | i o O ˚ U u
Semivowel	l r w y
Nasal	m n 4
Strong obstruent	C J s S z Z
Weak obstruent	b d D f g k p t T v
Silence	b d • g k ˆ p t h#

Table 3.2: Broad classes with example phonemes

Strident fricatives and affricate releases form the strong obstruent class. Weak fricatives and stop releases form the weak obstruent class. Closures are grouped with the silence class. In addition, allophones, such as syllabic and flapped realizations, are grouped with their corresponding classes.

Figure 3-3 shows mean broad class MFSCs averaged over the training set for each microphone.

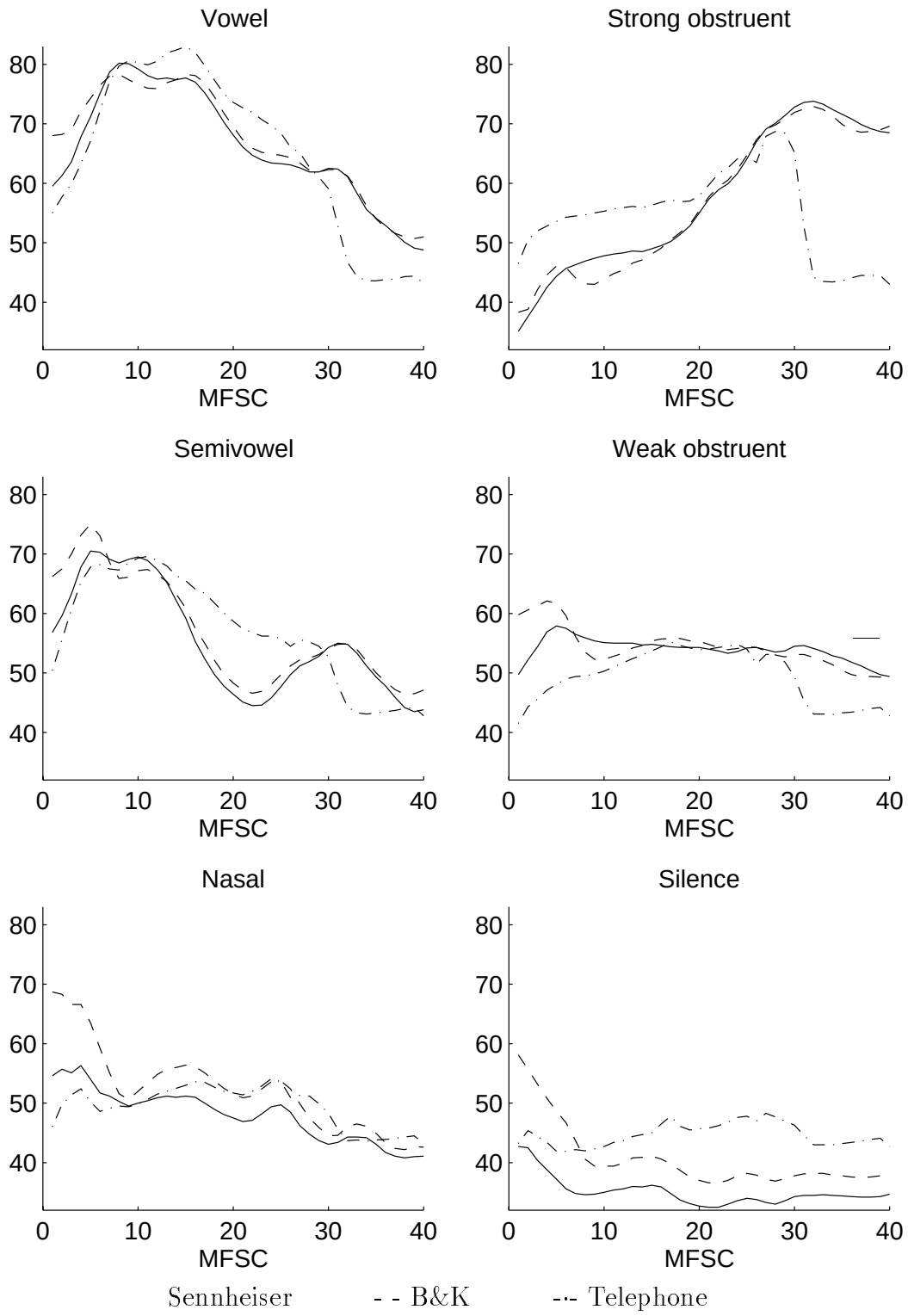

Figure 3-3: Mean broad class MFSCs over the training set for each microphone

As measured by normalized distance over 40 MFSCs, the B&K and Sennheiser differ mostly in the nasal and silence classes. In these classes, the B&K shows a low frequency peak that is presumably due to nasal resonances, pre-voicing and environmental noise. The B&K also differs in the strong obstruent class, which shows a low frequency variation presumably due to voicing and environmental noise. These deviations suggest confusions between the nasal and silence classes and within the obstruent class over voicing.

In comparison to the B&K and Sennheiser, the Telephone and Sennheiser show larger differences that suggest more between-class errors. As measured by normalized distance over the 30 lower MFSCs, the Telephone and Sennheiser mostly differ in the weak obstruent and silence classes presumably due to transmission effects and signal normalization. The Telephone also differs in the semivowel class, which shows a mid-frequency variation that can obscure formant structures. These deviations suggest confusions between the weak obstruent and silence classes and within the semivowel class.

3.3.4 Cepstral Characteristics

Figure 3-4 shows mean MFCCs averaged over the training set for each microphone. The first coefficient, MFCC[0], has a relatively large value that reflects total energy. The higher order coefficients, MFCC[1–13], have relatively smaller values that reflect spectral distribution. MFCC[0] is clipped in order to focus on the higher order MFCCs.

As with the MFSCs, the normalized distance between Telephone and Sennheiser, 2.7, is almost twice as large as the distance between the B&K and Sennheiser, 1.4, suggesting higher error rates under mismatched conditions involving the Telephone. An examination of the MFCC signal representation shows that the large cepstral deviations in the Telephone are mostly due to differences in spectral distribution. For example, the value of MFCC[2] is much more negative for the Telephone than for the Sennheiser and B&K. Figure 3-5 shows the cosine weighting function for MFCC[2]. In computing MFCC[2], the energy at middle frequencies, MFSC[10–30], is subtracted

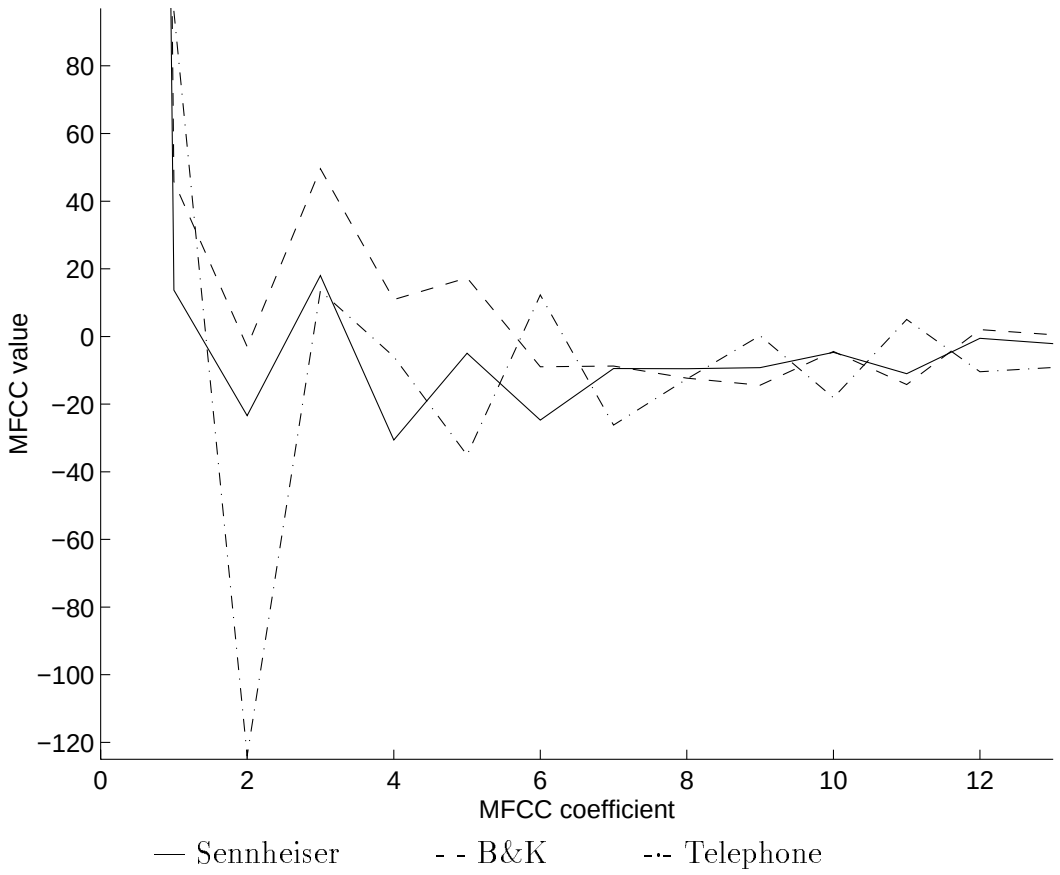

Figure 3-4: Mean MFCCs over the training set for each microphone

from the energy at low and high frequencies, MFSC[1–10] and MFSC[30–40]. Lack of energy at high frequencies, above MFSC[30], results in large negative values for MFCC[2].

Downsampling corresponds approximately to applying the cosine weighting function over only the lower 30 MFSCs. Figure 3-6 shows mean MFCCs averaged over the training set for each microphone after downsampling.

Downsampling decreases the normalized distance between the Telephone and Sennheiser to 1.3, suggesting that downsampling can reduce error rates under mismatched conditions when training on the Sennheiser and testing on the Telephone.

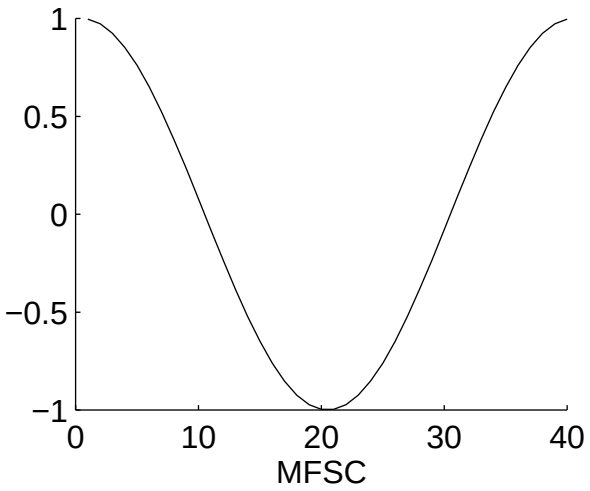

Figure 3-5: Cosine weighting function for MFCC[2]

3.4 Summary

The TIMIT microphones and data are analyzed in order to understand, speculate and explain experimental results. The differences between the B&K and Sennheiser are relatively small, with an increase in energy at low frequencies. In comparison, the differences between the Telephone and Sennheiser are larger, with a lack of energy at high frequencies and deviations within the Telephone bandwidth. In the following chapters, these differences are related to errors under mismatched microphone training and testing conditions.

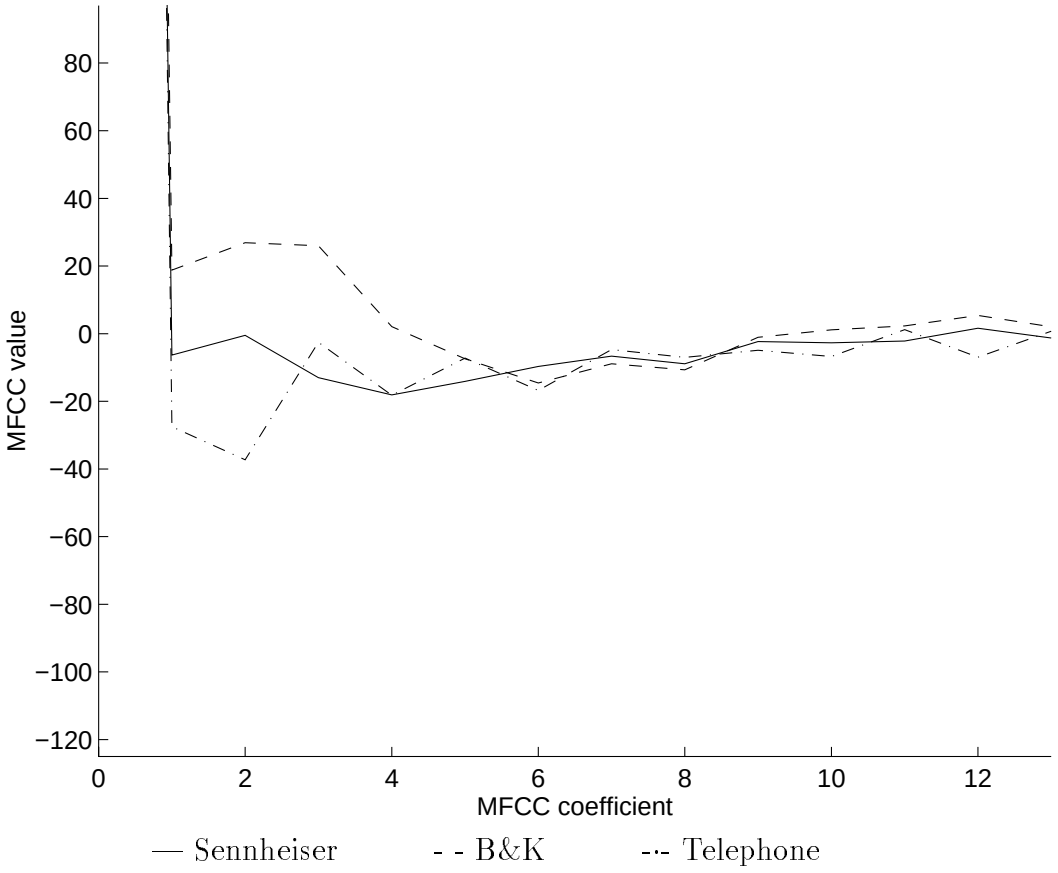

Figure 3-6: Mean MFCCs over the training set for each microphone after downsampling

Chapter 4

Baseline Experiments

Baseline experiments are performed in phonetic classification and recognition for different microphone conditions. These experiments are analyzed to understand the effects of microphone variations and used in comparison to experiments with compensation techniques. Analysis focuses on the realistic conditions when the system is trained on the high quality Sennheiser and tested on the lower quality B&K and Telephone.

4.1 Notation

The microphone is abbreviated by its initial. S, B and T respectively correspond to the Sennheiser, B&K and Telephone. Other notations are introduced for denoting microphone conditions and presenting experimental results.

4.1.1 Conditions

The microphone training and testing condition is denoted by a parenthesized ordered pair. (X, Y) corresponds to the condition when microphone X is used in training and microphone Y is used in testing. For clarity, the sampling rate may be specified by a third argument in the parenthesized notation. (X, Y, Z) is the condition when training on X and testing on Y at a Z kHz sampling rate. The 3 sets of TIMIT

microphone data provide 9 training and testing conditions at 16 kHz. Downsampling provides 9 downsampled training and testing conditions at 8 kHz.

4.1.2 Results

Error rates in percent are presented in a matrix. Table 4.1 shows an example error rate matrix.

	S	B	T
S	(S, S)	(S, B)	(S, T)
B	(B, S)	(B, B)	(B, T)
T	(T, S)	(T, B)	(T, T)

Table 4.1: Example error rate matrix

Rows show training microphones. Columns show testing microphones. Diagonal entries correspond to matched conditions when training and testing on the same microphone. Off-diagonal entries correspond to mismatched conditions when testing on a microphone different from the one used in training.

4.2 General Results

Baseline experiments are performed for all microphone conditions before and after downsampling. General classification and recognition results are presented and discussed.

4.2.1 Classification

Table 4.2 shows baseline classification error rates in percent.

The diagonal entry shows that the matched testing condition achieves the lowest error rate for each training microphone. Increases in error rate down the diagonal reflect decreases in the quality of the microphone. The lowest error rate is achieved by (S, S, 16), increasing 6% to (B, B, 16) and 33% to (T, T, 16). These results are

	S	B	T
S	31.2	38.9	66.7
B	35.6	33.1	67.5
T	81.1	76.9	41.6

Table 4.2: Baseline classification error rates in percent

consistent with those in the literature. For example, a previous study [5] also shows a 33% increase in error rate, from 25.2% under (S, S, 16) to 33.5% under (T, T, 16), when evaluating on 39 rather than 56 classes.

Error rates off the diagonal reflect differences between the training and testing microphones. Error rates increase moderately under mismatched conditions involving the Sennheiser and B&K. For example, the (S, S, 16) error rate increases by 25% to (S, B, 16). However, error rates increase severely under all mismatched conditions involving the Telephone. For example, the (S, S, 16) error rate increases by 114% to (S, T, 16). These results are consistent with the analyses in the previous chapter. In comparison to the Sennheiser and B&K, the Telephone shows large differences in signal-to-noise, spectral and cepstral characteristics. As downsampling reduces all of these differences, downsamping also reduces the severe performance degradations under mismatched conditions involving the Telephone. Table 4.3 shows baseline classification error rates in percent after downsampling.

	S	B	T
S	33.5	42.0	55.7
B	37.4	34.0	57.6
T	49.9	55.2	41.8

Table 4.3: Baseline classification error rates in percent after downsampling

Downsampling causes the speech recognition system to focus on acoustic information below 4 kHz, effectively bandlimiting the Sennheiser and B&K to match the transmission bandwidth of the Telephone. Increases in error rate for the Sennheiser and B&K reflect the loss of high frequency information. For example, the (S, S, 16)

error rate increases by 7% to (S, S, 8). However, downsampling does not cause a loss of high frequency information for the Telephone. For the matched condition, downsampling results in a statistically insignificant difference in error rate between (T, T, 16) and (T, T, 8). For mismatched conditions involving the Telephone, downsampling results in reductions in error rate. For example, the error rate under (S, T, 16) decreases by 16% to (S, T, 8). Although downsampling can reduce error rates under mismatched conditions, even after downsampling, mismatched conditions involving the Telephone still suffer large performance degradations. For example, the (S, S, 8) error rate increases by 66% to (S, T, 8). These degradations are presumably due to transmission effects other than bandlimiting in the Telephone.

4.2.2 Recognition

Table 4.4 shows baseline recognition error rates in percent. For each training and testing condition, the total error rate is shown first followed respectively by the substitution, deletion and insertion rates in parentheses.

	S				B				T			
S	51.2	(26.7	16.1	8.4)	59.0	(30.7	20.7	7.7)	79.1	(32.1	45.7	1.3)
B	54.9	(30.9	16.0	8.0)	52.9	(28.3	15.4	9.2)	79.3	(31.6	46.4	1.2)
T	82.3	(54.8	18.6	8.9)	82.6	(52.3	18.7	11.6)	59.8	(31.6	21.8	6.3)

Table 4.4: Baseline recognition error rates in percent

Increases in error rate reflect the complexity of the recognition task. Under (S, S, 16) and (S, B, 16), error rates increase by approximately 60% from classification to recognition, The substitution, deletion and insertion rates are respectively about 80%, 50% and 20% of the classification error rates.

In general, the recognition results follow the trends in classification. The (S, S, 16) error rate increases by 3% to (B, B, 16) and by 17% to (T, T, 16). These results are comparable to those of a previous study [26] which shows a 24% increase in error rate, from 47.3% under (S, S, 16) to 58.7% under (T, T, 16), when evaluating on 39 rather than 56 classes. Under mismatched conditions involving the Sennheiser and

B&K, error rates increase moderately, with the (S, S, 16) error rate increasing by 15% to (S, B, 16). Under mismatched conditions involving the Telephone, error rates increase severely, with the (S, S, 16) error rate increasing by 54% to (S, T, 16). The previous study [26] shows a 45% increase in error rate, from 47.3% under (S, S, 16) to 68.7% under (S, T, 16).

Table 4.5 shows baseline recognition error rates in percent after downsampling. Downsampling reduces the severe performance degradations under mismatched conditions involving the Telephone. For example, the error rate under (S, T, 16) shows an 8% decrease to (S, T, 8). Yet, even after downsampling, mismatched conditions involving the Telephone still suffer large performance degradations that are presumably due to transmission effects other than bandlimiting. For example, the (S, S, 8) error rate increases by 43% to (S, T, 8).

	S				B				T			
S	55.7	(31.5	12.6	11.6)	62.7	(35.5	15.4	11.7)	73.1	(31.4	38.2	3.4)
B	59.8	(31.0	22.6	6.2)	56.6	(30.4	17.0	9.2)	75.5	(31.6	41.8	2.1)
T	68.5	(41.9	14.2	12.4)	73.7	(45.5	13.3	14.9)	63.9	(34.8	19.4	9.7)

Table 4.5: Baseline recognition error rates in percent after downsampling

4.2.3 Discussion

The baseline phonetic classification and recognition systems perform comparably with those in the literature [5, 26]. Classification experiments show the effects of microphone variations on the signal representation, feature extraction and acoustic modeling components in the speech recognition system. These experiments are useful for analysis. Recognition experiments show the effects of microphone variations on the entire system, combining the classification with the segmentation and search components. These experiments are more complex and result in higher error rates that are more difficult to analyze.

In order to establish a complete set of benchmarks, experiments cover all microphone training and testing conditions before and after downsampling. The Sennheiser

is shown to be of higher quality than the B&K, which in turn is of higher quality than the Telephone. Mismatched conditions involving the Sennheiser and B&K suffer moderate performance degradations, while mismatched conditions involving the Telephone suffer severe degradations. Downsampling reduces these degradations by effectively bandlimiting the Sennheiser and B&K to match the transmission bandwidth of the Telephone, but even after downsampling, mismatched conditions involving the Telephone still suffer large degradations.

Analysis focuses on the relative differences between (S, B, 16) and (S, T, 8) in comparison to (S, S, 16). These mismatched conditions are considered to be realistic in deployment. In training, the system is assumed to have used the high quality Sennheiser to produce two sets of models, one at each sampling rate. In testing, the user is assumed to use the lower quality B&K or Telephone rather than the Sennheiser, and the system is assumed to automatically select the correct sampling rate and model, depending on the spectral characteristics of the input. Although the system can achieve low error rates under matched conditions, the system lacks microphone robustness and cannot maintain low error rates under mismatched conditions.

4.3 (S, B, 16)

In the previous chapter, the B&K is shown to deviate from the Sennheiser, especially by an increase in low frequency energy. These deviations are related to the differences in performance when testing on the B&K.

4.3.1 Classification

When testing on the B&K rather than the Sennheiser, the classification error rate increases by 25%, from 31.2% under (S, S, 16) to 38.9% under (S, B, 16). Table 4.6 shows the breakdown in percent of the phone tokens between (S, S, 16) and (S, B, 16).

As the B&K shows small deviations from the Sennheiser, 82% of the tokens are either correctly or incorrectly classified on both the B&K and Sennheiser. These

	(S, B, 16) correct	(S, B, 16) incorrect
(S, S, 16) correct	56	13
(S, S, 16) incorrect	5	26

Table 4.6: Breakdown in percent of tokens between (S, S, 16) and (S, B, 16)

tokens do not contribute to information about relative performance and are not examined. Instead, the remaining 18% of the tokens that are correctly classified under one condition and incorrectly classified under the other are extracted for examination.

Analysis focuses on the 13% of the tokens that are incorrectly classified under (S, B, 16) but correctly classified under (S, S, 16). These tokens are the additional classification errors due to testing on the B&K rather than the Sennheiser. In order to examine general errors, these additional misclassifications are grouped into the broad phonetic classes described in Table 3.2. Table 4.7 shows the frequency in percent of the most frequent misclassifications between various broad classes that occur due to testing on the B&K.

Class	Class	Frequency
Vowel	Vowel	29.9
Weak obstruent	Weak obstruent	15.6
Strong obstruent	Strong obstruent	9.6
Vowel	Semivowel	7.6
Nasal	Silence	4.5
Weak obstruent	Silence	4.3
Nasal	Weak obstruent	4.0

Table 4.7: Frequency in percent of the most frequent broad class misclassifications due to the B&K

Overall, approximately 60% of the misclassifications occur within broad classes, while the remaining 40% occur between broad classes. The table shows the most frequent within and between class errors, totaling to approximately 75% of the additional misclassifications between (S, B, 16) and (S, S, 16).

The within-class errors almost all occur within one of the vowel or obstruent classes. These results are consistent with the observations in the previous chapter that

the B&K vowels and obstruents show small deviations from those of the Sennheiser, leading to errors within broad classes. For example, vowels can be confused with each other due to variations in formant frequencies. Obstruents can be confused due to variations in the voicing feature.

The between-class errors mostly occur between the nasal, silence and weak obstruent classes or the vowel and semivowel classes. These results are also consistent with the observations in the previous chapter that the B&K nasals and silences show larger deviations from those of the Sennheiser, leading to errors between broad classes. For example, nasals, silences and weak obstruents are similar in that they have low levels of energy and can be confused due to variations in nasal resonances, voicing, and environmental noise. Vowels and semivowels are similar in formant structure and can be confused due to variations in spectral amplitudes.

In order to examine specific errors, the additional misclassifications are sorted by phoneme. Table 4.8 shows the frequency in percent of the most frequent misclassifications of various phonemes with their most frequent substitutions that occur due to testing on the B&K.

Phoneme	Frequency	Substitution
cl	14.1	n
s	11.2	z
t	6.2	d
p	4.9	b
v	4.9	m
E	4.4	I
f	4.1	v

Table 4.8: Frequency in percent of the most frequent misclassifications with their most frequent substitutions due to the B&K

The table shows the seven phonemes that are most often misclassified, totaling to approximately 50% of the additional misclassifications between (S, B, 16) and (S, S, 16). Some phonemes are often misclassified as phonemes in other broad classes. For example, the silence class, /cl/, accounts for the largest percent of additional misclassifications and is most frequently misclassified as the nasal, /n/. The weak voiced

fricative, /v/, is often misclassified as the nasal, /m/. These errors are presumably due to confusions over nasal resonance, voicing and room noise.

Other phonemes are often misclassified within their broad class. For example, the unvoiced fricatives, /s/ and /f/, and stops, /t/ and /p/, are often misclassified as their voiced counterparts, /z/, /v/, /d/ and /b/. In fact, 67% of the within-obstruent errors are misclassifications between unvoiced and voiced phonemes, presumably due to confusions over voicing and noise energy. In addition, the lax vowel, /E/, is often misclassified as another lax vowel, /I/, presumably due to confusions over formant frequencies.

Four of the most frequent confusion pairs, the misclassifications of /cl/ as /n/, /s/ as /z/, /v/ as /m/ and /E/ as /I/, are extracted for further examination. Figure 4-1 shows mean MFSCs averaged over the training set for these frequent confusion pairs that occur due to testing on the B&K. The B&K and Sennheiser recordings of the misclassified phoneme are respectively denoted by solid and dotted lines. The Sennheiser recording of the substituted class is denoted by a dashed line.

The occurrence of a confusion means that the B&K recording of the misclassified phoneme differs from the Sennheiser recording of that phoneme and resembles the Sennheiser recording of the substituted phoneme. As measured by normalized distance, the B&K phoneme, denoted by the solid line, is often closer to the incorrect Sennheiser phoneme, denoted by the dotted line, than the correct Sennheiser phoneme, denoted by the dashed line. Although the confusions involve different broad classes with different spectral characteristics, they all show low frequency deviations. The between-class confusions of /cl/ as /n/ and /v/ as /m/ can be explained by the peak in low frequency energy in the B&K /cl/ and /v/. The within-class confusions of /s/ as /z/ and /E/ and /I/ can be explained by the variations at low frequencies in the B&K /s/ and /E/.

Having analyzed the additional errors due to the B&K, analysis is shifted to the 5% of the tokens that are incorrectly classified under (S, S, 16) and correctly classified under (S, B, 16). These are the additional errors due to testing on the Sennheiser rather than the B&K. Table 4.9 shows the frequency in percent of the most frequent

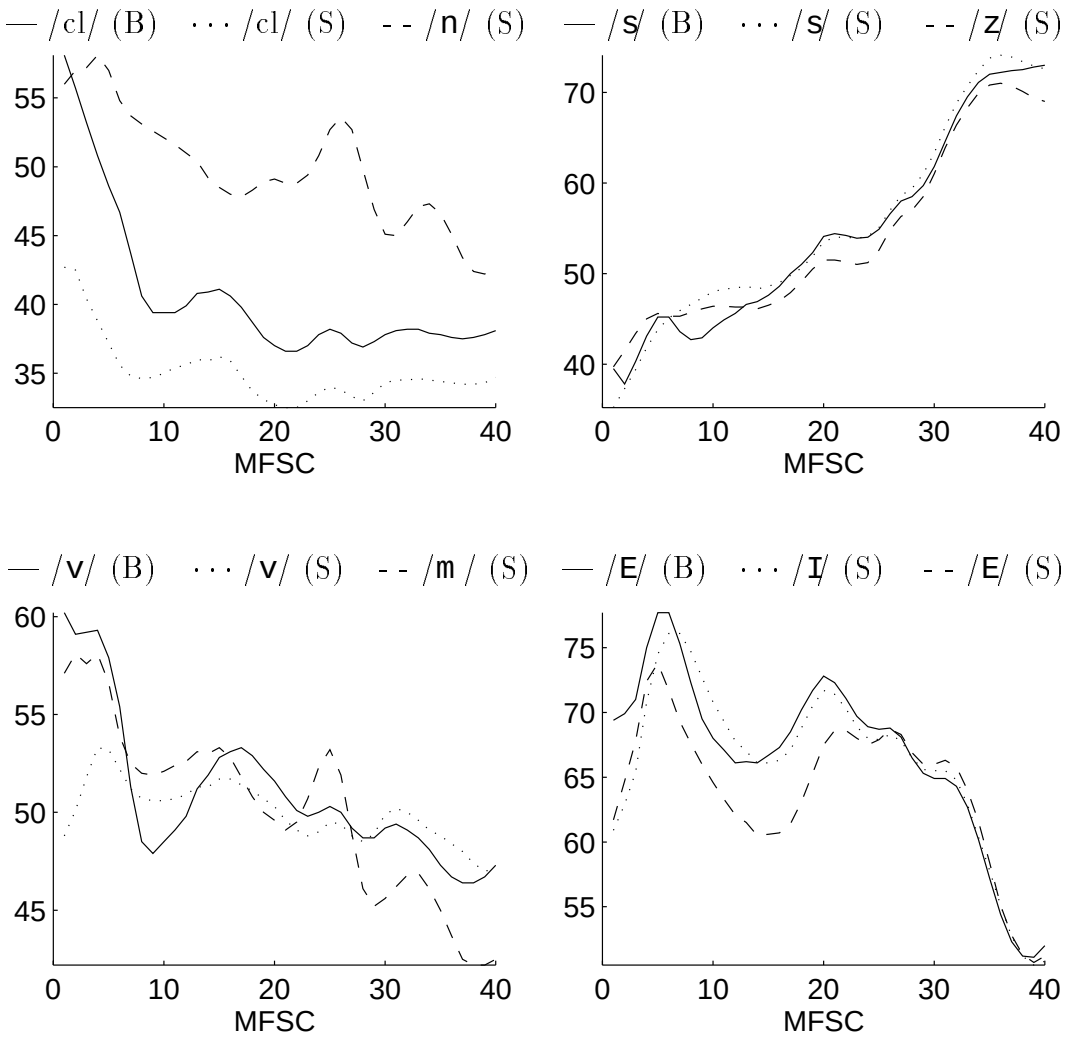

Figure 4-1: Mean MFSCs over the training set for frequent confusion pairs due to the B&K

misclassifications of various phonemes with their most frequent substitutions due to testing on the Sennheiser.

Phoneme	Frequency	Substitution
z	30.1	s
i	18.6	I
u	6.2	w

Table 4.9: Frequency in percent of the most frequent misclassifications with their most frequent substitutions due to the Sennheiser

In comparing (S, S, 16) to (S, B, 16), /z/ accounts for the largest percent of additional misclassifications and is most frequently misclassified as /s/. In contrast, misclassifications of /s/ as /z/ are common in comparing (S, B, 16) to (S, S, 16). This trend describes many of the misclassifications that occur under matched conditions but not under mismatched conditions. In general, under mismatched conditions, classes that are often substituted for other classes may also be correctly classified more often. On the other hand, these occurrences may result from incorrect transcription of the corpus. For example, English voiced fricatives, such as /z/, are often devoiced, especially at the end of the sentence. Some of these devoiced phonemes may have been mislabeled by human transcribers due to phonological compensation.

4.3.2 Recognition

When testing on the B&K rather than the Sennheiser, the recognition error rate increases by 15%, from 51.2% under (S, S, 16) to 59.0% under (S, B, 16). The substitution and deletion rates respectively increase by 15% and 29%, while the insertion rate decreases by 8%. These differences are separated for further examination.

Substitutions

47% of the additional errors are substitutions. Table 4.10 shows the frequency in percent of the most frequent substitutions of various phonemes with their most frequent misclassifications due to testing on the B&K.

Phoneme	Frequency	Misclassification
cl	20.2	n
s	10.1	z
k	7.7	t
t	7.2	d
p	4.6	b

Table 4.10: Frequency in percent of the most frequent substitutions with their most frequent misclassifications due to the B&K

The table shows the five phonemes that account for the largest increase in substitutions, totaling to approximately 50% of the additional substitutions between (S, B, 16) and (S, S, 16). /cl/ accounts for the largest percent of additional substitutions and is most frequently confused with /n/. The remaining substitutions are unvoiced obstruents, /s/, /t/ and /p/, which are often confused with their voiced counterparts, /z/, /d/ and /b/. These substitutions are presumably due to low frequency differences between the B&K and Sennheiser that obscure nasal resonance, voicing and environmental noise. Furthermore, these substitutions resemble the errors in classification and suggest that some of the previous analyses can be applied to recognition.

Deletions

The remaining 53% of the additional errors are deletions. Table 4.11 shows the frequency in percent of the most frequent deletions due to testing on the B&K.

The table shows the phonemes that account for the largest increase in deletions, totaling to approximately 50% of the additional deletions between (S, B, 16) and (S, S, 16). /cl/ accounts for the largest percent of additional deletions. Other weak events, such as the nasal, /n/, and the reduced vowels, /^/, /|/ and /}/ are also often deleted. These deletions are presumably due to effects such as environmental noise that obscure the presence of weak events. In addition, the obstruents, /s/, /p/ and /t/, are often deleted. These deletions also resemble the errors in classification and suggest that phonemes that have low classification scores may be deleted rather

Phoneme	Frequency
cl	20.3
n	4.9
^	4.7
s	4.4
}	4.2
|	4.2
p	4.2
t	3.9

Table 4.11: Frequency in percent of the most frequent deletions due to the B&K

than substituted.

Insertions

In comparing (S, B, 16) to (S, S, 16), the insertion rate decreases. Table 4.12 shows the frequency in percent of the most frequent insertions that do not occur when testing on the B&K.

Phoneme	Frequency
t	18.0
k	11.5
cl	9.8
?	7.4
•	5.7

Table 4.12: Frequency in percent of the most frequent insertions that do not occur on the B&K

The phonemes, such as /t/ and /cl/, that account for large percentages of the decrease in insertions, also account for large percentages of the increase in deletions. This trend describes many of the insertions that occur under matched conditions but not under mismatched conditions. In general, under mismatched conditions, classes that are often deleted may not be inserted as often.

4.4 (S, T, 8)

In comparison to the B&K, the Telephone deviates from the Sennheiser both at high frequencies and within the Telephone bandwidth. These larger deviations are related to the larger increases in error that occur when using the Telephone rather than the Sennheiser.

4.4.1 Classification

When testing on the Telephone, the classification error rate increases by 79%, from 31.2% under (S, S, 16) to 55.7% under (S, T, 8). Table 4.13 shows the breakdown in percent of phone tokens between (S, S, 16) and (S, T, 8).

	(S, T, 8) correct	(S, T, 8) incorrect
(S, S, 16) correct	38	31
(S, S, 16) incorrect	6	25

Table 4.13: Breakdown in percent of tokens between (S, S, 16) and (S, T, 8)

In comparison to the B&K, only 63% of the tokens are either correctly or incorrectly classified on both the Telephone and Sennheiser. The 6% of the tokens that are incorrectly classified under (S, S, 16) and correctly classified under (S, T, 8) are distributed over many classes, none of which show significant trends. As a result, analysis focuses on the 31% of the tokens, a 138% increase from the B&K, that are the additional classification errors due to testing on the Telephone. Table 4.14 shows the frequency in percent of the most frequent misclassifications between various broad classes that occur due to testing on the Telephone.

The table shows the most frequent broad class errors, totaling to approximately 75% of the additional misclassifications between (S, T, 8) and (S, S, 16). In comparison to the B&K, approximately 60% of the additional errors occur between, rather than within, broad classes, while the remaining 40% occur within broad classes. These results are consistent with the previous observations that the Telephone shows larger deviations from the Sennheiser that suggest more errors between broad classes.

Class	Class	Frequency
Vowel	Vowel	18.8
Weak obstruent	Weak obstruent	11.9
Nasal	Silence	11.0
Vowel	Semivowel	10.7
Weak obstruent	Silence	7.6
Strong obstruent	Weak obstruent	6.7
Nasal	Weak obstruent	5.1
Strong obstruent	Silence	4.1

Table 4.14: Frequency in percent of the most frequent broad class misclassifications due to the Telephone

The between-class errors mostly occur between the vowel and semivowel class or the nasal, silence and obstruent classes. Vowels and semivowels are often confused due to the deviations in formant frequencies. Nasals, silences and obstruents are often confused due to the lack of high frequency energy and the deviations within the Telephone bandwidth. The within-class errors mostly occur within the vowel or weak obstruent classes. These errors are also due to variations in spectral energy caused by transmission effects and signal normalization.

Table 4.15 shows the frequency in percent of the most frequent misclassifications of various phonemes with their most frequent substitutions that occur due to testing on the Telephone.

Phoneme	Frequency	Substitution
cl	23.8	n
l	4.7	r
i	4.7	|
b	3.5	d
w	3.1	r
p	3.1	d
s	3.1	f
z	3.1	v

Table 4.15: Frequency in percent of the most frequent misclassifications with their most frequent substitutions due to the Telephone

The table shows the eight phonemes that are most often misclassified, totaling to approximately 50% of the additional misclassifications between (S, T, 8) and (S, S, 16). The silence class, /cl/, accounts for almost 25% of additional misclassifications and is most frequently misclassified as the nasal, /n/. /cl/ is also often confused with other weak events, such as the weak and strong voiced fricatives, /v/ and /z/. These errors constitute many of the between-broad-class errors.

The stops, /b/ and /p/ and fricatives, /s/, /z/ and /f/ are often misclassified as other stops and fricatives, /d/, /f/, /v/ and /s/. In comparison to the B&K obstruents, which are often misclassified along the voicing dimension due to low frequency variations, the Telephone obstruents are often misclassified along the place-of-articulation dimension due to the lack of high frequency information. The tense vowel, /iy/, and semivowels, /w/ and /l/, are often misclassified as the reduced vowel and semivowels, /ix/ and /r/. In comparison to the B&K vowels and semivowels, which show reasonable misclassifications due to small variations at low frequencies, the Telephone vowels and semivowels are often coarsely misclassified due to larger deviations over the formant region.

Four of the most frequent confusion pairs, the misclassifications of /cl/ as /n/, /l/ as /r/, /iy/ as /ih/ and /b/ as /d/, are extracted for further examination. Figure 4-2 shows mean MFSCs averaged over the training set for these frequent confusion pairs that occur due to testing on the Telephone.

As measured by normalized distance over the 30 lower MFSCs, the misclassified Telephone phoneme is often closer to the incorrect Sennheiser phoneme than the correct Sennheiser phoneme. Even within the Telephone bandwidth, the Telephone shows deviations at all frequencies. The confusion of /cl/ as /n/ can be explained by the increase in background energy due to transmission and normalization effects. The confusions of /l/ as /r/ and /iy/ and /ih/ can be explained by the variations in spectral energy within the formant region. The confusion of /b/ as /d/ can be explained by the variations in low frequency energy.

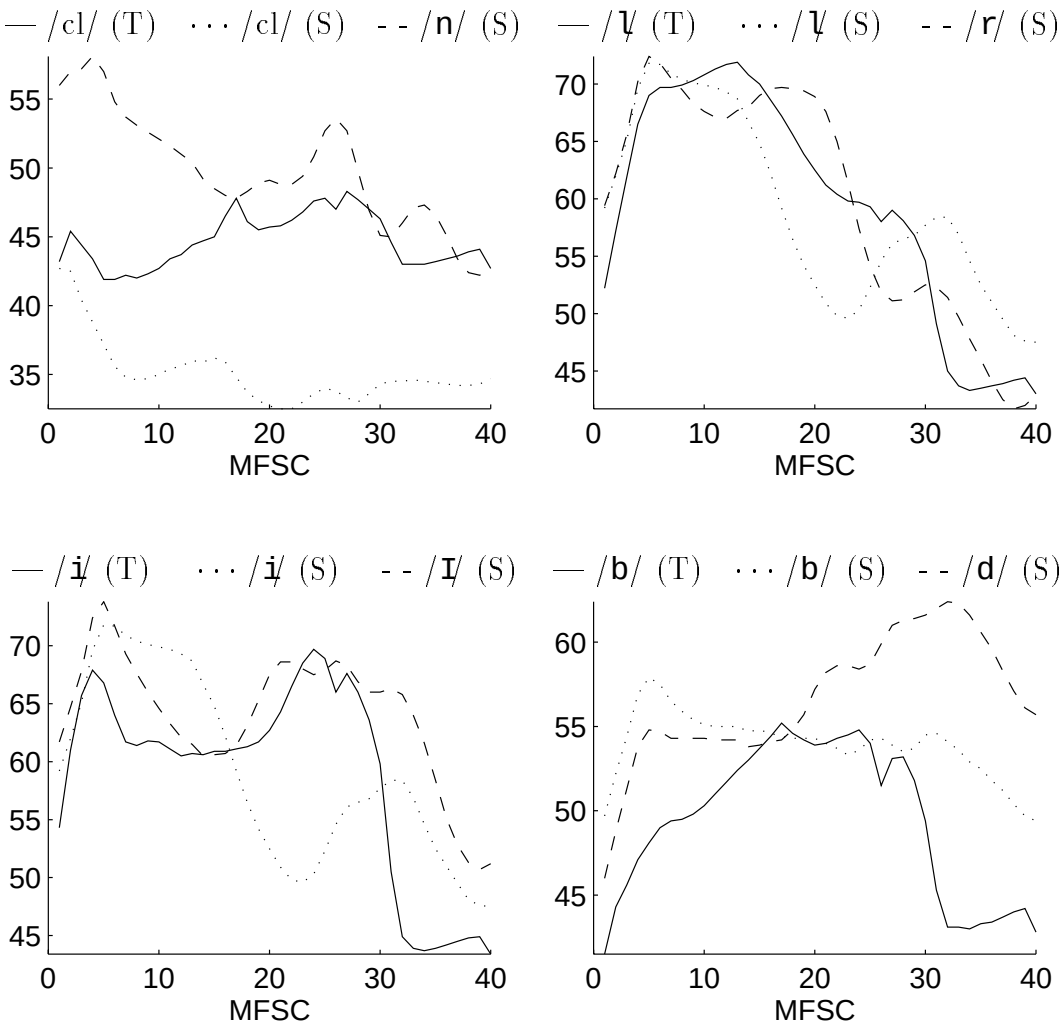

Figure 4-2: Mean MFSCs over the training set for frequent confusion pairs due to the Telephone

4.4.2 Recognition

When testing on the Telephone, the recognition error rate increases by 15%, from 51.2% under (S, S, 16) to 73.1% under (S, T, 8). The substitution and deletion rates respectively increase by 18% and 137%, while the insertion rate decreases by 60%.

Substitutions

18% of the additional error are substitutions. Table 4.16 shows the frequency in percent of the most frequent substitutions of various phonemes with their most frequent misclassifications due to testing on the Telephone.

Phoneme	Frequency	Misclassification
cl	38.3	f
t	9.0	r
s	6.0	r

Table 4.16: Frequency in percent of the most frequent substitutions with their most frequent misclassifications due to the Telephone

/cl/ accounts for almost 40% of additional substitutions and is most frequently confused with the weak unvoiced fricative, /f/. /cl/ is also confused with other weak events such as the lax vowel, /I/, and the nasal, /n/. These errors are presumably due to deviations that obscure weak features. In addition, the unvoiced stop, /t/, and fricative, /s/, are often confused with /r/. These errors are presumably due to deviations that obscure mid and high frequency features.

Deletions

The remaining 82% of the additional errors are deletions. Table 4.17 shows the frequency in percent of the most frequent deletions of various phonemes due to testing on the Telephone.

The table shows the phonemes that account for the largest increase in deletions, totaling to approximately 50% of the additional deletions between (S, T, 8) and

Phoneme	Frequency
cl	22.5
z	4.2
|	4.1
s	3.7
I	3.7
t	3.2
b	3.1
D	3.0
n	2.9

Table 4.17: Frequency in percent of the most frequent deletions due to the Telephone

(S, S, 16). /cl/ accounts for almost 25% of the additional deletions. Other weak events, such as the bandlimited obstruents, /z/, /s/, /t/, /b/ and /D/, the reduced vowels, /|/ and /I/, and the nasal, /n/, are also often deleted. These deletions are presumably due to transmission effects that obscure the presence of weak events. These deletions also resemble the errors in classification and suggest that some of the previous analyses can be applied to recognition. The large differences between the Telephone and Sennheiser suggest very low classification scores that may lead to large increases in deletions.

Insertions

In comparing (S, T, 8) to (S, S, 16), the insertion rate decreases. Table 4.18 shows frequency in percent of the most frequent insertions of various phonemes that do not occur when testing on the Telephone.

As with the B&K, phonemes that are often deleted may not be inserted as often. For example, /n/, /cl/ and /t/ account for large percentages of the decrease in insertions and the increase in deletions.

Phoneme	Frequency
n	21.4
cl	15.8
h#	9.9
t	9.6

Table 4.18: Frequency in percent of the most frequent insertions that do not occur on the Telephone

4.5 Summary

Baseline experiments are performed in phonetic classification and recognition for all microphone conditions before and after downsampling. Analysis focuses on the realistic mismatched conditions when the system is trained on the Sennheiser and tested on the B&K or Telephone. The classification and recognition systems suffer moderate performance degradations when tested on the B&K and severe performance degradations when tested on the Telephone. Although recognition is more difficult to analyze, with the use of consistent systems, the additional recognition errors, mostly deletions and substitutions, are shown to resemble the additional classification errors, and the understanding gained from classification may generalize to recognition. Both mismatched conditions suffer large increases in error due to closures and other weak events and vowels and semivowels that can be easily confused. In addition, the B&K shows large increases in error between voiced and unvoiced obstruents due to its deviation from the Sennheiser at low frequencies, and the Telephone shows large increases in error rate between all obstruents due to its deviation at high frequencies.

Chapter 5

Preprocessing Techniques

Experiments are performed with preprocessing techniques that do not require microphone-specific data. These experiments are compared and analyzed to understand the effects of preprocessing on the speech recognition system. Analysis focuses on the techniques that achieve the lowest error rates under the mismatched conditions when training on the Sennheiser and testing on the B&K or Telephone.

5.1 Description

Of the techniques developed towards microphone robustness, the most common preprocessing techniques apply signal processing algorithms directly to the recorded signal in order to compensate for the effects of microphone variations before input to the speech recognition system. Preprocessing techniques vary from requiring no microphone-specific data to requiring simultaneously recorded microphone training data. This thesis focuses on preprocessing techniques that do not use microphone-specific data.

5.1.1 Convolutional Effects

Mean Normalization (MN) [3, 23] and Relative Spectral Processing (RASTA) [14] are two related techniques that focus on compensating for effects that are modeled by

convolution in the time domain. Broad Class Mean Normalization (BCMN) applies MN to segment-based systems and uses a preliminary broad class hypothesis for each segment. All of these techniques assume that the convolutional effects on the recorded signal do not vary over the interval of interest, and estimate compensation vectors to subtract in the cepstral domain.

MN

In MN [3, 23], Equation 5.1 is applied to each cepstral frame in the utterance.

$$\bar{y}[m] = \bar{x}[m] - \frac{1}{N}\sum_{n=1}^{N}\bar{x}[n] \tag{5.1}$$

where

$\bar{x}[m]$: cepstral vector for frame m

$\bar{y}[m]$: compensated cepstral vector for frame m

N : number of frames in the utterance

$1 \leq m \leq N$

The MN compensation vector for each frame in the utterance is the cepstral mean vector averaged over all frames in the utterance. Subtracting this vector removes the non-varying component over each utterance and normalizes the mean across all utterances. The development set is used to experiment with various mean estimates. For example, shorter term means can be obtained by averaging over specified numbers of frames, and longer term means by averaging over all eight utterances spoken by each speaker. The results presented are obtained using the cepstral mean vector estimated from each utterance, which achieves comparable or lower error rates under mismatched conditions than other estimates investigated.

RASTA

In RASTA [14], Equation 5.2 is applied to each cepstral frame in the utterance.

$$\bar{y}[m] = \bar{x}[m] - \bar{x}[m-1] + \alpha \bar{y}[m-1] \tag{5.2}$$

where

$\bar{x}[m]$: cepstral vector for frame m

$\bar{y}[m]$: compensated cepstral vector for frame m

α : exponential decay factor

RASTA applies an exponentially decaying highpass filter to each utterance in order to remove the slowly varying cepstral components over each utterance. In comparison to MN, the RASTA compensation vector varies for each frame and is a weighted average computed over an interval that depends on the exponential decay factor, α. The development set is used to experiment with different values of α. As α approaches 1, RASTA achieves lower error rates under mismatched conditions. In the limit, when α is 1, the average is computed over the entire utterance and removes the non-varying component, as in MN. The results presented are obtained using an α of 0.99.

BCMN

As shown in Chapters 3 and 4, microphones have different effects on different phonetic classes depending on their spectral characteristics, suggesting that techniques should also vary for different classes. BCMN attempts to account for broad class effects in a segment-based system by making a preliminary broad class hypothesis for the segment in the utterance and using these hypotheses to estimate cepstral compensation vectors to subtract. Experiments show that, given the correct broad class instead of the hypothesis, BCMN is very effective, suggesting that the technique can reduce confusions over features that discriminate phonemes within broad classes. In comparison to MN, the BCMN compensation vector varies for each broad class and

is the cepstral mean averaged over the segments that are hypothesized to belong to that broad class. The development set is used to experiment with different broad class groupings. As the number of broad classes approaches 1, BCMN achieves lower error rates under mismatched conditions. The results presented are obtained using 2 broad classes, one containing the sonorant vowels, semivowels and nasals and the other containing the obstruents and silences.

5.1.2 Additive Effects

Log-Spectral Subtraction (SUB) [33] focuses on compensating for effects that are modeled by addition in the time domain. Segment-based subtraction (SSUB) applies SUB to segment-based systems. Both techniques assume that the additive effects on the recorded signal are uncorrelated with speech and do not vary over the interval of interest, and estimate compensation vectors to subtract in the log spectral domain.

SUB

In SUB [33], Equation 5.3 is applied to each log spectral frame in the utterance.

$$\bar{Y}[m] = \bar{X}[m] + \max(\log(1 - 10^{\bar{X}[m]-\bar{N}[m]}), \bar{C}_{max}[m]) \tag{5.3}$$

where

$\bar{X}[m]$: log spectral vector for frame m
$\bar{Y}[m]$: compensated log spectral vector for frame m
$\bar{N}[m]$: estimated log spectral noise vector for frame m
$\bar{C}_{max}[m]$: maximum log spectral compensation vector for frame m

The SUB compensation vector for each frame in the utterance is a function of the recorded log spectral vector and the estimated noise vector. Adding this compensation vector removes the effects of noise in the log spectral domain. The development set is used to experiment with various noise estimates. For example, noise can be estimated

by setting thresholds on histograms or by averaging over silence regions. The results presented are obtained using the noise vector estimated from the beginning and ending silence regions for each utterance.

SSUB

Segment subtraction (SSUB) attempts to make more accurate estimates for a segment-based system by averaging over the frames in each segment. In comparison to SUB, the SSUB log spectral compensation vector varies for each segment rather than for each frame.

5.1.3 Combined Effects

The linear cascade of SUB and MN (SUBMN) uses independent compensation vectors to account for convolutional and additive effects. Codeword-Dependent Cepstral Normalization (CDCN) [1] uses joint compensation vectors to account for the combined effects. Both techniques assume that the additive effects on the recorded signal are uncorrelated with speech.

SUBMN

SUBMN compensates for additive and convolutional effects by applying SUB followed by MN. Experiments show that techniques that are effective individually are often enhanced when combined and that the removal of additive followed by convolutional effects is often more effective. The results presented are obtained using SUB in cascade with MN, which achieves comparable or better results than other combinations investigated.

CDCN

CDCN [1] compensates for joint effects by applying Maximum Likelihood (ML) estimation to determine convolutional and additive parameters and Minimum Mean

Square Error (MMSE) estimation to determine compensation vectors. Of all the techniques described, CDCN is by far the most complex and is the only technique not explictly implemented for this thesis. For details, refer to Acero's doctoral thesis [1].

5.2 Comparison

Classification and recognition experiments under different microphone conditions are performed with the preprocessing techniques. Results are compared in order to understand the ability of preprocessing techniques to increase microphone robustness.

5.2.1 Classification

Table 5.1 shows classification error rates in percent for various preprocessing techniques. Under (S, S), the baseline and compensated classification error rates are shown

	(S, S)	(S, B)	(S, T)
Baseline	31.2	38.9	55.7
MN	31.3	34.6	52.5
RASTA	32.8	35.9	55.6
BCMN	32.0	35.5	55.4
SUB	31.9	39.5	55.1
SSUB	31.9	39.1	54.0
SUBMN	31.9	34.6	51.7
CDCN	31.7	34.2	55.2

Table 5.1: Classification error rates in percent for various preprocessing techniques

in the first column. For each training microphone, the baseline is the lower bound on error rate, and preprocessing may increase error rates under matched conditions. These increases in error rate for the matched conditions reflect the cost of using the techniques to increase microphone robustness. With the exception of RASTA, none of the techniques cause statistically significant changes in error rate, suggesting that the cost of using the techniques to increase microphone robustness in classification is

small if not insignificant.

For each testing microphone, the baseline is the upper bound on error rate, and preprocessing usually decreases error rates under mismatched conditions. These decreases in error rate with preprocessing reflect the ability of the techniques to reduce the degradations due to mismatched testing. However, preprocessing usually does not decrease error rates under mismatched conditions below error rates under matched conditions. With respect to the matched condition, the remaining error rate increases reflect the inability of the techniques to achieve complete microphone independence.

Under (S, B), the baseline and compensated classification error rates are shown in the second column of Table 5.1. For reference, the baseline classification error rate under (B, B) is 33.1%. All of the techniques significantly decrease error rate, with the exception of SUB and SSUB. Since SUB and SSUB focus on additive effects, and the combined techniques achieve comparable results to techniques that focus on convolutional effects, the differences between recording on the B&K and Sennheiser in a noise-isolated environment may be mostly convolutional rather than additive. Of the techniques, MN, SUBMN and CDCN insignificantly differ from each other in achieving the largest reductions in error rate, decreasing error by more than 10% and resulting in less than 5% increases in error from (B, B). This suggests that preprocessing can effectively compensate for the differences between the B&K and Sennheiser.

Under (S, T), the baseline and compensated classification error rates are shown in the third column. For reference, the baseline classification error rate under (T, T) is 41.4%. None of the techniques significantly decrease error rate, with the exception of MN and SUBMN. MN and SUBMN insignificantly differ from each other in achieving the largest reductions in error rate, decreasing error by approximately 5%. Nevertheless, they are less effective for the Telephone than for the B&K, resulting in approximately 25% increases in error from (T, T). This suggests that preprocessing cannot compensate for the larger differences between the Telephone and Sennheiser.

Overall, MN and SUBMN are the most effective techniques in classification. Under matched conditions, these techniques do not degrade performance. Under mis-

matched conditions, they significantly reduce the moderate degradations in the B&K and slightly reduce the severe degradations in the Telephone.

5.2.2 Recognition

Table 5.2 shows recognition error rates in percent for various preprocessing techniques. Recognition error rates are higher than classification error rates but follow the same

	(S, S)	(S, B)	(S, T)
Baseline	51.2	59.0	73.1
MN	51.1	54.8	71.0
RASTA	52.8	56.2	72.1
BCMN	53.0	58.8	76.1
SUB	50.9	58.1	74.2
SSUB	51.4	58.3	73.9
SUBMN	51.7	55.3	73.2
CDCN	51.1	54.6	67.7

Table 5.2: Recognition error rates in percent for various preprocessing techniques

trends. Under (S, S), the preprocessing techniques cause small changes in error rate, reflecting that the cost of using the techniques to increase microphone robustness is small.

Under (S, B), all of the techniques substantially decrease recognition error rate, with the exception of SUB and SSUB, which do not focus on compensating for the convolutional differences between the B&K and Sennheiser. MN and CDCN achieve the largest decreases in error rate, by more than 7%, lowering the (S, B) error rate to within 2% of the baseline (B, B) error rate of 52.9%.

Under (S, T), the large differences between the Telephone and Sennheiser render most of the techniques ineffective in recognition, and only MN, RASTA and CDCN substantially decrease error rate. CDCN achieves the largest decrease in error rate, by 7%, but still results in a 13% increase from the baseline (T, T) error rate of 59.8%.

The recognition results are similar to those in classification except that CDCN is more effective than the linear cascade of SUB and MN. Overall, MN and CDCN are

the most effective techniques in recognition. These techniques maintain performance under (S, S), significantly improve performance under (S, B) and slightly improve performance under (S, T).

5.2.3 Discussion

Preprocessing techniques are benchmarked in phonetic classification and recognition. Of the techniques, MN, SUBMN and CDCN are most effective in increasing microphone robustness for the TIMIT microphones, while RASTA and BCMN are moderately effective, and SUB and SSUB have insignificant effects. Without increasing error rates for the training microphone, the most effective techniques can maintain lower error rates for testing microphones that are relatively similar to the training microphone, but more severely mismatched conditions still suffer significant performance degradations. These techniques are further analyzed in the following section.

As an additional note, the preprocessing techniques are selected and implemented based on a study of previous work and may not cover all approaches or use optimal algorithms. The thesis does not aspire to achieve comprehensive coverage or optimal implementation, nor does it attempt to determine whether one technique is absolutely better than another. Rather, a consistent methodology is used to make comparisons in order to improve understanding of the effects of preprocessing techniques on the speech recognition system. Regardless, results always reflect the systems, corpora and microphones used. For example, RASTA [14] is developed for a system that uses a linear predictive front end, unlike the Mel-frequency cepstral front end used in SUMMIT. SUB [33] and SSUB are developed for corpora that mainly differ in additive effects, unlike the Sennheiser and B&K corpora in TIMIT.

Furthermore, researchers have made and continue to make improvements towards increased microphone robustness. In this thesis, modifications have been made, algorithms have been developed, and improvements have been proposed. For example, both SSUB and BCMN are developed for segment-based systems. These techniques rely on segment hypotheses and can be improved with more robust segmentation. BCMN also relies on broad class hypotheses and can be improved with more robust

classification. In addition, instead of applying preprocessing techniques prior to and separate from recognition, these techniques can be incorporated with other components in the speech recognition system. For example, BCMN can be combined with acoustic and language modeling to produce more robust broad class hypotheses and more effective compensation. Preprocessing can also be combined with training to produce more robust models, as discussed in the next chapter.

5.3 Analysis

Of the techniques, MN, SUBMN and CDCN are most effective, but since the cascade of SUB and MN does not offer significant advantages over using MN alone, analysis focuses on the MN and CDCN techniques, which represent the extremes with regard to algorithmic and computational complexity. The MN [3, 23] algorithm estimates a compensation vector using the mean averaged over each frame in the utterance. With this simple algorithm, MN consistently achieves comparable or lower classification and recognition error rates under all microphone conditions, except (S, T) in recognition. The CDCN [1] algorithm estimates a compensation vector using iterative ML and MMSE estimation techniques. With this complex algorithm, CDCN achieves comparable results to MN, with superior performance under (S, T) in recognition. Analysis focuses on MN, but CDCN is analyzed where its abilities exceed those of MN.

By subtracting the cepstral mean from each utterance, MN normalizes the cepstral mean averaged over the training set for each microphone to zero. Since the cepstral and log spectral coefficients are related by a linear transformation, the log spectral mean is likewise normalized. The general effects of this compensation can be analyzed by broad class. Figure 5-1 shows mean broad class MFSCs averaged over the training set for each microphone after MN. In comparison to Figure 3-3, MN reduces the difference between the B&K and Sennheiser for all broad classes, as measured by normalized distance over 40 MFSCs. For example, MN compensates for the low frequency peak in the B&K that is presumably due to nasal and glottal resonances

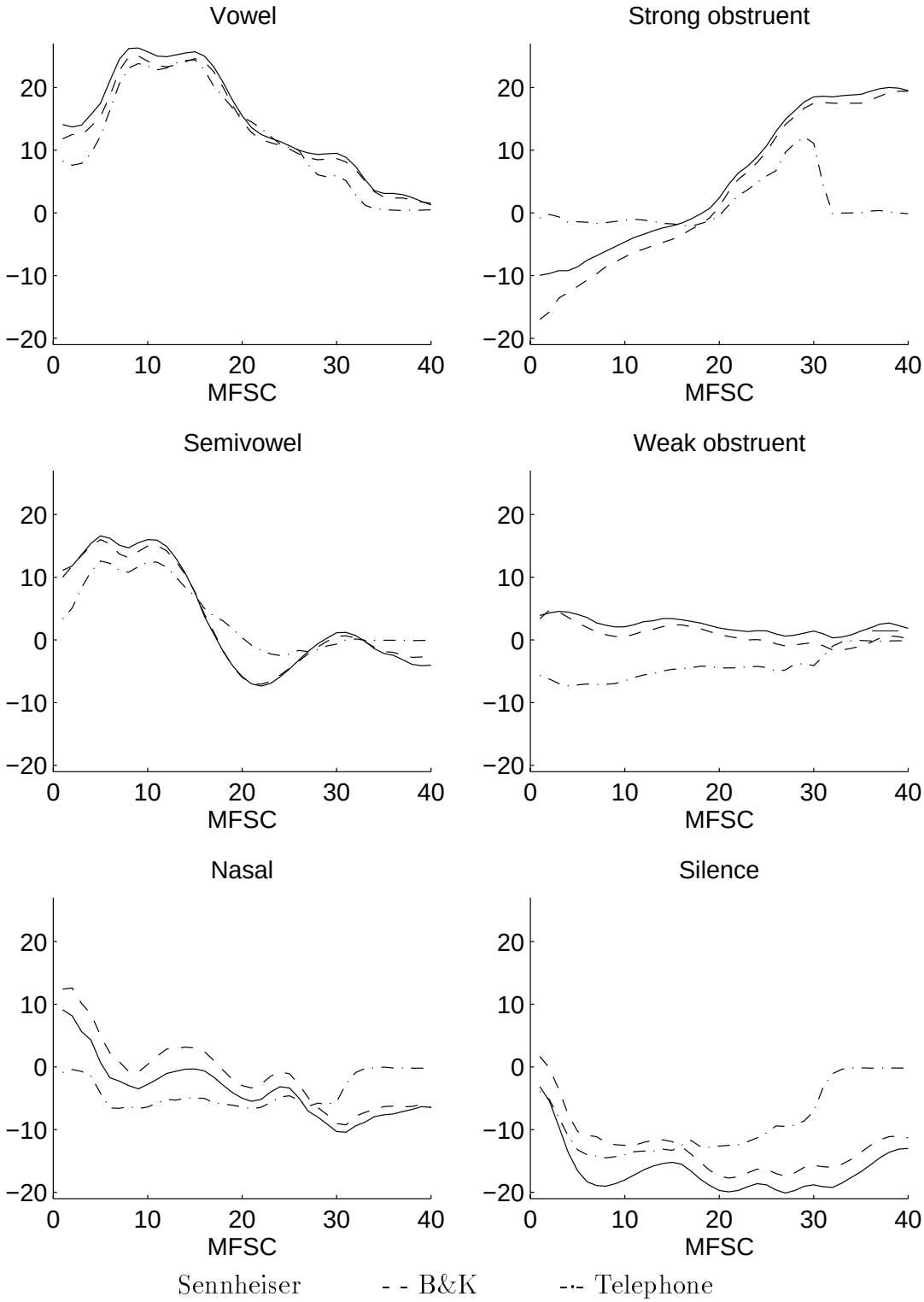

Figure 5-1: Mean broad class MFSCs over the training set for each microphone after MN

and environmental effects, suggesting fewer confusions between the nasal and silence classes and within the obstruent and vowel classes. As measured by normalized distance over the lower 30 MFSCs, MN reduces the differences between the Telephone and Sennheiser for all broad classes, except the weak obstruent class. For example, MN compensates for the variations in the vowel and semivowel classes that may obscure formant energies, suggesting fewer confusions within and between the vowels and semivowels. Overall, the effects of the preprocessing suggest reductions in error under mismatched conditions, although in comparison to the B&K, the Telephone still shows significant residual deviations even after preprocessing. A more detailed analysis of (S, B) and (S, T) in classification and recognition show the types of errors that can be effectively compensated for and the types of errors that still occur despite preprocessing.

5.3.1 (S, B)

In the previous chapter, the baseline (S, B) and (S, S) conditions are compared to analyze the additional errors due to mismatched testing. In this section, a comparison is made of the (S, B) condition, before and after preprocessing, to show how the baseline errors are affected by preprocessing. Analysis focuses on MN, since CDCN does not offer significant advantages for (S, B).

Classification

MN decreases the classification error rate by 11%, from 38.9% under the baseline to 34.6% after preprocessing. Analysis focuses on those tokens that are incorrectly classified under the baseline but correctly classified after MN. Table 5.3 shows the frequency in percent of the most frequent misclassifications of these tokens with their most frequent substitutions. The table shows the eight phonemes that are most often misclassified under the baseline but not after MN, totaling to approximately 50% of the reduction in error caused by preprocessing. In comparison to Table 4.8, these corrections are similar to the errors caused by mismatched testing on the B&K rather

Phoneme	Frequency	Substitution
s	9.9	z
p	7.1	d
E	6.9	I
t	6.5	d
v	5.8	m
f	5.5	v
O	4.8	o
e	4.4	i

Table 5.3: Frequency in percent of the most frequent misclassifications with their most frequent substitutions that do not occur on the B&K after MN

than the Sennheiser. For example, MN reduces the misclassifications between unvoiced obstruents, such as /s/, /t/ and /f/, and their voiced counterparts, /z/, /d/ and /v/, between weak obstruents and nasals, such as /v/ and /m/, and between vowels, such as /E/ and /I/. This suggests that the preprocessing effectively compensates for some of the differences at low frequencies, allowing improved discrimination of voicing, nasal, formant and noise energies.

Despite these error reductions, mismatched testing on the B&K, even after MN, results in an 8% increase in error rate from 32.1% under (S, S) to 34.6% under (S, B). Although MN corrects many of the frequent misclassifications, some misclassifications still occur after preprocessing. For example, MN does not correct many of the errors involving closures, which account for the largest percentage of the additional errors under the baseline. After MN, the total number of misclassifications decreases, but the percent of the additional misclassifications of closures increases to 25%. This suggests that this technique does not effectively compensate for all of the deviations due to prevoicing and noise.

Figure 5-2 shows mean MFSCs averaged over the training set for the four frequent confusion pairs in Figure 4-1 that occur due to testing on the B&K after MN. In comparison to the baseline, the compensated B&K phoneme, denoted by the solid line, is often closer to the target Sennheiser phoneme, denoted by the dashed line, rather than the misclassified Sennheiser phoneme, denoted by the dotted line. MN

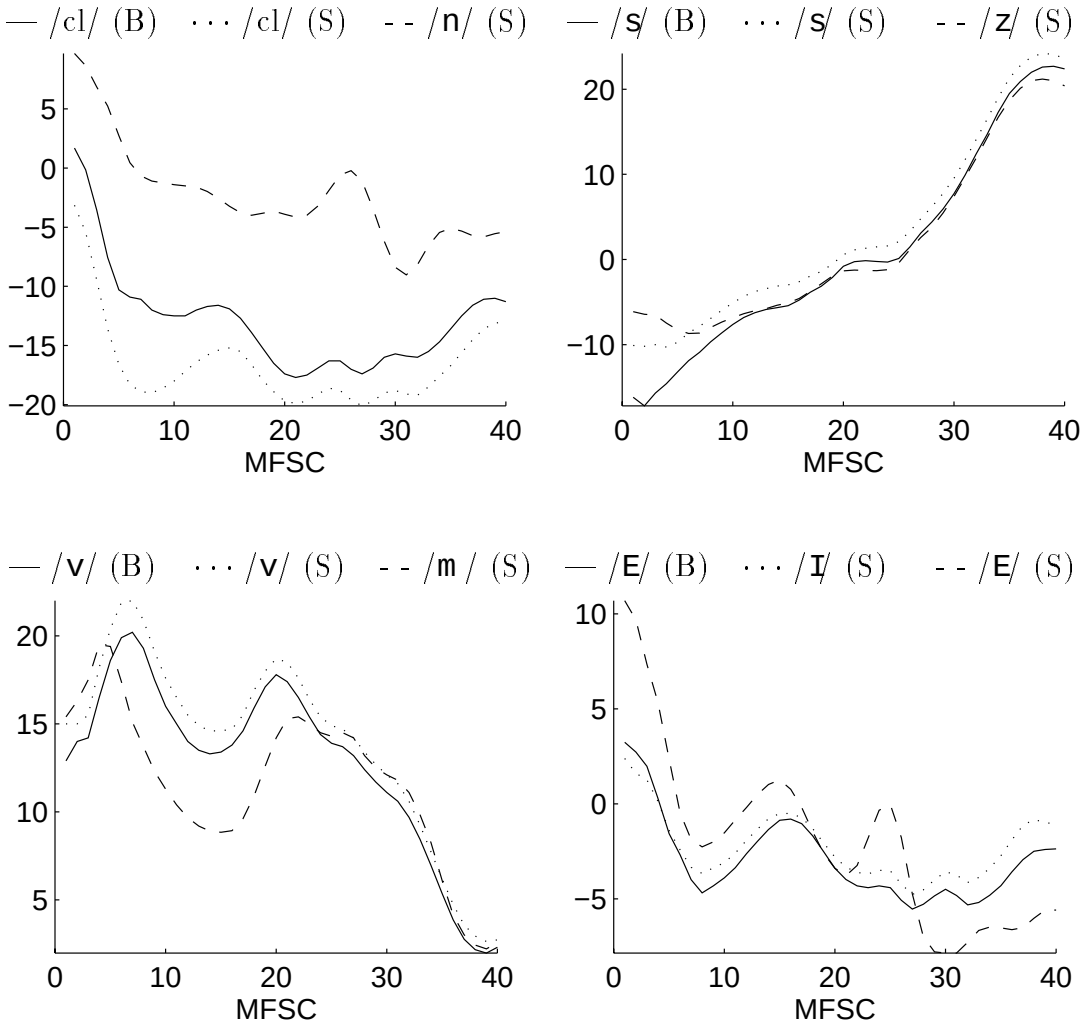

Figure 5-2: Mean MFSCs over the training set for frequent confusion pairs due to the B&K after MN

effectively compensates for many of the differences at low frequencies, suggesting fewer misclassifications after preprocessing. For example, the low frequency variation that confuses the voicing feature between /s/ and /z/ is removed. The low frequency peaks that confuse the voiced and nasalized events, /v/ and /m/, are separated. The low frequency peaks that confuse the formants of /E/ and /I/ are also discriminated. Relative to these corrections, the confusion between /cl/ and /n/ is not as effectively compensated for, suggesting misclassifications even after preprocessing.

Recognition

Table 5.4 shows recognition error rates in percent before and after MN. MN decreases

	Total	Substitution	Deletion	Insertion
Baseline	59.0	30.7	20.7	7.7
MN	54.8	29.0	16.2	9.6

Table 5.4: Recognition error rates in percent before and after MN

the total recognition error rate by 7%. As in classification, MN effectively compensates for many, but not all, of the additional errors caused by mismatched testing on the B&K rather than the Sennheiser. After preprocessing, testing on the B&K still causes a 7% increase in recognition error rate from 51.2% under (S, S) to 54.8% under (S, B). Analysis focuses on the more frequent substitutions and deletions.

Decreases in substitutions account for 27% of the reduction in error. Table 5.5 shows the frequency in percent of the most frequent substitutions of various phonemes with their most frequent misclassifications that do not occur when testing on the B&K after MN. The table shows the seven phonemes that are most often substituted un-

Phoneme	Frequency	Misclassification
s	10.1	z
n	9.4	m
k	7.9	t
cl	7.2	D
E	6.8	I
r	6.5	5
t	6.1	d

Table 5.5: Frequency in percent of the most frequent substitutions with their most frequent misclassifications that do not occur on the B&K after MN

der the baseline but not after MN, totaling to approximately 50% of the decrease in subsitutions caused by preprocessing. MN effectively compensates for many of the baseline substitutions shown in Table 4.10. For example, MN reduces the substitutions within obstruents, such as /s/ and /z/, /k/ and /t/, and /t/ and /d/. MN also

reduces some of the substitutions involving closures. Despite these error reductions, MN only corrects some of the baseline errors, and many of the same substitutions still occur after preprocessing. Even after MN, mismatched testing on the B&K results in a 9% increase in substitution rate.

Decreases in deletions account for the remaining 73% of the reduction in error. Table 5.6 shows the frequency in percent of the most frequent deletions of various phonemes that do not occur when testing on the B&K after MN. The table shows

Phoneme	Frequency
cl	14.6
n	7.7
t	4.9
l	4.4
p	4.4
E	4.1
a	3.8
|	3.8
}	3.8

Table 5.6: Frequency in percent of the most frequent deletions that do not occur on the B&K after MN

phonemes that are most often deleted under the baseline but not after MN, totaling to approximately 50% of the decrease in deletions. MN effectively compensates for most of the baseline deletions shown in Table 4.11. For example, MN reduces the deletions of weak events such as /cl/, /n/, /t/, /p/ amd /}/. After MN, although mismatched testing on the B&K results in increased substitution and insertion rates, the deletion rate does not change. This suggests that preprocessing can effectively compensate for effects, such as additive noise, which may obscure the presence of weak events. This also suggests that preprocessing may significantly improve classification scores, resulting in fewer deletions.

5.3.2 (S, T)

In comparison to B&K, the Telephone shows larger deviations both at high frequencies and within the Telephone bandwidth that cause larger increases in error when using the Telephone. In this section, a comparison is made of the (S, T) condition, before and after preprocessing. Although MN achieves lower error rates in classification, CDCN achieves lower error rates in recognition. Analysis attempts to reveal the advantages of each technique.

Classification

MN decreases the classification error rate by 6%, from 55.7% under the baseline to 52.5% after preprocessing. Of those tokens that are incorrectly classified under the baseline but correctly classified after MN, Table 5.7 shows the frequency in percent of the most frequent misclassifications with their most frequent substitutions. The

Phoneme	Frequency	Substitution
l	18.8	r
cl	13.1	n
z	7.4	v
l	6.8	r
{	6.0	r

Table 5.7: Frequency in percent of the most frequent misclassifications with their most frequent substitutions that do not occur on the Telephone after MN

table shows the phonemes that are most often misclassified under the baseline but not after MN, totaling to approximately 50% of the error reduction. In comparison to Table 4.15, these corrections are similar in kind to the errors caused by mismatched testing on the Telephone, but preprocessing only compensates for a small fraction of the large numbers of baseline errors. Even after MN, using the Telephone results in a 68% increase in error rate from 32.1% under (S, S) to 52.5% under (S, T), and all of the baseline misclassifications still occur after preprocessing. This suggests that the preprocessing can only partially compensate for the large differences between the

Telephone and Sennheiser.

Figure 5-3 shows mean MFSCs averaged over the training set after MN for two of the frequent confusion pairs in Figure 4-2 that occur due to testing on the Telephone after MN. In comparison to Table 4-2, MN reduces some of the differences between

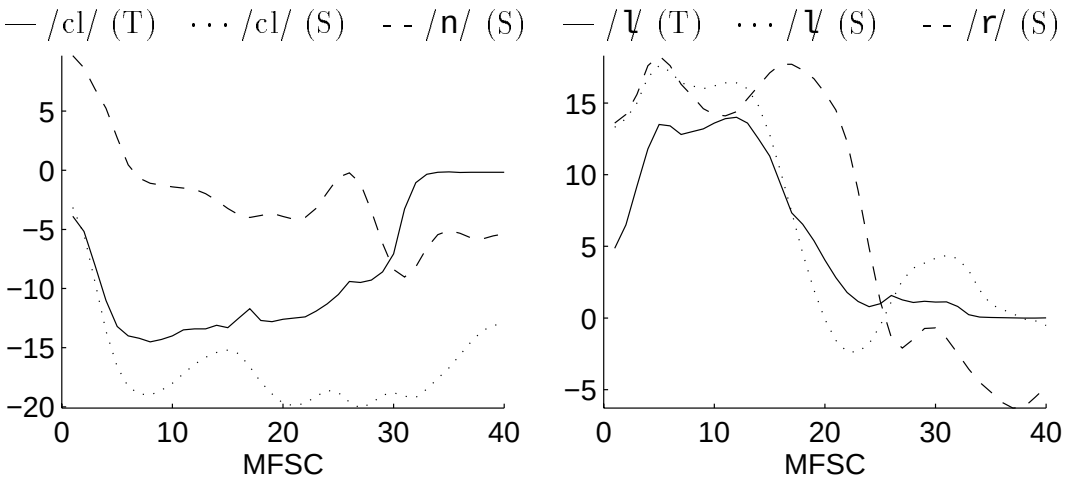

Figure 5-3: Mean MFSCs over the training set for frequent confusion pairs due to the Telephone after MN

the Telephone and Sennheiser, but preprocessing is not as effective for the Telephone as the B&K. For example, the Telephone /cl/ is closer to the Sennheiser /cl/ than the Sennheiser /n/, but the Telephone still deviates widely from the Sennheiser. Similarly, the Telephone /ɻ/ is closer to the Sennheiser /ɻ/ than the Sennheiser /r/, but the formants are not clearly discriminated. This suggests difficulty in classification for the Telephone even after preprocessing.

Recognition

Table 5.8 shows recognition error rates in percent before and after preprocessing. With such poor results in classification, MN is even less effective in recognition, decreasing the error rate by only 3%. As in classification, MN effectively compensates for only a fraction of the large numbers of additional errors caused by mismatched testing on the Telephone. Furthermore, MN is not able to compensate for the deletions

	Total	Substitution	Deletion	Insertion
Baseline	73.1	31.4	38.2	3.4
MN	71.0	30.8	37.2	3.1
CDCN	67.7	32.7	31.2	3.8

Table 5.8: Recognition error rates in percent before and after preprocessing

of closures that account for the largest percentage of the error under the baseline.

Although CDCN results in a higher total error rate in classification than MN, it effectively compensates for many more of the baseline misclassifications of closures. This difference is presumably responsible for the ability of CDCN to achieve lower error rates in recognition. In recognition, CDCN is the most effective of all the techniques in compensating for mismatched testing on the Telephone, decreasing the error rate by 7%. All of the error reduction is in deletions. Table 5.9 shows the frequency in percent of the most frequent deletions of various phonemes that do not occur when testing on the Telephone after CDCN. The table shows the two

Phoneme	Frequency
cl	44.3
k	5.3

Table 5.9: Frequency in percent of the most frequent deletions that do not occur on the Telephone after CDCN

phonemes that are most often deleted under the baseline but not after CDCN, totaling to approximately 50% of the decrease in deletions. CDCN effectively compensates for the baseline deletions shown in Table 4.17, especially the deletion of /cl/. Despite these error reductions, CDCN cannot compensate for the large numbers of errors caused by testing on the Telephone, and even after CDCN, the recognition error rate increases by 32% from 51.2% under (S, S) to 67.7% under (S, T).

5.4 Summary

Experiments with preprocessing techniques are compared and analyzed to understand their effects on the speech recognition system. The most effective techniques compensate for most of the relatively small differences in the B&K but only a fraction of the larger differences in the Telephone. As a result, preprocessing can increase microphone robustness to small mismatches between training and testing conditions but cannot maintain low error rates under severely mismatched conditions.

Chapter 6

Training Techniques

Although preprocessing techniques can reduce performance degradations under mismatched conditions, they do not enable the speech recognition system to achieve microphone robustness. By nature, preprocessing techniques attempt to compensate for microphone effects on the recorded signal without affecting the recognition process. Since the data used to train the acoustic models may be recorded using a microphone different from the one used in testing, regardless of the compensation, the recognizer parameters may still be suboptimal, thus leading to performance degradations. With the availability of microphone-specific training data in TIMIT, a more effective technique may be to apply algorithms that directly account for microphone variations in training, thereby reducing mismatch and performance degradations.

This chapter explores the increases in microphone robustness that can be achieved by using microphone-specific data in training. Experiments are conducted using multi-style training and microphone selection. Multi-style training [22] involves pooling the data into one set of models. Microphone selection involves training a separate set of models for each microphone and automatically selecting the best models during testing. These techniques are compared with the baseline and preprocessing techniques to understand the advantages of using microphone-specific training techniques.

6.1 Description

6.1.1 Multi-style Training

Multi-style training (MULTI) [22] was originally used to train a speech recognition system on multiple speaking styles in order to increase robustness to mismatched conditions when the system is trained on normal speaking styles and tested on abnormal speaking styles, such as speaking under stress. In general, these experiments show that incorporating multiple styles in training improves performance under different styles in testing. Multi-style training decreases the mismatch between training and testing conditions by incorporating different conditions in training. As a result, the model parameters, such as the means and variances use in the previous chapters, are averaged over different data and better matched to variations in testing. In addition, multi-style training may produce more robust models by causing the system to focus on acoustic features that are consistent across conditions. As a result, the models may be richer and provide a better description of more robust features.

For microphone robustness, multi-style training involves training the system on a composite training set consisting of one third of the original training sets from each of the TIMIT microphones. By using only one third of the data for each microphone, the size of the composite set is identical to the baseline in order to ensure a fair comparison. The thirds do not overlap and contain the same number, plus or minus one, of male and female speakers and utterances.

In order to determine the effects of downsampling on multi-style training, experiments are performed at both 16 and 8 kHz. Table 6.1 shows classification and recognition error rates in percent for multi-style training before and after downsampling.

In classification, the combination of multi-style training with downsampling does not cause statistically significant changes in error rate for the Telephone and slightly decreases the error rate for the Sennheiser and B&K. Similarly, in recognition, the combination results in comparable but slightly higher error rates. Since multi-style training accounts for the differences between the Sennheiser and Telephone by training

	Classification			Recognition		
	Sennheiser	B&K	Telephone	Sennheiser	B&K	Telephone
16 kHz MULTI	34.3	35.8	45.5	54.6	56.9	63.9
8 kHz MULTI	35.8	37.2	45.5	57.3	59.5	66.5

Table 6.1: Classification and recognition error rates in percent for multi-style training before and after downsampling

on utterances with both spectral distributions, the models are more robust and can capture information in different spectral regions. Therefore, it is not necessary to downsample in order to focus the system on acoustic features in the low frequency regions.

In order to determine the effects of preprocessing on multi-style training, experiments are also performed in combination with the preprocessing techniques. Table 6.2 shows classification and recognition error rates for multi-style training in combination with various preprocessing techniques.

	Classification			Recognition		
	Sennheiser	B&K	Telephone	Sennheiser	B&K	Telephone
MULTI	34.3	35.8	45.5	54.6	56.9	63.9
MULTI-MN	33.4	34.5	44.5	52.9	55.9	64.5
MULTI-RASTA	35.2	36.6	45.7	54.3	57.2	65.4
MULTI-BROAD	35.0	36.1	46.1	54.2	57.0	66.5
MULTI-SUB	35.0	35.9	45.3	53.9	56.8	64.5
MULTI-SSUB	34.3	35.5	44.7	54.4	56.9	64.6
MULTI-SUBMN	34.0	34.5	44.6	53.4	55.8	65.1
MULTI-CDCN	34.7	35.3	45.9	52.1	55.1	64.0

Table 6.2: Classification and recognition error rates in percent for multi-style training in combination with various preprocessing techniques

In classification, the combination of multi-style training with MN or SUBMN results in small decreases in error rate, while the combination with other preprocessing techniques does not cause statistically significant changes in performance. In recognition, the use of preprocessing in addition to multi-style training also does not cause large decreases in error rate. Multi-style training directly accounts for microphone

variations in training rather than compensating for effects on the signal prior to recognition. The reductions in error achieved by multi-style training seem to subsume most of the reductions in error achieved by preprocessing techniques.

6.1.2 Microphone Selection

Microphone selection (SELECT) involves training separate models to match each different testing microphone. Like multi-style training, microphone selection also decreases the mismatch between training and testing conditions by incorporating different data in training. Unlike multi-style training, microphone selection provides a separate model to directly match each condition rather than pooling the data and averaging the model parameters over different conditions. As a result, microphone selection may be able to model and represent more diverse testing conditions, but the models are not more robust in the sense that training separate models does not cause the system to focus on robust features that are consistent across different conditions.

An added complexity in microphone selection is that the matching model must be determined in testing. One method of automatic selection is to use the model that corresponds to the highest scoring output. Another method is to train a separate microphone classifier to determine microphone identity. In the limit, if the microphone can always be determined correctly, for example by the user, the system can always perform under matched conditions, and the robustness issue is avoided.

For our microphone experiments, microphone selection involves using the full training set to train three models, one for each TIMIT microphone. For each testing utterance, the microphone model that corresponds to the highest scoring output is automatically selected. Since separate models are trained to directly match each testing condition, downsampling in order to improve performance under mismatched conditions is not necessary with microphone selection. In fact, the effects of transmission bandlimiting on the Telephone only facilitate the process of automatic microphone selection.

Experiments are performed in combination with preprocessing in order to determine the effects of preprocessing on microphone selection. Table 6.3 shows classifica-

tion and recognition error rates in percent for microphone selection in combination with various preprocessing techniques.

	Classification			Recognition		
	Sennheiser	B&K	Telephone	Sennheiser	B&K	Telephone
SELECT	31.6	33.0	51.0	54.3	54.0	59.8
SELECT-MN	31.0	32.2	40.7	51.0	53.4	60.0
SELECT-RASTA	33.0	33.7	54.4	53.3	56.6	60.8
SELECT-BROAD	32.5	33.6	54.5	53.2	57.1	62.2
SELECT-SUB	32.2	33.3	48.5	51.1	55.4	60.9
SELECT-SSUB	32.4	33.1	49.4	52.2	54.8	60.4
SELECT-SUBMN	31.0	32.8	41.7	51.8	54.4	59.9
SELECT-CDCN	32.5	32.9	46.5	52.0	55.1	59.2

Table 6.3: Classification and recognition error rates in percent for microphone selection in combination with various preprocessing techniques

In classification, the combination of microphone selection with preprocessing can significantly decrease error rates when testing on the Telephone but does not cause statistically significant changes when testing on the Sennheiser and B&K. In recognition, the use of preprocessing in addition to microphone selection achieves small decreases in error rates. Of the techniques, MN is most effective in combination, reducing the classification error rate by 20% for the Telephone and the recognition error rate by 4% for the Sennheiser.

6.2 Comparison

6.2.1 Classification

Table 6.4 shows classification error rates in percent before and after preprocessing and training. MN is used for preprocessing and in combination with multi-style training and microphone selection.

For the Sennheiser, increases in error rate from the baseline reflect the cost of using techniques to increase microphone robustness. Preprocessing and microphone selection do not cause statistically significant changes in error rate, but multi-style

	Sennheiser	B&K	Telephone
Baseline	31.2	38.9	55.7
Preprocessing	31.3	34.6	52.5
Multi-style training	33.4	34.5	44.5
Microphone selection	31.0	32.2	40.7

Table 6.4: Classification error rates in percent before and after preprocessing and training

training increases the error rate by 10%. Preprocessing and microphone selection both produce models that only use high quality training data. These techniques maintain the lowest error rates under the high quality matched condition and can improve microphone robustness at no cost. Multi-style training produces models that use both low and high quality training data. This technique introduces some mismatch when using the Sennheiser and results in performance degradations under high quality conditions.

For the B&K, decreases in error rate from the baseline reflect the ability of the technique to increase microphone robustness between the Sennheiser and B&K. Without microphone-specific data, preprocessing can decrease the error rate by 11%. In comparison, with microphone-specific data, multi-style training does not result in a statistically significant change, but microphone selection achieves an additional 7% decrease in error rate. As discussed in previous chapters, the B&K differs from Sennheiser mainly by a peak in low frequency energy. The use of preprocessing to compensate for these relatively small differences is very effective, at least comparable to the use of pooled models with multi-style training, and only slightly less effective than the use of separate models with microphone selection.

For the Telephone, preprocessing decreases the error rate by 6%. In comparison, multi-style training and microphone selection respectively achieve additional 13% and 22% decreases in error rate. In comparison to the B&K, the differences between the Telephone and Sennheiser are larger, resulting in larger performance degradations. Under such conditions, preprocessing without microphone-specific data is not as effective, and training with microphone-specific data achieves significant reductions in

error rate.

Of all the techniques, microphone selection is most effective. With microphone selection, performance for the three microphones is not statistically different from the baseline matched (S, S), (B, B) and (T, T) conditions. In this sense, microphone selection enables the classification system to achieve microphone robustness between the Sennheiser, B&K and Telephone.

6.2.2 Recognition

Table 6.5 shows recognition error rates in percent before and after preprocessing and training. Recognition error rates are higher but follow the trends in classification.

	Sennheiser	B&K	Telephone
Baseline	51.2	59.0	73.1
Preprocessing	51.1	54.8	71.0
Multi-style training	52.9	55.9	64.5
Microphone selection	51.0	53.4	60.0

Table 6.5: Recognition error rates in percent before and after preprocessing and training

Under matched conditions, preprocessing does not increase the error rate. Under mismatched conditions, preprocessing compensates for the differences between the B&K and Sennheiser and effectively decreases the (S, B) error rate, but preprocessing cannot compensate for the larger deviations for the Telephone and only slightly reduces the (S, T) error rate. In comparison, multi-style training can significantly decrease the error rate when using the Telephone, at the cost of increasing the error rates when using the Sennheiser and B&K. This averaging of error rates corresponds to the pooling of data in one set of models. Overall, microphone selection achieves the best results, further reducing the error rates when using the B&K and Telephone without increasing the error rate when using the Sennheiser. Since performance with microphone selection is comparable to the baseline (S, S), (B, B) and (T, T) conditions. this technique comes the closest to achieving microphone robustness, at least

in the context of these experiments.

6.3 Summary

Experiments with training techniques are compared and analyzed to understand their effects on the speech recognition system. Multi-style training averages performance across different conditions. In comparison to preprocessing, this results in improvements under the severely mismatched conditions but comes at the cost of degradations under other conditions. On the other hand, microphone selection results in significant improvements under all mismatched conditions without degrading matched conditions. With the availability of microphone specific data, this training technique enables the system to achieve large increases in microphone robustness.

Chapter 7

Conclusion

7.1 Summary

This thesis seeks to improve our understanding of the effects of microphone variations and compensation techniques on the speech recognition system. A methodology is designed to enable the isolation of microphone effects and the benchmarking and comparison of techniques. The tasks of phonetic classification and recognition are studied in order to reduce the effects of confounding task, corpus and system dependent variables. The TIMIT [10] corpus and SUMMIT [36] system are configured for classification and recognition experiments on microphone variations. The Sennheiser, B&K and Telephone recordings of the commonly accepted TIMIT acoustic-phonetic corpus are found to be particularly useful for comparative studies in microphone robustness.

The microphones and data are analyzed in order to understand the effects of microphone variations on the recorded signal. The deviations between the B&K and Sennheiser are relatively small, with an increase in energy at low frequencies due to differences between the non-gradient boom-mounted far-field and gradient close-talking noise-canceling microphones. The deviations between the Telephone and Sennheiser are larger, with a lack of energy at high frequencies and other variations within the Telephone bandwidth due to transmission effects and signal normalization.

Baseline experiments are performed in phonetic classification and recognition for all microphone conditions before and after downsampling. Downsampling reduces error rates for mismatched conditions involving the Telephone by effectively bandlimiting the Sennheiser and B&K to match the transmission bandwidth of the Telephone. Analysis focuses on the realistic mismatched conditions when the system is trained on the Sennheiser and tested on the B&K or Telephone. Mismatched testing on the B&K causes moderate performance degradations that can be explained by the low frequency deviations. Mismatched testing on the Telephone, even after downsampling, causes severe degradations that are more difficult to analyze due to the higher levels of distortion in the Telephone.

Towards increasing microphone robustness, the thesis focuses on preprocessing techniques that compensate for microphone effects on the recorded signal prior to recognition. Several preprocessing techniques that do not require specific-microphone data are implemented and developed for comparison and analysis. Of the techniques, simple Mean Normalization [23] is found to effectively compensate for most of the low frequency deviations and significantly reduce performance degradations for the B&K. In comparison, the complex Codeword-Dependent Cepstral Normalization [1] is found to more effectively compensate for some of the larger deviations and slightly reduce degradations for the Telephone. Overall, preprocessing can increase microphone robustness to small mismatches between training and testing conditions but cannot maintain low error rates under severely mismatched conditions.

The thesis also explores training techniques that directly account for microphone variations in training rather than compensating for effects prior to recognition. These techniques incorporate microphone-specific data in training to reduce microphone mismatches and further increase microphone robustness, especially for Telephone.

7.2 Future Work

Future work includes experiments in word recognition to verify that improvements gained at the phonetic level generalize to the word level. In addition, techniques

can be improved to achieve greater microphone robustness. Preprocessing techniques may be able to take advantage of microphone-specific data in order to more effectively compensate for severe microphone mismatches. Training techniques may be able to use less than full sets of microphone-specific training data and produce more robust models. The remaining parts of the speech recognition system can also be investigated in the context of microphone robustness. More robust segmentation and search algorithms may reduce the large error rates in recognition. Other representations and features may be more robust than the cepstral coefficients, which are sensitive to microphone effects such as transmission bandlimiting and noise [29]. For example, duration is microphone-invariant and can potentially reduce errors, such as confusions between voiced and unvoiced phonemes. Representations based on auditory models [32] have also been shown to be more robust to additive noise. Overall, more integrated strategies may enable speech recognition systems to achieve microphone robustness.

Appendix A

More on Preprocessing Techniques

This appendix contains more experimental results for various preprocessing techniques. Table A.1 shows an example error rate table that is used to present classification and recognition error rates in percent before and after downsampling under various microphone conditions for each preprocessing technique. The rows and columns respectively show training and testing microphones. The top three and fourth rows respectively show results at 16 and 8 kHz. The left and right halves respectively show results in classification and recognition.

	Classification			Recognition		
	S	B	T	S	B	T
S	(S, S, 16)	(S, B, 16)	(S, T, 16)	(S, S, 16)	(S, B, 16)	(S, T, 16)
B	(B, S, 16)	(B, B, 16)	(B, T, 16)	(B, S, 16)	(B, B, 16)	(B, T, 16)
T	(T, S, 16)	(T, B, 16)	(T, T, 16)	(T, S, 16)	(T, B, 16)	(T, T, 16)
S	(T, S, 8)	(T, B, 8)	(T, T, 8)	(T, S, 8)	(T, B, 8)	(T, T, 8)

Table A.1: Example error rate table

Table A.2 shows error rates in percent for MN [23].

	Classification			Recognition		
	S	B	T	S	B	T
S	31.3	34.6	61.7	51.1	54.8	77.0
B	32.8	32.6	56.8	51.3	53.1	75.0
T	79.5	75.9	40.6	84.6	83.8	60.0
S	34.0	36.3	52.5	56.0	58.4	71.0

Table A.2: Error rates in percent for MN

Table A.3 shows error rates in percent for RASTA [14].

	Classification			Recognition		
	S	B	T	S	B	T
S	32.8	35.9	63.5	52.8	56.2	77.7
B	34.3	33.6	59.9	54.3	55.4	76.4
T	81.4	76.7	41.8	81.2	80.3	60.8
S	35.2	38.4	55.6	57.6	60.1	71.1

Table A.3: Error rates in percent for RASTA

Table A.4 shows error rates in percent for BCMN.

	Classification			Recognition		
	S	B	T	S	B	T
S	32.0	35.5	66.5	53.0	58.8	79.2
B	34.2	33.3	67.1	55.1	54.3	80.5
T	81.4	81.8	42.4	83.1	84.1	61.2
S	34.5	37.2	54.5	57.3	59.1	74.9

Table A.4: Error rates in percent for BCMN

Table A.5 shows error rates in percent for SUB [33].

	Classification			Recognition		
	S	B	T	S	B	T
S	31.9	39.5	65.9	50.9	58.1	77.4
B	35.9	33.4	68.4	55.6	54.0	79.7
T	75.3	73.7	42.0	80.5	80.4	60.9
S	34.2	42.4	55.1	55.5	62.8	74.2

Table A.5: Error rates in percent for SUB

Table A.6 shows error rates in percent for SSUB.

	Classification			Recognition		
	S	B	T	S	B	T
S	31.9	39.1	66.0	51.4	58.3	78.2
B	35.8	33.1	67.5	55.1	53.5	79.7
T	75.4	73.5	41.3	82.1	82.8	60.4
S	33.7	41.8	54.0	55.1	62.0	73.9

Table A.6: Error rates in percent for SSUB

Table A.7 shows error rates in percent for the cascade of SUB and MN.

	Classification			Recognition		
	S	B	T	S	B	T
S	31.9	34.6	59.9	51.7	55.3	76.9
B	33.4	33.2	58.0	51.9	53.9	75.8
T	76.8	73.6	41.2	87.5	86.4	59.9
S	34.5	37.5	51.7	55.7	59.2	73.2

Table A.7: Error rates in percent for SUBMN

Table A.8 shows error rates in percent for CDCN [1].

	Classification			Recognition		
	S	B	T	S	B	T
S	31.7	34.2	58.4	51.1	54.6	70.9
B	33.6	32.7	59.6	51.4	54.0	73.0
T	63.2	63.2	40.8	74.5	75.2	59.2
S	35.4	37.9	55.2	55.7	58.5	67.7

Table A.8: Error rates in percent for CDCN

Bibliography

[1] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition,* Ph. D. Thesis, CMU, 1990.

[2] A. Acero and R. Stern. "Towards environment-independent spoken language systems," *Proc. DARPA Speech and Natural Language Workshop*, 157–162, 1990.

[3] T. Anastasakos, F. Kubala, J. Makhoul and R. Schwartz. "Adaptation to new microphones using tied-mixture normalization," *Proc. ICASSP* I:433–435, 1994.

[4] L. Beranek. *Acoustics*. Acoustical Society of America, 1954.

[5] B. Chigier. "Phonetic classification on wide-band and telephone quality speech," *Proc. DARPA Speech and Natural Language Workshop*, 291–295, 1992.

[6] S. Das, A. Nadas, D. Nahamoo and M. Picheny. "Adaptation techniques for ambience and microphone compensation in the IBM Tangora speech recognition system," *Proc. ICASSP*, I:21–24, 1994.

[7] Y. Ephraim. "A minimum mean square error approach for speech enhancement," *Proc. ICASSP*, 829–832, 1990.

[8] A. Erell and M. Weintraub. "Estimation using log-spectral-distance criterion for noise-robust speech recognition," *Proc. ICASSP*, 853–876, 1990.

[9] W. Fisher, G. Doddington, and K. Goudie-Marshall. "The DARPA speech recognition database: specification and status," *Proc. DARPA Speech Recognition Workshop*, 93–99, 1986.

[10] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett and N. Dahlgren. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM.* National Institute of Standards and Technology, 1990.

[11] L. Gillick and S. Cox. "Some statistical issues in the comparison of speech recognition algorithms," *Proc. ICASSP*, 532–535, 1989.

[12] O. Ghitza. "Auditory neural feedback as a basis for speech processing," *Proc. ICASSP*, 91–94, 1988.

[13] J. Glass. *Finding acoustic regularities in speech: applications to phonetic recognition,* Ph. D. Thesis, MIT, 1988.

[14] H. Hermansky, N. Morgan, A. Bayya and P. Kohn. "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," *Proc. Eurospeech*, 1367–1370, 1991.

[15] H. Hermansky, N. Morgan and H. Hirsch. "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *Proc. ICAASP*, II:83–86, 1993.

[16] H. Hirsch, P. Meyer and H. Ruehl. "Improved speech recognition using high-pass filtering of subband envelopes," *Proc. Eurospeech*, 413–416, 1991.

[17] M. Hunt and C. Lefebvre. "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *Proc. ICAASP*, 262–265, 1989.

[18] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz. "N-TIMIT:a phonetically balanced, continuous speech, telephone bandwidth speech database," *Proc. ICASSP*, 109–112, 1990.

[19] B. Juang, L. Rabiner and J. Wilpon. "On the use of bandpass liftering in speech recognition," *IEEE Trans. ASSP*, Vol. 35, Jul. 1987.

[20] L. Lamel, R. Kassel and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, 100–109, 1986.

[21] J. Lim, ed., *Speech Enhancement*, Prentice-Hall, 1983.

[22] R. Lippmann, E. Martin and D. Paul, "Multi-style training for robust isolated-word speech recognition," *Proc. ICASSP*, 17.4.1–4, 1987.

[23] F. Liu, R. Stern, A. Acero and P. Moreno. "Environment normalization for robust speech recognition using direct cepstral comparisons," *Proc. ICASSP*, II:61–64, 1994.

[24] P. Melmerstein and S. Davis. "Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences," *IEEE Trans. ASSP*, Vol. 28, Aug. 1980.

[25] H. Meng. *The Use of Distinctive Features for Automatic Speech Recognition,* M. S. Thesis, MIT, 1991.

[26] P. Moreno and R. Stern. "Sources of degradation of speech recognition in the telephone network". *Proc. ICASSP*, I:109–112, 1994.

[27] H. Murveit, M. Butzberger and M. Weintraub. "Reduced channel dependence for speech recognition," *Proc. DARPA Speech and Natural Language Workshop*, 280–284, 1992.

[28] A. Nadas, D. Nahamoo and M. Picheny. "Adaptive labeling: normalization of speech by adaptive transformations based on vector quantization," *Proc. ICASSP*, 521–524, 1988.

[29] J. Openshaw and J. Mason. "On the limitations of cepstral features in noise," *Proc. ICASSP*, II:49–52, 1994.

[30] D. Pallett. "Benchmark tests for DARPA resource management performance evaluations," *Proc. ICASSP*, 536–539, 1989.

[31] M. Phillips and V. Zue. "Automatic discovery of acoustic measurements for phonetic classification," *Proc. ICSLP*, 795–798, 1992.

[32] S. Seneff. *Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model,* Ph. D. Thesis, MIT, 1985.

[33] D. Van Compernolle. "Increased noise immuninity in large vocabulary speech recognition with the aid of spectral subtraction," *Proc. ICASSP*, 1143–1146, 1987.

[34] D. Van Compernolle. "Spectral estimation using a log-distance error criterion applied to speech recognition," *Proc. ICASSP*, 258–261, 1989.

[35] K. Yu. *A Study of Microphone and Signal Representation Effects on Speech Recognition,* B. S. Thesis, MIT, 1992.

[36] V. Zue, J. Glass, M. Phillips and S. Seneff. "Acoustic segmentation and phonetic classification in the SUMMIT system," *Proc. ICASSP*, 389–392, 1989.

[37] V. Zue, J. Glass, D. Goodine, M. Phillips and S. Seneff. "The SUMMIT speech recognition system: phonological modelling and lexical access," *Proc. ICASSP*, 49–52, 1990.