Speech Recognition In Adverse Environments: Improvements To IMELDA

by

David R. Starks, B.A.Sc.

A thesis submitted to the

School of Graduate Studies and Research

in partial fulfillment of the requirements

for the degree of

Master of Applied Science

Ottawa-Carleton Institute for Electrical Engineering

Department of Electrical Engineering

Faculty of Engineering

University of Ottawa

January, 1995

ISBN 0-612-15765-2

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Abstract

This thesis deals with speech recognition in adverse environments. The primary problem is the mismatch between training and test conditions. Cases of mismatch include the recording channel, acoustic noise and the speaker. Noise shaping and subband filtering are two noise suppression techniques that work by utilizing properties of speech. Dynamic speech analysis and some feature extraction methods are inherently robust to the influence of noise. Linear discriminant analysis (LDA) can be used to combine disparate sets of speech parameters and obtain the optimal set of features. IMELDA [45] is one such method.

In this thesis, we analyse the effectiveness of IMELDA under various training and test scenarios. Theoretical results are first derived and substantiated by simulations. It will be shown that LDA provides a form of noise shaping and the root-deconvolution technique is inappropriate for IMELDA. A new algorithm for predicting recognition performance is proposed and verified. Optimal cross-condition recognition is obtained by utilizing samples of noisy test speech in the within-class covariance, in the so-called QNT IMELDA transform. In the event that the noise is stationary, and can be modelled, we derive an equivalent transform by artificially modifying quiet speech samples. This suffices for the simplest instances. For the extreme helicopter case, we show the best approach to be a combination of band-pass filtering and dynamic analysis of the Mel-scale subbands. Unknown channel noise and additive noise are reduced through the respective subband processing algorithms. Finally, practical issues of applying LDA and integrating subband filtering in a speech recognition system are addressed.

ACKNOWLEDGMENTS

I thank my thesis advisor, Dr. Tyseer Aboulnasr, for her valuable guidance on the signal processing aspects reported here, and for the time she spent reviewing the thesis.

My thanks are also directed to Canadian Marconi Company, (CMC) in particular, Mr. John Thomas, product manager in the Aerospace business unit, for his continued support of my graduate studies. The National Research Council of Canada, (NRC), CMC and the Neil Squire Foundation (NSF) have long collaborated on research and development for speech recognition and I would like to acknowledge their cooperation. Dr. S.R.M Sinclair, head of the NRC IAR Flight Research Laboratory and Mr. Carl P. Swail deserve my gratitude for their long-term support of the NRC/CMC/NSF Speech Group and for providing the computer resources that were used in part for this thesis. I also thank Farouk Laichi, from the Speech Group, for his feedback on the thesis, and for the time he spent preparing training and testing material that was used in the research of Section 3.2 of this thesis.

Finally, I would like to deeply thank my wife, Firouzeh Maleki for her constant encouragement and spiritual support throughout my studies.

SUMMARY

Automatic speech recognition systems have undergone extensive research for the past twenty years while the transition to real world applications has accelerated in the last three years. The average consumer is able to use operator-less telephone applications and voice-assisted personal computers. As speech recognition technology moves out of the laboratory and into the field, the influence of the environment on performance has become a major concern. Performance levels of most current speech recognisers drops significantly when unpredictable noise is present during use. The term *adverse environment* is used to describe the unknown, mismatched and often severe differences in environment between training and testing. The goal is to have the recogniser operate as if it were trained and tested under the same conditions.

Noise can take the form of: i) distortion and filtering in the recording channel; ii) additive noise as coloured or white noise; iii) speech distortion or *Lombard effect* occurs when a talker is forced to shout because of noise in their ear. Variability within and between conditions is a chief concern. Different microphones used between training and testing is another example of variability.

Areas where speech recognition is currently being applied in noisy environments include: telephone networks, cellular phones in automobiles, and aerospace for both helicopters and fixed wing aircraft. Table I.1 summarises the noise sources found in these environments and illustrates the variability in noise characteristics.

Environment	Noise Sources	Characteristics
office [50]	i) disks, fans, printer ii) conversation	i) 45-70 dBA ii) noise spectrum falls-off with higher frequency
telephone [50]	i) channel ii) non-linear carbon microphone	i) 10 dB attenuation ii) attenuation increases with higher frequency ii) wide range of variation in attenuation
car [50]	i) engine ii) cooling fan iii) wind iv) tire and road	i) SNR can be lower than -5dB with hands-free microphone ii) noise level increases with velocity iii) noise spectrum falls off 90 dB to 45 dB between 500 Hz - 1 kHz, 45 dB flat 1 kHz - 4 kHz
helicopter	i) rotor blade ii) rotational machinery iii) wind & ground flap iv) power supply harmonics v) vibrations	i) 5 Hz spectrum modulation ii) noise floor increases with velocity iii) 80-95 dBA iv) pink noise throughout spectrum v) body resonances
fighter cockpit [50]	i) wind ii) oxygen mask iii) g-force	i) 90 dB ii) noise spectrum falls-off with higher frequency

Table I.1: Examples of noisy environments

Speech recognition is particularly attractive in high workload aerospace applications and this is the motivation for the research being conducted at the NRC IAR Flight Research Laboratory. When Dr. M.J. Hunt used *LDA* to *I*ntegrate the static and dynamic outputs of a *Mel*-scale filter-bank (*IMELDA*), it was found that the features were noise-resistant [45]. Since this time (1989), the IMELDA speech recognition system has been successfully deployed in the NRC's research helicopter. Best performance is obtained when the system is trained to the individual user (i.e. Speaker Dependent mode) with speech data collected in the helicopter. A practical concern is the high cost of flight time incurred while gathering speaker data in-situ. Speaker Independent (SI) systems offer

the advantage of not having to train the system to the individual user's voice but at the expense of lower performance (most notably in noise). Under laboratory conditions with artificially generated noise, SI performance can range from 98% in 15 dB SNR white noise, to a low of 80% for channel distortions, with the cross-conditional average being 92.6%. When apriori knowledge of the test condition is used in the LDA, in a process we call QNT, the average rates improve to 98.7%. Refer to Table I.2.

IMELDA Recognition Method	SI Recognition Rate (%)			
	Quiet	15 dB SNR	Channel Noise	Average
standard	99.3	98.7	80.0	92.6
QNT	99.7	98.0	98.3	98.7
subband filtered	98.7	98.0	94.3	97.0

Table I.2: Summary of SI recognition rates for IMELDA. Standard and QNT methods were derived prior to this thesis. Subband filtering is better than the standard method for channel noise. Optimal rates are obtained through apriori knowledge as in the QNT method.

SI systems for the helicopter have not been considered to date because of the expected low performance and because of the requirement for many speakers recording in-situ speech data.

Although the performance of the IMELDA speech recognition system was high, it was suspected that it could be improved and this was the aim of the research described here. The objective of this thesis was to conduct research that would:

1) Improve the cross-condition performance.

2) Eliminate the necessity for apriori knowledge of the test environment.

3) Develop analytical measures to compare methodologies and therefore reduce the amount of testing time.

4) Eliminate the in-situ training requirement for the helicopter.

5) Assess SI recognition in the helicopter.

6) Reduce the consequences of between-speaker differences on SI recognition.

In Chapter 1 the reader is introduced to the basic speech analysis issues which are required for an understanding of how a speech recognition system works. Sections on speech signal processing, LDA and pattern classification are given particular attention because of their relevance to the body of the thesis. Chapter 2 provides a background theory of the speech signal in noise, and the problems that environmental changes impose on speech recognition. As part of the background research, a thorough literature search into methods for robust speech recognition was conducted. Chapter 2 summarises this investigation and directs the reader's attention to those methods that appear to be the most promising. Four themes were selected for a more intensive evaluation and these are covered in the remainder of the thesis:

i) **Noise shaping**: Noise and speech statistics can be used to weight regions of the spectrum to mitigate the influence of noise. Root-deconvolution is investigated in Chapter 2 and shown to be inappropriate for IMELDA. In Chapter 3 we show that the noise masking algorithm used in IMELDA is a form of noise shaping. Of interest

to speech researchers is our examination, on a subband level of the within-class and between-class covariance used in LDA. From this, we can depict those frequency bands that do not contribute to speech discrimination. Our conclusion is that LDA produces output eigenvectors with low weighting for these subbands. Therefore, LDA is a form of noise shaping.

ii) **Dynamic feature analysis:** This widely used method of representing dynamic features of speech is known to be resistant to noise. It achieves this by integrating current and past speech samples over a long interval. The resultant features are uncorrelated to noise. In our examination of IMELDA we show that dynamic features help us with goals 1, 2 and 6.

iii) **QNT IMELDA Transform:** LDA is able to construct noise resistant speech features when noisy speech samples are used in the within-class covariance. Although this method was previously described by Hunt, a thorough analysis on the subband level had never been conducted (to the best of our knowledge). Chapter 3 is devoted to this deliberation and it is believed to be the most in-depth study of the noise resistant properties of IMELDA taken to date. In Table I.2 we see that the QNT transform produces the best cross-conditional result (goal 1), provided we can obtain the noisy specimens. In this thesis we outline an algorithm that is based on the speech recording model where noise is added to normal quiet speech. This satisfies goal number 4. Objective number 3 was reached by comparing separability metrics with

a new algorithm that is based on dissimilarity scores.

iv) **Subband filtering**: The frequency characteristics of communication channels are often fixed or at least slowly varying in time. The rate of change of the human vocal tract is typically much higher than that of perceptually irrelevant components. With these two principles in mind, researchers have shown that filtering (high-pass or band-pass) spectral subband envelopes can reduce the effects of noise and speaker variability [38], [36]. In Chapter 4 of the thesis we set up filters on the subbands of the Mel-scale filterbank. LDA is used to integrate the subband and dynamic acoustic representations. Subband filtering resolves goals 1 and 2 according to Table I.2. SI recognition experiments in the helicopter represent a gross mismatch between training and test. Subband filtering shines in such cases. Recognition rates increased from 79.2% to 92.1%. Goals 4 and 5 were achieved in this study. Subband filtering is also reported to help with goal number 6 but we were not able to substantiate this.

Finally, in Chapter 5, we include some conclusions and discussions about this work. Some suggestions for further research in the analysis of subband filtering and LDA are provided.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Glossary

Acronyms and Definitions

AM	Amplitude Modulated. Multiplication in the time domain of a low frequency information bearing signal by a higher frequency signal (carrier).
B	Between-Class covariance. Referring to the covariance matrix of inter-class reference patterns.
$\mathbf{B}_t$	Vector representing a linear distortion or tilt.
$\mathbf{B}_n$	Vector representing addittive noise component.
BPF	Band-Pass Filtering.
CDCN	Codeword-Dependent Cepstral Normalization.
CEPSTRUM	Speech feature representation that is generated by a cosine transformation of the log spectrum.
CR	Cruise flight mode in the NRC helicopter. Cruising at 60-80 knots.
dB	Decibel. Unit for measuring ratio of two powers or intensities on the logarithmic scale (base 10).
dBa	Adjusted decibel. A unit used to show relationship between interfering effects of a noise frequency, or band of noise frequencies and a reference noise power level of -85 dBm.
DFT	Discrete Fourier Transform.
DLCE	Delta Log Channel Energy. A speech representation that expresses dynamic changes in speech. It is computed by taking a linear difference

	from several past samples of LCE. It is also called the first time-derivative or regression feature.
DTW	Dynamic Time Warping. A standard pattern matching technique for speech that is based on principles of dynamic programming.
FFT	Fast Fourier Transform.
FIR	Finite Impulse Response. A non-recursive filter implementation.
F-ratio	Ratio of **B** to **W** variance in the LDA transform space.
GQ	Ground Quiet test condition in the NRC helicopter. Inside the hangar with engines off and power systems on.
HMM	Hidden Markov Models. A standard stochastic pattern matching technique for speech.
HPF	High Pass Filtering.
HV	Hover flight condition in the NRC helicopter. Hovering less than 5m above the ground.
IIR	Infinite Impulse Response. A recursive filter implementation.
IMELDA	Integrated Mel-scale filterbank representations with Linear Discriminant Analysis. An acronym for our speech recognition system.
I70	Idle flight condition in the NRC helicopter. Engines idling at 70% main rotor RPM.
LCE	Log Channel Energy. The 20 channel (filters) output of a Mel-scale filterbank after a logarithmic deconvolution. This speech representation is considered to be static since it does not explicitly express dynamic changes

in speech.

LDA — Linear Discriminant Analysis. Pattern classification technique used in multi-variate statistical analysis. It reduces the number of dimensions of the analysis space by selecting a subspace which contains the most important information for discriminating between classes.

LMS — Least Means Square adaptive filtering algorithm.

Lombard — The Lombard effect refers to the change in vocal effort (eg. shouting) which occurs when a talker is speaking in a noisy environment.

LPC — Linear Predictive Coding. An all-pole model of the vocal tract, this is one of the most common forms of speech analysis and synthesis.

LPF — Low Pass Filtering.

Mel — Mel-scale of frequency. A technical scale that is based on human perception of hearing. The scale is linear below 1 kHz and logarithmic above.

NRC — The National Research Council of Canada.

NSS — Nonlinear Spectral Subtraction.

PCA — Principal Components Analysis is the process of de-correlating a set of features by applying the algebraic technique of diagonalization.

PLP — Perceptually weighted Linear Prediction. A perceptually motivated speech analysis technique that combines a Bark-based representation with spectral root (1/3) compression and predictive analysis.

PSD — Power Spectral Density is the power of a waveform as a function of

frequency. This can be measured by using a filterbank and taking the average power at the output of each filter.

Q	We use the term Q to describe the standard covariance matrix which is composed of quiet data only.
QNT	We use the term QNT to describe the process of applying degradations to the covariance matrix.
RASTA	RelAtive SpecTrAl. An acronym for an approach to sub-band filtering that uses band-pass type filters.
RPM	Revolutions Per Minute.
SD	Speaker Dependent. Mode of operation for a speech recognition system where the reference models of the system are based on speech from the speaker under test.
SI	Speaker Independent. Mode of operation for a speech recognition system where the test speaker has not contributed their speech to the reference models of the system.
SMC	Short-term Modified Coherence. Method of increasing SNR of a speech signal by utilizing the coherence of adjacent segments.
SNR	Signal to Noise Ratio.
SPA	'A' weighted sound level expressed in dB. It is the ratio of this sound pressure to a reference noise pressure.
SPL	Linear weighted sound level expressed in dB. It is the ratio of this sound pressure to a reference noise pressure.

W	Within-Class covariance. Referring to the covariance matrix of intra-class samples.

Conventions and Notations

I	Identity matrix.
F_0	Fundamental frequency, or pitch of a speaker.
J_1	Separability measure called Fishers Linear Discriminant.
J_4	Separability measure. Ratio of $\mathbf{B}$[1] to **W**.
ln	Natural log term.
Lv	Number of speech samples.
N	Vector size or filter order.
M	Number of elements in acoustic vector.
tr(**W**)	Represents the trace of matrix **W**.
<.>	Denotes the time average of the sample function enclosed within.
⊛	Denotes the convolution operation.
≐	Denotes an approximation.
|**W**|	Determinant of matrix **W**.
$\mathbf{W}^{-1}$	The inverse of a square matrix **W**.
$\mathbf{W}^t$	The transpose of matrix **W**.
σ_j	Variance of parameter j.

---

[1] Boldface letter denotes a vector or matrix.

ρ_{kj}	Correlation between parameters k and j.
Γ	Diagonal matrix of standard deviations.
P	Matrix of correlation coefficients
T_a	Time interval to analyse a single glottal cycle.
T_d	Time interval for dynamic analysis of speech spectrum.
T_s	Input sampling period of the speech signal.
T_w	Time interval for windowing the input speech signal.
z	Denotes the operator *z* of the Z-transform.
1-pass	time-alignment with cosine transform
2-pass	time-alignment with transform result from 1st pass

CHAPTER 1

SPEECH ANALYSIS FOR RECOGNITION PURPOSES

1.1 Introduction

The architecture of speech recognition systems comprises two processing stages: the *front-end* and *back-end* components. Figure 1.1 illustrates this in block diagram form. The acoustic features of the speech signal are extracted in the front-end component. The extraction process consists of two phases. First, a spectral analysis is performed, then a transformation, is applied with the purpose of retaining the relevant information for speech recognition. The derived spectral information is subsequently passed on to the back-end component which performs the comparison of the extracted features for recognition. Popular front-ends for speech recognition are Linear Predictive Coding (LPC) and filterbanks while the most common back-ends are Hidden Markov Models (HMM) and Dynamic Time Warping (DTW). While this thesis concentrates on one particular combination, an FFT-based filterbank followed by a linear discriminant transformation and DTW pattern matching, the concepts of subband filtering and LDA can be applied to both types of front-end and back-end.

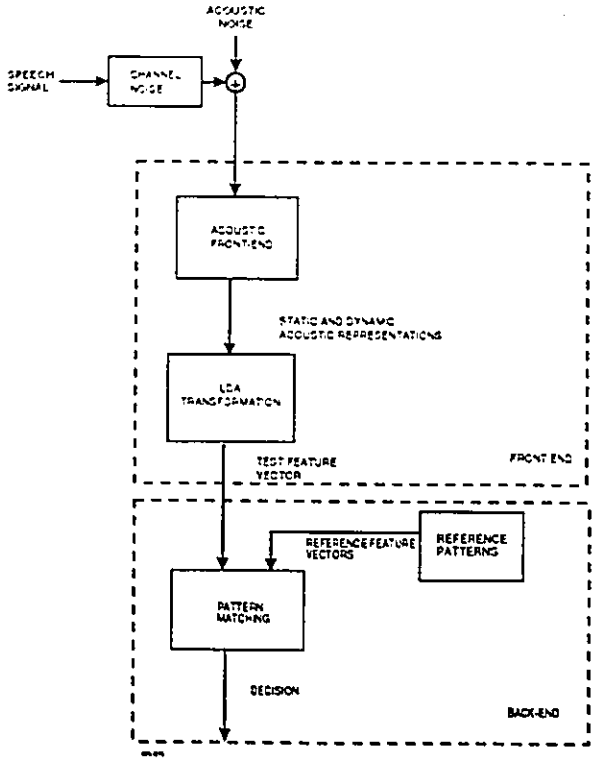


Figure 1.1: Block diagram of a speech recognition system

1.2 The speech signal

1.2.1 Speech generation

The organs of speech are composed of three subsystems:

i) **Lungs and trachea**: The lungs supply compressed air to the system via the trachea and these organs control the loudness of the resulting speech.

ii) **Larynx**: The larynx contains and controls the primary sound-generating mechanism, the vocal cords.

iii) **Vocal tract**: This modulates the resulting sound.

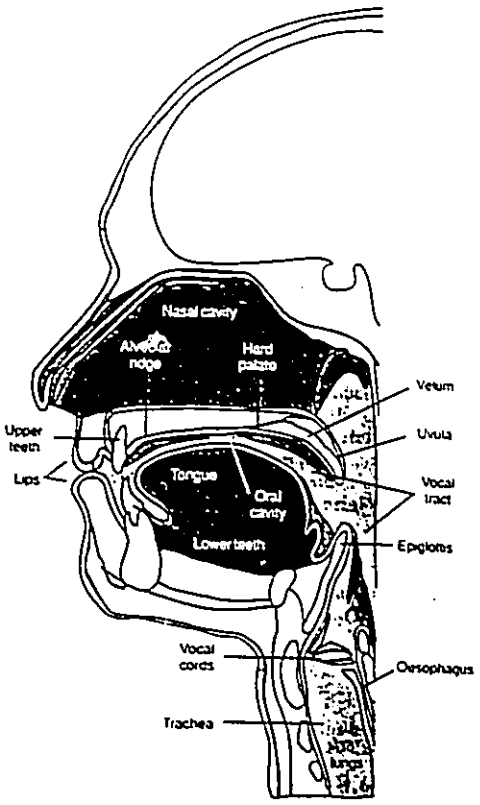


Figure 1.2: The principal organs of speech (from [40])

Referring to Figure 1.2 the vocal tract is everything past the vocal cords, while the space between the vocal cords is the glottis. The function of the vocal tract is to colour and articulate the voice, it achieves this by movement of the tongue, lips and lower jaw. Radiation of the speech is also controlled by elements of the vocal tract. The vocal tract in the adult male is approximately 17 cm long. The frequency content of the speech signal is affected by cavity resonances in the vocal tract. As the signal passes through the vocal tract the changing shapes gives rise to differing resonances. The shape of the oral cavity and oral pharynx is altered as the speaker moves the tongue. By raising the velum, the nasal cavity can be sealed off from the pharynx.

1.2.2 Speech production and perception

Speech production can be broken down into two areas:

i) **Excitation**: The glottis controls excitation and this consists of phonation, whispering, frication, compression and vibration.

ii) **Modulation**: The vocal tract modulates the speech signal and radiates the speech out of the mouth and this is how the speaker imparts information on the glottal output.

Phonation is the oscillation of the vocal cords and this is the most important exictation source. Forcing air through the vocal cords causes them to vibrate and opening and closing these cords separates the airstream into pulses. The repetition rate of the pulses is called *pitch* and this is controlled by tension on the vocal cords while regulation is achieved through feedback via the ears and brain. Pitch is not always regular nor does it need to have a 50% duty cycle, in fact the pitch can double or be halved in certain cases. *Voiced* speech are those sounds associated with phonation. Voiced sounds include the vowels and some consonants, such as, 'w', 'l', 'm' and 'n'.

Whispering is produced by drawing the vocal cords together but maintaining a small triangular shaped opening to force the air through. The resulting excitation signal is wideband noise and the sound produced are called *consonants* (e.g. 'c', 'g').

Frication occurs when the vocal tract is constricted at any other point and the air flow past the constriction is turbulent resulting in a broadband noise excitation. Sounds

produced this way are called *fricatives* or *sibilants* (e.g. 's' and 'f'). When phonation and frication occur at the same time the sound is termed a voiced fricative, examples are 'z' and 'v'.

Compression takes place when the vocal tract is completely closed while the speaker continues to try and exhale, consequently internal pressure builds up and when the vocal tract is finally reopened a small eruption occurs. This process is referred to as a release and if it is abrupt and clean the sound is a *stop* or *plosive* (e.g. 'p' or 't'), otherwise it is an *affricate*.

Vibrations are set up when air is forced through the tongue rather than the vocal cords. An example is the French trilled 'r' sound.

The sound is modulated by moving the tongue, lips and jaw in order to change the quality of the voice. The speech signal is filtered by the vocal tract and it is this filtering that is often modelled in speech recognition systems. The glottal waveform contains many harmonics and when this waveform is modified by the shape of the vocal tract which itself has many natural frequencies the result is natural resonances which are called *formants*. Formants account for all vowels and some consonants and are the most important acoustical characteristic of the vocal tract.

1.2.3 Acoustic phonetics

Fig 1.3 compares the long-time averaged Power Spectral Density (PSD) for ensembles of male speakers and female speakers with that of an indivdual male speaker. The low-pass characteristic above 500 Hz indicates the high occurance of low frequency

speech sounds such as sustained vowels. The high-frequency sounds are lower in power but they provide information about unvoiced consonants, fricatives and stop sounds. The average SNR for voiced speech is about 34 dB and for unvoiced speech 14 dB.

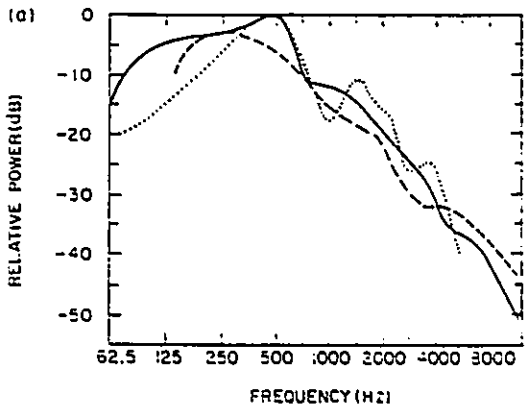


Figure 1.3: Long-time averaged PSD. Solid line is composite of males (6), dashed line for females (5) and dotted line for a male speaker (from [49]).

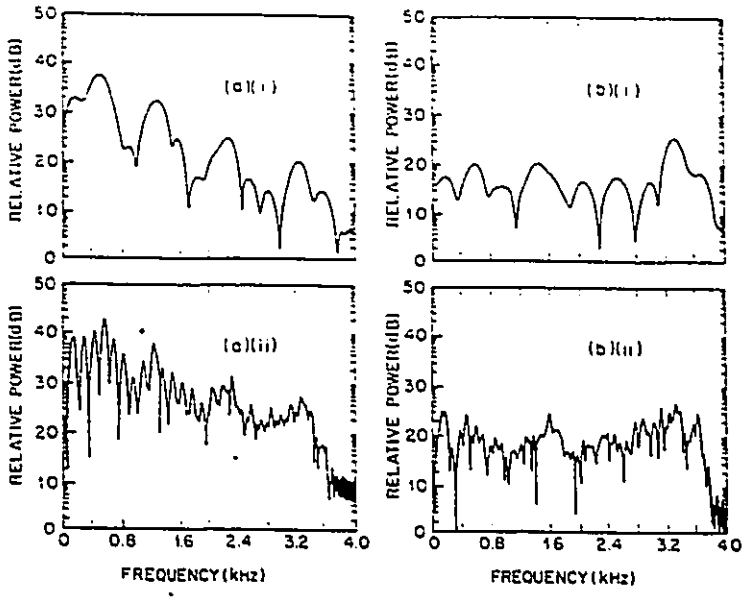


Figure 1.4: Comparison of short-time spectrum for voiced (left) and unvoiced speech (right). Low (64) and high resolution (256) PSD shown top to bottom (from [49]).

Figure 1.4 illustrates short-time averaged spectral plots for a) voiced ("A") and b) unvoiced ("S") speech sounds. The voiced speech sound is relatively low-pass like the long-term PSD in Figure 1.3 whereas the unvoiced sound is somewhat highpass (i.e. there is more high frequency content). Low (64-point FFT) and high resolution (256-point FFT) in frequency PSDs are compared in parts (i) and (ii) of Figure 1.4. The glottal excitation, depicted by the rapid oscillations in the voiced speech PSD, is especially apparent in the higher resolution FFT. The voiced speech contains four distinct peaks, these are the formants. Formants span the region of 0 to 4 kHz and they decrease in amplitude with increasing frequency. This decrease in amplitude comes from a combination of the impulse response of the vocal tract, radiation from the lips and the excitation. The net result is a -6 dB/octave slope to the spectrum.

Pre-emphasising the speech signal compensates for the falloff at high frequncies. The form of pre-emphasis is a single-zero high-pass filter (1.1). The zero location, α is chosen to be in the range of 0.9 to 1.0. The consequence of attenuation for frequencies below 200 Hz can be large depending on the zero but this is irrelevant because the speech signal is usually filtered to remove these frequencies during the signal acquisition stage.

$$y(n)=x(n)-\alpha x(n-1) \tag{1.1}$$

The -6dB/octave falloff applies to voiced speech and unvoiced speech tends to have a flat spectrum at high frequencies (as per Figure 1.4). Ideally pre-emphasis would only be applied to voiced speech and so to limit the outcome on unvoiced speech, α is chosen to be 0.9 [71].

Formants are identified by number in order of increasing frequency: F_1, F_2, F_3 etc and pitch is sometimes referred to as fundamental frequency, F_0. In determining the bandwidth of a speech recognition system at least three formants must be covered. Table 1.1 depicts the frequency range for the first three formants. These are typical values only and F_1 can lie below F_0.

	Frequency Ranges (Hz)		Bandwidths (Hz)
	Adult males	Adult females	
F_1	200-800	250-1000	40-70
F_2	600-2800	700-3300	50-90
F_3	1300-3400	1500-4000	60-180

Table 1.1: Frequency range of the formants

1.3 Effect of speaker differences

The major sources of variation between the voices of different speakers are physiological differences and distinctions in usage for both the vocal tract and the larynx. The difference in physiology of the vocal tract and larynx is most evident between men and women. The male larynx is larger than that of the female and so the average fundamental frequency of 100 Hz (range 80 to 160 Hz) is about half that of a female (range of 160 to 400). Children have much smaller larynx and so the pitch frequency is even higher. The vocal tract resonances in female speech are approximately 15% higher in frequency because a woman's vocal tract is typically 15% shorter than a man's [43]. If the fundamental frequency is raised the separation between harmonics becomes greater. The peaks of the spectral envelopes will also be higher if the vocal tract resonances are higher. The long-term PSD of Figure 1.3 also illustrates the different location of formant regions.

Another difference is the dynamic range of the speech signal. Variance in average short-term PSD is measured to be about 20 dB across speakers [49]. The rate at which we speak exhibits a high degree of variability. Word speaking rates for normal speech can vary from 118 to 193 words per minute [75].

1.4 Spectral representation of speech

One of the major properties of the speech signal that is utilized in speech analysis is the relative stationarity of the signal during the short interval between glottal cycles.

It is assumed that the relevant speech parameters are fixed during this time. Thus, short-time windows of speech can be analysed and the relevant properties can be extracted.

Most analysis techniques perform some form of averaging over the time window and in order to model dynamic parameters the speech signal is divided into succesive windows called analysis windows. The duration of the analysis window is subject to several temporal aspects of speech. During slow speech, the vocal tract shape and the glottal excitation may not change for up to 200 ms, however in most speech sounds the durations are about 80 ms. Rapid events such as stop releases dictate windows of 5 to 10 ms in order to avoid averaging the burst spectra with an adjacent sound.

In speech recognition applications it is desirable to analyse a single glottal cycle at a time because two or more cycles could lead to harmonics of the fundamental being captured. The analysis interval, T_a is bounded by (1.2).

$$\frac{1}{F_o} < Ta < \frac{2}{F_o} \qquad (1.2)$$

The choice of T_a is a compromise between the 5 ms (200 Hz F_0) suitable for females and 10 ms (100 Hz F_0) for males. Ideally, the analysis interval would be synchronised to the glottal excitation but this is not always practical.

In order to accurately represent the slowly varying physiological aspects of vocal tract movements it is necessary to average adjacent analysis intervals. The averaging takes place over a period of time, called the analysis window, T_W. The length of the window is dictated by:

i) Short enough to encompass the stationarity so that significant changes do not take place.

ii) Long enough to calculate the desired parameters and if additive noise is present to average out the random noise component [71].

iii) Long enough so that short speech sounds are not missed as the window is advanced periodically.

The set of speech samples in the interval is shaped by multiplying the speech with the window function. The window length is typically chosen to be twice the analysis interval, thus allowing a 50% overlap to satisfy condition iii). The shape and duration of the window are balanced in order to satisfy both conditions i) and ii). The window shape is also chosen to reduce the effect of spectral distortion when truncating an infinite sequence prior to spectral analysis. The Hanning type window, or Raised Cosine Window is most often used in speech analysis because the edges of this window function are tapered and allows for pitch period boundaries or sudden speech changes like voiced onsets. For a given window shape the duration of the window is inversely proportional to its spectral bandwidth and so the choice of duration is a tradeoff between frequency and time resolution. Most speech analysis algorithms use a fixed window size of 5 to 10 ms with a duration of 10 to 25 ms. This enables individual harmonics to be resolved for fundamental frequencies as low as 40 Hz. Time behaviour is averaged over a few pitch periods. Windows longer than this may smooth out the rapid spectral changes that are important to speech recognition.

Modulation frequency is a term used to describe the temporal fluctuation of energy in frequency bands. Only modulation frequencies of up to 25 Hz exist in speech signals, with the long-term maximum being located at about 3 Hz. For speech perception, the relevant parts of the modulation frequency spectrum are between 10 and 25 Hz, while speaker specific information dominates for frequencies below 10 Hz [39]. Table 1.2 summarises the three analysis rates that are required in speech recognition.

	Sample Rate	Requirements
T_s	8 kHz	Minimum case is to cover the first 3 to 4 formants (Table 1.1)
T_a	100 Hz	Analysis of a single glottal cycle (1.2). Should be between 100 Hz for males and 200 Hz for female talkers.
T_w	40 Hz	Slow variation of vocal tract. Duration of 10 to 25 ms (100 Hz to 40 Hz). The band from 3 Hz to 25 Hz is perceptually relevant.

Table 1.2: Typical analysis rates used in speech recognition

1.5 Functional description of a speech recognition system

This section provides a detailed description of the block diagram of Figure 1.1. The frequency range of the human auditory system for adults is about 16Hz to 16kHz. A 4 kHz bandwidth encompasses the first three to four formants and the beginning of fricatives and so most speech sounds can be understood. Standard systems use a sample rate of 8 kHz and band-pass the signal for a 300 Hz to 3400 Hz range. The reason for band-pass filtering is because speech is highly correlated below 300 Hz and so high-pass filtering this component reduces the overall variance and hence the number

of encoding bits. Even though the high-pass filter removes the pitch frequency (and power supply hum) the listener is still able to perceive the pitch and identify the speaker.

The dynamic range of a talker can exceed 40 dB over a short interval (one second) and sustain a 30 to 40 dB range [49]. Typical speech recognition systems will use a 12 bits linear resolution encoder for a dynamic range of about 72 dB.

1.5.1 Acoustic representations

The following description is for one of the most common forms of spectral analysis used in speech recognition, the Mel-scale FFT-based filterbank. Figure 1.5 depicts the acoustic front-end used in IMELDA. The principles of filterbank based speech analysis are described in [54] and several signal processing options are compared in [16].

Spectral analysis is performed on the windowed signal with an FFT. The frequency distributed power (we call spectra) is then computed by squaring and summing pairs of real and imaginary components. This is then normalized to the sampling rate. The phase component is ignored because it is not relevant to speech recognition. The spectra are then weighted by overlapping triangular shaped filters that are distributed along the technical Mel-scale of frequencies. Refer to [72] for a more complete description of this scale. This distribution is linear below 1 kHz and logarithmic thereafter. Table 1.3 depicts the frequency spacing for 20 filters used in IMELDA. It is well-known that the ear's critical band-widths vary along this perceptual scale and capture the phonetically important characteristics of speech. The minimum bandwidth of the bandpass filters is 200 Hz and this allows at least one harmonic of the fundamental

frequency to be included in each filter. This reduces the effect of pitch differences between speakers (recall the range is 100 to 200 Hz). Davis and Mermelstein, [17] first described this filterbank and showed that the Mel-scale distribution produced a better suppression of insignificant spectral variation in the higher frequency bands than a linear spaced filterbank. The first three formants typically lie in the lower frequency, higher resolution filters while voicing onsets and voiceless sounds are covered by the higher frequency, lower resolution filters. The shape of the filter is not as important as the spacing. Rectangular, triangular, trapezoidal and other filter shapes have been investigated and neither showed any particular advantage in speech recognition. Allthough it may be argued that filter shape can be optimised to reduce correlation, speech researchers typically use the triangular shape.

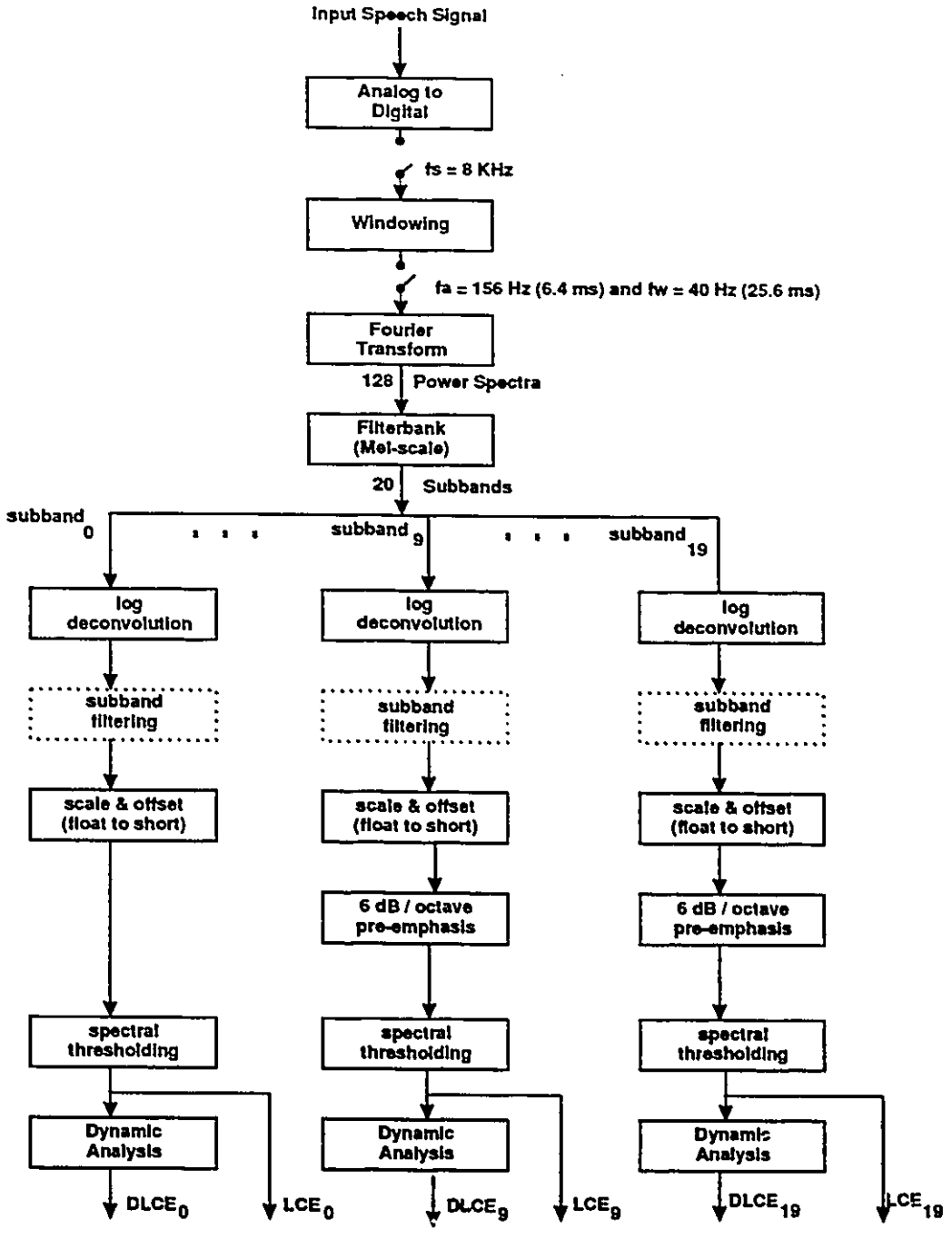
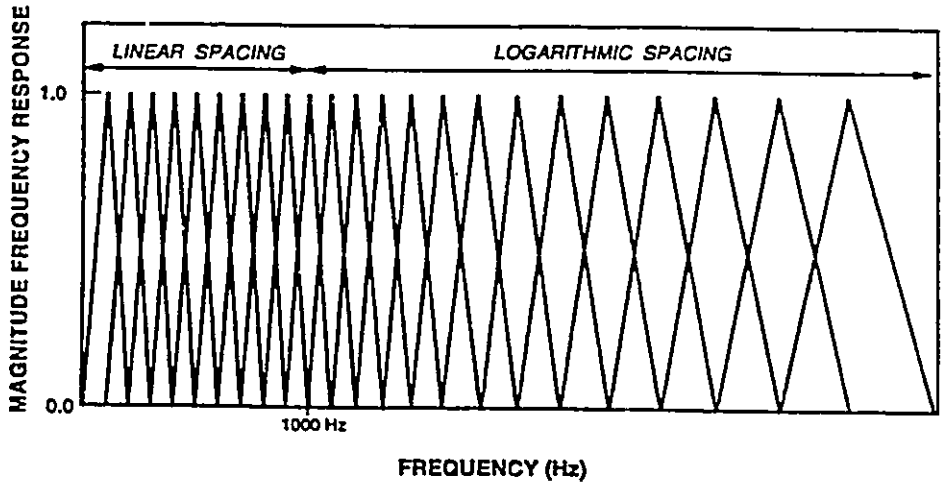


Figure 1.5: Flow diagram of the acoustic front-end used in IMELDA



Subband (i)	Low Freq	Centre Freq	Upper Freq
0	0	100	200
1..8	i*100	(i+1)*100	(i+2)*100
9	900	1000	1150
10..18	$1000*1.15^{i-10}$	$1000*1.15^{i-9}$	$1000*1.15^{i-8}$
19	3500	4000	4000

Table 1.3: Mel-scale filterbank characteristics.

The resolution of the spectral analysis is determined by the sampling rate and length of the FFT. Most researchers use a 256 point FFT when using an 8 kHz sampling rate, for a frequency resolution of 31 Hz per FFT output every 25.6 ms. All subsequent processing in the front-end is with the individual filter outputs. The filters are also referred to as *subbands*. The next stage of processing is a non-linear transformation. The log

function is appropriate for speech and the basis for this is that higher value signals do not contain as much perceptual information as the lower value signals. Thus, the information bearing portion of the speech signal is expressed in the linear region of the logarithmic curve and the 'louder' part is compressed. The log domain is also less sensitive to alterations in absolute energy levels. In Chapter 2 we will see that some researchers have been investigating alternative functions such as the root operator (1.3) where, r is less than 1.0. One of the fundamental aspects of subband filtering is the selection of the linear or log domain for the filters since the former is suitable for removing additive noise components and the latter for reducing convolutional noise (e.g. channel distortion).

$$y = x^r \tag{1.3}$$

In order to reduce variance in the spectral representations caused by low level speech and non-speech sounds the speech signal level can be truncated by setting a lower limit on the energy values, this is called spectral thresholding and was first described by Klatt [54]. In this way, low energy portions that are influenced by random noise will be set to the same value for both test and reference vectors and thus not contribute to the accumulated difference. The operation is simple, if the input energy value falls below a pre-determined value called the threshold, then the input value is set to the threshold value, this is performed on both test and reference.

Both instantaneous and spectral transitions play an important role in speech perception. The logarithmic representation of the filterbank is considered to be the

instantaneous or *static* acoustic speech representation. We call this LCE. Spectral transitions or *dynamic* features, DLCE can be portrayed by applying regression analysis over a defined time interval, T_D, of static features. The dynamic representation was first introduced by Furui [24]. Furui determined that the optimal time interval for preserving transitional information associated with changes from one speech sound to another was in the range of 56 ms to 72 ms. Using m to represent the subband (total M), the linear regression coefficients (i.e. the first-order orthogonal polynomial) are given in (1.4).

$$DLCE_m(t) = \frac{\sum_{n=-\frac{T_D-1}{2}}^{n=\frac{T_D-1}{2}} n * LCE_m(n)}{\sum_{n=-\frac{T_D-1}{2}}^{n=\frac{T_D-1}{2}} n^2} \qquad (1.4)$$

The combination of dynamic and static acoustic representations has proven to be very constructive in mapping vowels and consonants and for discriminating between confusable words [24]. A third acoustic representation, the 2nd order regression, is currently receiving much attention because it helps discrimination in noise [33]. The number of acoustic parameters can be in the order of 3*M and so much effort has been made into reducing the parameters and for finding the optimal combination. Since

adjacent filterbank outputs tend to be correlated, a compact, uncorrelated representation of the spectral envelope is taken by applying a linear transformation. The resulting outputs, are now in multi-dimensional feature space which is more efficient and seemly for pattern matching. This is described in the following section.

1.5.2 Pattern Classification

1.5.2.1 Introduction

This section provides an introduction into the theory of pattern classification and focuses on those aspects that are important to speech recognition. In pattern classification the unknown or *test* pattern is identified by comparing it with a set of previously characterized or *reference* patterns. Each pattern is composed of a group of vectors where each vector contains a set of features. The features are measurable characteristics of the class that has been found useful for identification. In word recognition the vectors are stored time sequentially over the length of the word. This type of pattern is a template.

A speech recognition system operates by processing the sounds of the test word into its acoustic representations, extracts features from these and forms the test pattern. The system has a library of reference patterns, one for each word in the vocabulary. All reference patterns are compared with the unknown and the word with the reference pattern providing the best match is considered the identity of the unknown. The system has a learning phase (also called *training*) for building the library of reference patterns.

During training, data for known classes are applied to the system, which then averages together all examples for each class producing the reference library. Word patterns may be represented by sub-word units (e.g. speech sounds like syllables or phonemes) or in their entirety as whole-words. Therefore, the remainder of this section will use the more general term of "class" when referring to word patterns.

In the following two sections, the subjects of decision rules and feature selection are addressed. The problem of time-aligning the test pattern with the reference patterns is described in the section on dynamic programming.

1.5.2.2 Similarity and Distance Measures

In speech recognition, features are normally considered with groups of vectors or frames. For the sake of simplicity this section does not consider timing considerations. Each frame of speech is viewed as spanning a multidimensional feature space. The pattern vectors are usually assumed to have multivariate gaussian densities. Features are not always gaussian but there is generally insufficient data to prove otherwise and this assumption also allows the use of well-understood mathematics. If the densities associated with different classes overlap then there will be recognition errors. Therefore, statistical methods are used in selecting a set of features that minimise overlap and for choosing the most appropriate decision rule.

The probability that an unknown feature vector **x** represents class *i* is P[*i*|**x**]. The learning process provides the statistics of the features for each class and from this the probability that word *i* has given rise to the features **x**, P[**x**|*i*]. Bayes theorem relates the

conditional probabilities and with the assumption that all classes are equally probable it can be shown that the rule is, for any unknown feature vector **x**, to choose the class that maximises P[**x**|*i*]. Assuming that the pattern vectors for each class are multivariate-gaussian, then the probability density function for each class is given in (1.5).

$$P_i(\boldsymbol{x}) = (2\pi)^{-\frac{M}{2}} |\boldsymbol{W}_i|^{-\frac{1}{2}} \exp\left(\frac{(\boldsymbol{x}-\mu_i)^T \boldsymbol{W}_i^{-1}(\boldsymbol{x}-\mu_i)}{2}\right) \qquad (1.5)$$

The probability density function for class *i* is fully described by a mean vector μ and a covariance matrix $\mathbf{W}^{-1}$. Most of the features used in speech recognition have unimodal distributions resembling gaussian and so this is a reasonable assumption. In the equation above, $\mathbf{W}_i$ is the covariance matrix for class i and $|\mathbf{W}_i|$ is its determinant while μ_i is the mean vector for class i. The covariance matrix (1.6) has the following properties:

i) The covariance is estimated by pooling all samples of class i, $\mathbf{f}_i$ and removing the mean for the class.

$$\boldsymbol{W}_i = \langle (\boldsymbol{f}_i - \mu_i)(\boldsymbol{f}_i - \mu_i)^T \rangle \qquad (1.6)$$

ii) There are M acoustic parameters per sample. The variance of the jth parameter, σ_j^2 is situated on the diagonal of the matrix. The covariance between two

parameters, k and j is $\rho_{kj}\sigma_k\sigma_j$ where ρ_{kj} is the correlation which is contained in the off-diagonal terms of the matrix.

iii) The covariance matrix can also be expressed as the product of the diagonal matrix of standard deviations, Γ_i and the matrix of correlation coefficients, P_i. This matrix is symmetric.

$$\boldsymbol{W}_i = \Gamma_i P_i \Gamma_i \qquad (1.7)$$

It is difficult to obtain accurate estimates for the covariance matrix of each word from limited training data. Furthermore, the matrices are similar for all classes, hence, most recognition systems use a single **W**. A significant memory saving is realised when only a single matrix is used.

For a given test vector **x**, class i is selected if (1.8) is satisfied.

$$P_i(\boldsymbol{x}) > P_j(\boldsymbol{x}) \qquad \text{for all classes } j \neq i \qquad (1.8)$$

For computational simplicity a logarithmic transformation can be applied and terms that are constant across classes can be eliminated (e.g. the common $|\mathbf{W}|$). The result is the basic maximum-likelihood criterion, where we seek to minimise the distance between the test and the reference (1.9).

$$D_i(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{W}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln |\boldsymbol{W}_i| \qquad (1.9)$$

The similarity between the unknown and the reference is inversely proportional to their distance of separation in the feature space. A distance measure or *metric*, d(**x**,**y**) between test pattern **x** and reference pattern **y** is required. To classify the patterns two approaches can be taken: i) The feature space is partitioned into regions, one for each class, and boundaries are defined by hyperplanes and the distance is computed between the planes. ii) The simpler alternative is to compute the distance from the unknown input to the centroid of each class ($\boldsymbol{\mu}_i$ represented here as **y**) and select the class for which the distance is minimum.

The maximum likelihood criterion is not often used in its fullest form (1.9) because of the difficulty in obtaining reliable estimates of the covariance matrices. Simplifications of the general form lead to the following distance metrics. The covariance-weighted distance, or *Mahalanobis* distance sets $\mathbf{W}_i$ to be the autocovariance matrix corresponding to the reference vector with the underlining assumption that this is the same for all classes. The natural log term, ln |**W**| is then constant and can be ignored (1.10).

$$d(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{W}^{-1} (\boldsymbol{x} - \boldsymbol{y}) \qquad (1.10)$$

By assuming that the features are uncorrelated, the off-diagonal terms of the covariance matrix are thus zero and the matrix reduces to the terms along the diagonal. These terms are simply the variance of the reference features. Combining this assumption with the approach that the variance is the same for all classes results in the *normalized Euclidean* distance (1.11). The variance of the *M* features is σ_j^2.

$$d(x, y) = (x - y)^T \, diag\,(\sigma_1^2, \sigma_2^2, \ldots)(x - y) = \sum_{j=1}^{M} \left(\frac{x_j - y_j}{\sigma_j}\right)^2 \qquad (1.11)$$

By further assuming that all of the features have equal variances (i.e. **W** is the identity matrix **I**) the result is the *Euclidean* distance (1.12). The Euclidean distance is optimal however only if the M features are uncorrelated (mutually independent) and contribute equally to the distance measure. This is an important consideration in speech recognition systems and is the subject of Chapter 3. A diagonal covariance matrix can be obtained through a linear transformation which preserves the normal distribution and employs a coordinate system where the features are independent. Linear transformations are described in detail in Section 1.5.2.5.

$$d(x, y) = \sum_{j=1}^{M} (x_j - y_j)^2 \qquad (1.12)$$

The Euclidean metric is the most common distance measure. The tradeoff with this metric is accuracy for reduced computational complexity (M multiplies as compared to M^2 for the Mahalanobis metric).

1.5.2.3 Feature Selection

Not all of the information contained in the acoustic parameters is relevant in recognition of speech. The process of selecting those parameters or combination of parameters that contribute to the recognition process is described in this section. Feature

selection results in a set of reduced data, and this is beneficial in terms of storage and computational load. The features should have the following properties (from [72]):

i) Vary widely from class to class.

ii) Not be correlated with the other features.

iii) Have similar values for different renditions of the same pattern.

iv) Insensitive to external variables such as ambient noise, stress or fatigue in the speaker's voice.

v) Stable over long periods of time.

vi) Easy to measure.

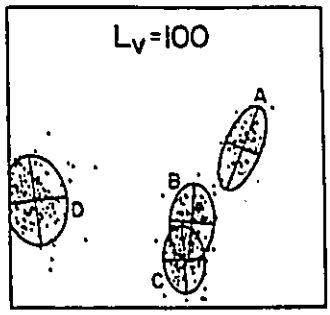


Figure 1.6: Scatter diagram showing within and between class dispersion for 4 classes of 100 samples (Lv).

Figure 1.6 illustrates the dispersion within four classes: A, B, C and D and the spread between the classes on a 2 dimensional scatter diagram. The horizontal and vertical coordinate space represent two acoustic parameters. Samples of the acoustic

parameters are shown as dots while the surrounding elipse indicates the distribution about the mean. In this particular example, classes B and C overlap and thus cannot be discriminated.

The covariance matrix for a set of five acoustic parameters (subbands 10 through 14) derived from a Mel-scale filterbank is plotted in Figure 1.7. The correlation between parameters (off-diagonal terms) is evident in the overlap as well as the dual peaks.

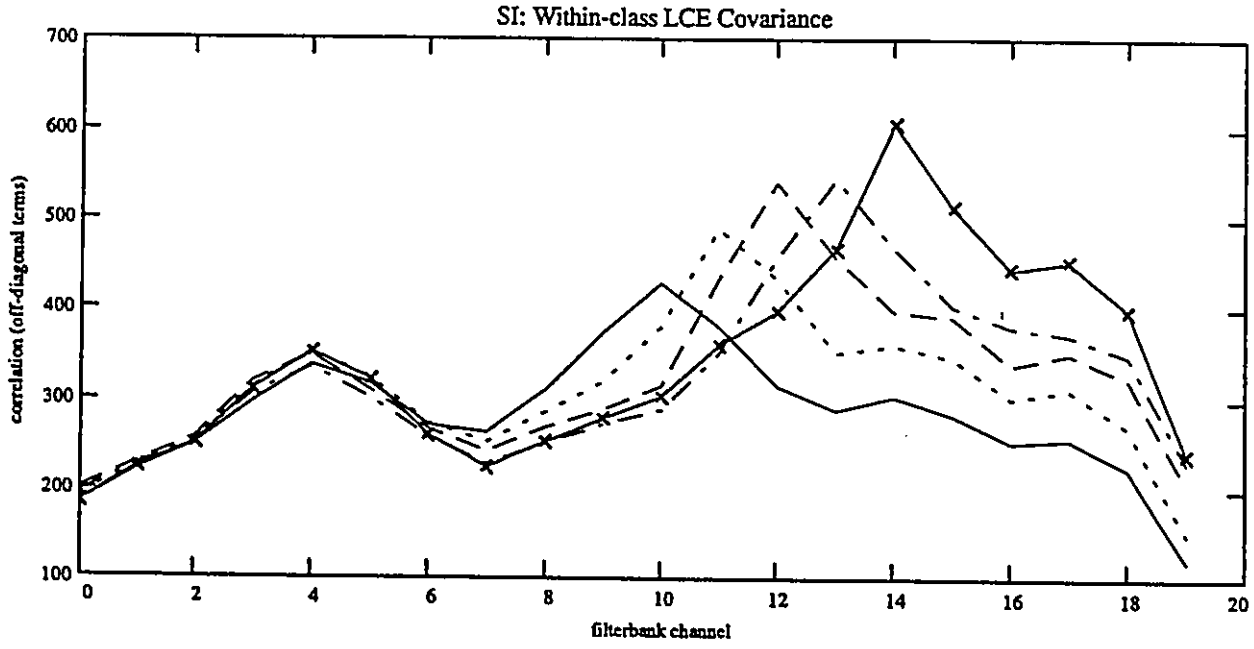


Figure 1.7: Five acoustic parameters extracted from Covariance Matrix **W**

Repetitions of the same speech class should have consistent feature measurements and thus be tightly clustered (i.e. small dispersion) in the feature space.

Different classes should provide very distinct measurements and a corresponding wide separation in the feature space. Hence the best features exhibit little within class variance and large between class variance.

In the Mahalanobis distance (1.10), the covariance matrix $\mathbf{W}^{-1}$ has the between class variance reflected along the main diagonal [71]. Features that are poor in discriminating classes are represented by small values in the matrix so that they contribute less to the overall distance. If the features are orthogonal, only the diagonal is considered (1.11) and if the features are orthonormal then $\mathbf{W}^{-1}$ reduces to I (1.12). A Principal Components Analysis (PCA) is required to transform the acoustic representations in order to obtain orthogonal features. This is described in further detail in Section 1.5.2.5. The transformation from the acoustic parameter representation to a set of orthogonal features is a key component in the pattern matching process. This transformation can be viewed as a de-correlation of the acoustic parameters.

1.5.2.4 Measures of Separability

This section describes methods for evaluating combinations of features to see how well they separate the classes. This is of particular interest in our work since the traditional method is to run time-consuming word recognition experiments and we are seeking a more efficient way of doing this. The basic measurement is called Fisher's linear discriminant. Here the class-to-class separation of a feature over all classes is represented by the ratio of the variance of the means over all classes to the mean of the variances within the classes. For vectors of features, the following measures are used:

$$J_1 = tr(\mathbf{W}^{-1}\mathbf{B}) \tag{1.13}$$

$$J_2 = |(\mathbf{W}^{-1}\mathbf{B})| \tag{1.14}$$

$$J_4 = tr(\mathbf{B})/tr(\mathbf{W}) \tag{1.15}$$

The within-class covariance matrix pooled over all classes is represented by **W** and this describes the variation within the classes. By representing all of the data for a class by its mean the between-class covariance matrix **B** then describes the scatter between classes regardless of the scatter within the class. The trace of a covariance matrix provides a measure of the total variance of the matrix parameters, hence J_4 compares the scatter between classes to the scatter within classes. The larger the ratio the better discriminating power of the features. Measures J_1 and J_2 are invariant under rotations and scaling while J_4 is only invariant under scaling. These measures will be used in Chapter 3 where we compare the expected usefullness of various feature sets with experimental results.

1.5.2.5 Linear transformations

This section deals with modifications to the coordinates of the feature space in order to identify those features or combinations of features which contribute the most to separating the speech classes. The first transformation to be described is PCA, it was derived by Hotelling (1933). This is also called the Karhunen-Loève transform or

diagonalization where the objective is to uncorrelate the features by rotating the cluster of features through some matrix. The relationship between the transformed set of features, **y** and the original correlated set **x** is through matrix **A** (1.16).

$$\mathbf{y} = \mathbf{A}^T\mathbf{x} \tag{1.16}$$

The cosine transform (1.17) is often used to represent a PCA in speech recognition. Its basis vectors are good approximations to the eigenvectors of a Toeplitz matrix resembling the covariance matrices of speech signals.

$$CT_m(n) = \cos\left(\frac{m\pi}{2N}(2n+1)\right) \qquad for\ m=0,1..M-1 \quad for\ n=0,1..N-1 \tag{1.17}$$

Low-order coefficients of this transform are sensitive to gross structure (e.g. the overall intensity) and so the first term (n=0) is dropped. Higher terms are sensitive to fine structure (individual channels) and so these terms are not used. An example of the problem with having too fine a resolution or structure is when harmonics of the fundamental (pitch) are only separated by adjacent channels and the fundamental happens to rise during the test condition, this then differs from the reference and causes a mismatch during recognition. Typically 8 to 12 terms (N) per set of spectral parameters are used (order M). The cosine transform is competent in quiet conditions but it is susceptible to degradations, consequently many researchers have attempted to improve

the robustness of the cosine transform. Cepstral analysis is the term used when referring to the log power spectra that have undergone a cosine transformation. In Section 3.3 we inspect the cosine coefficients.

The general procedure for finding **A** in (1.16) is by solving the determinant equation below. This is described in [72] and [23] and in most algebra text books and is not repeated here. The principle is to find the eigenvector matrix, **A** of the covariance matrix of **x**, $\mathbf{W}_x$ via (1.18).

$$| \mathbf{W}_x - \lambda \mathbf{I} | = 0 \tag{1.18}$$

This is a linear transformation which rotates and scales the axes so that the hyperellipsoid shaped class distributions become hyperspeheres (recall Figure 1.6). A summary of the properties of PCA is provided below (from [14] and [23]):

i) The eigenvector matrix, **A** provides a coordinate transformation.

ii) The eigenvalues are the variances of the transformed variables.

iii) Each component has maximum variance and is statistically uncorrelated and orthogonal to the preceding components.

iv) Much of the variance is contained in the first few principal components.

v) The principal components have zero means.

vi) This is an orthonormal transformation and Euclidean distances are preserved.

Cooley, [14] outlines the importance of using a large enough set of samples for computing **W** so as to minimize the errors in estimating the principal components. This is considered in Chapter 3 where analysis of IMELDA is carried out.

In multi-dimensional feature space the contours of probability density (2D case shown in Figure 1.6) are shown as hyperellipsoids and the features form clouds around these contours. The transformation rotates the coordinates to align directions of maximum variance along the new coordinate axes. The separability measures, J_2 and J_4 compare the size of the clouds of the between and within class covariance.

To achieve the goal of maximising separability between classes, the process of statistical discriminant analysis is used [21]. This technique finds a subspace which separates the class centroids maximally relative to the variation within the classes. This transformation differs from the one described above because the directions of maximum separability rather than maximum variance are aligned with the feature axis.

Discriminant analysis can be summarised as follows (from [72]):

1) Rotate the within class matrix by finding the principal components.
2) Neutralize the distracting influence of the individual cluster shapes by applying a scaling (or *whitening*) transformation to the result from step 1. This turns the ellipsoidal shapes into spheres.
3) Rotate the interclass distribution **B** by applying the results from steps 1 and 2. Diagonalize the transformed **B**.

In practise the whitening step can be eliminated and the other two steps consolidated into one step of simultaneous diagonalization (1.19). This is well described in [23] and is not repeated here. We call this Linear Discriminant Analysis (LDA).

$$| \mathbf{W}^{-1}\mathbf{B} - \lambda \mathbf{I} | = 0 \qquad (1.19)$$

This is similar to the PCA except that the transformed between class matrix $\mathbf{W}^{-1}\mathbf{B}$ is not symmetric and the resulting eigenvector matrix, **A** maximises separability of the classes. In multidimensional space, the Euclidean distance (1.12) is now valid since **A** transforms the input parameters to a set of de-correlated set of features that have equal variance (while maximising separability of the classes). Separability measure J_1 is the optimization criterion used in the diagonalization. The resulting eigenvectors are not mutually orthogonal but are orthogonal with respect to **W** and can be made orthonormal by appropriate scaling.

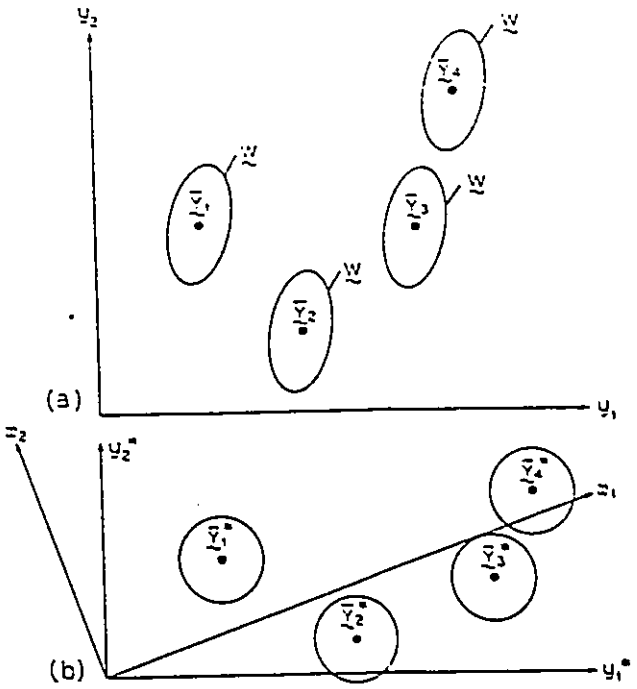


Figure 1.8: Geometrical Interpretation of Discriminant Analysis (from [10])

A geometrical interpretation of the discriminant analysis for the case of two features y_1 and y_2 is shown in Figure 1.8. Centroids of the classes,Y_1 to Y_4 are shown in 1.8a surrounded by an ellipse indicating the distance metric for **W**. Figure 1.8b shows the whitened and rotated ellipses as circles in the reoriented space with axis y_1^* and y_2^*. This is as far as the PCA would go. Rotating this space to maximise separability between the centroids produces new coordinate x_1 with the highest separation and x_2 with the next highest separation. The improved separability is clear. The degree of rotation achieved by the discriminant analysis is influenced by the between class distribution and so this

works more effectively than the PCA which only has knowledge of the within-class distribution. We refer to the new coordinate space as the discriminant space.

The diagonalization performed in the LDA places much of the variance in the first few components. This can be measured by observing the eigenvalues (F-ratios) of **A**. Those vectors with high valued F-ratio can be preserved, subsequently reducing the dimensionality of the feature space. This is of extreme importance to recognition systems where the input acoustic parameters may be of dimension 64 and the transformed features being reduced to as few as eight (e.g. [44]). The computational and memory savings is tremendous. It has been shown by several researchers [19], [8], including our own experience, that the reduced number of features actually improves performance. This has been attributed to statistical noise in the lower F-ratio terms which becomes detrimental to the pattern matching process. Further analysis of LDA and the practical considerations of using spoken words as the classes is provided in Chapter 3.

1.5.2.6 Discriminative training

Discriminant analysis is also applied in a technique called discriminative training where the premise is to increase the separability between pairs of confusing classes. A modified form of the Fisher discrimiant ratio is optimised in order to find a weighting function that best separates the two classes. Regions of the classes that provide the best discrimination are given higher weighting. This approach is receiving a great deal of attention. A tutorial level of detail is available in [50].

1.5.3 Pattern matching through dynamic time warping

For the sake of brevity this section is kept to a minimum. An excellent tutorial on pattern matching and a comparison of HMM and DTW is provided in [69] while algorithm details for DTW are summarised in [79] and for HMM in [73].

The most common method of matching a reference pattern with an unknown speech pattern is that of dynamic programming and is often referred to as Dynamic Time Warping. This algorithm is an efficient means of compensating for the non-linear temporal distortions of spoken speech. The figure below illustrates the temporal differences between the unknown pattern and a reference pattern (spectrogram representations). There are numerous potential ways of aligning the two patterns. The optimal path is shown in the grid. The alignment paths are called *warping* paths and they could conceivably start for all frames in each reference pattern. Consequently, constraints are put in place to reduce the number of potential path candidates.

The difference between the input frame of unknown speech and the frames of reference speech are computed at every frame by using a distance metric (1.12). The current distance is accumulated with the past differences, which is called the *score*. This process continues for each input frame in a sequential fashion as the reference and unknown frames are compared along the warping paths. The greater the score the more distortion and hence less similarity. The current score can be used to reduce the candidate paths if a threshold is exceeded.

One of the most common warping constraints is the symmetric step, here the path is restricted to have come from either the horizontal, diagonal or vertical directions. This

restriction is a 'global constraint' and allows at most a factor of two stretching or compressing. In Figure 1.9 we see that the path portions are diagonal (i-1,T-1) during the best frame match and horizontal (i,T-1) during a time compression of the reference and vertical (i-2,T-1) in order to stretch the template. The lowest of the three possibilities is taken at each frame and added to the current score. This is expressed in (1.21).

Limiting the range over which the warp paths can travel reduces the number of paths and hence the computational cost. This is reflected in the subset of allowed steps (vertical, diagonal, horizontal). Adding penalties to the non-diagonal steps increases the overall cost function of those reference templates that require use of these steps. Penalty weights are denoted as values 'pen_1' and 'pen_2' in (1.21).

$$score[i][T] = dist + \min\begin{pmatrix} score[i][T-1] + pen_1, \\ score[i-1][T-1], \\ score[i-2][T-1] + pen_2 \end{pmatrix} \qquad (1.21)$$

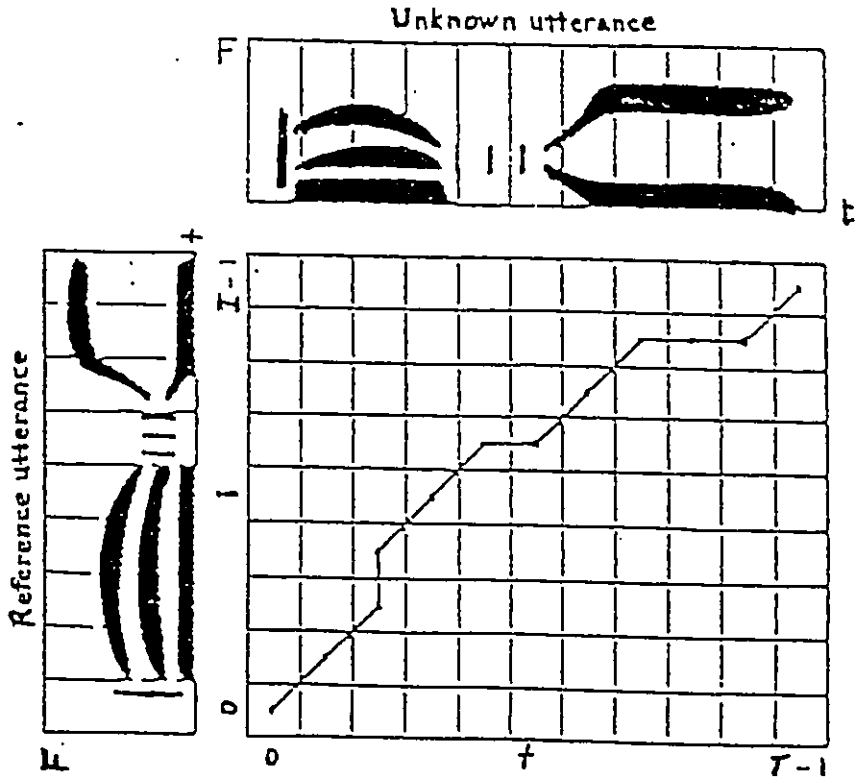


Figure 1.9: DTW Time-alignment between unknown and reference patterns (from [74])

Since the pattern matching is by far the most time consuming aspect of a speech recognition system there is great advantage if the analysis rate can be reduced. Optimal rates are in the range of 10 ms to 32 ms. The DTW algorithm has a high degree of parallellism to it and this has led to many efficient implementations. The author of this thesis has applied systolic array techniques to the DTW algorithm [81] as well as coarse-grain commercially available digital signal processors [83].

A disadvantage with DTW that is addressed by HMM is that it is sensitive to words that are affected by co-articulation (refer to Section 1.2.2). One problem is the mis-classification of words with plosives such as final *t-release* sounds (see Section 1.2.3). For example, 'feet' could be mistaken by 'eight' if the reference pattern for 'feet' contains

the 't' sound while it is not there under test conditions. Co-articulation effects from succeeding words can cause this problem. If the reference pattern for 'eight' does not contain the 't' sound because of the way it was trained then it becomes the closest match. HMM gets around this common problem by modelling sub-word elements and assigning transition probabilities to them. In the case above, the 't' sound would be assigned a low transition probability allowing the word to be spoken with or without this sound. One work-around solution for DTW is to provide multiple templates, with and without the release sounds. This issue arises during the recognition experiments of Chapter 4.

CHAPTER 2

SPEECH RECOGNITION IN ADVERSE CONDITIONS

2.1 Introduction

This chapter provides a theoretical background on the speech signal in noise. The problems that environmental changes impose on speech recognition and the major methods for coping with them are outlined.

Acoustic variability in the speech signal arises from three broad categories: the speaker, his environment and the input equipment. The major acoustic variables that lead to degradation in speech recognition performance are: additive noise, speaker individuality, environment-dependent speaking styles, microphone characteristics and transmission channels. These are summarised in Table 2.1.

Category	Causes
Environment	Speech-correlated noise: i) reverberation
	Uncorrelated noise issues: i) additiivity & stationarity
Speaker	Attributes of speakers: i) dialect ii) gender iii) age
	Manner of speaking: i) breath and lip noise ii) stress iii) Lombard effect iv) rate of delivery v) level & dynamic range vi) pitch vii) cooperativeness
Input Equipment	Microphone
	Distance to microphone
	Filter
	Transmission system: i) channel distortion ii) noise iii) echo
	Recording equipment

Table 2.1: The main causes of speech variation (from [25])

Figure 2.1 illustrates in flow-chart form the methods for dealing with microphone problems, additive noise and speaker variation. These methods are explained in sections 2.3, 2.4 and 2.5 respectively. For further details, the reader is referred to three detailed surveys on speech recognition in adverse conditions, [50], [25], [29].

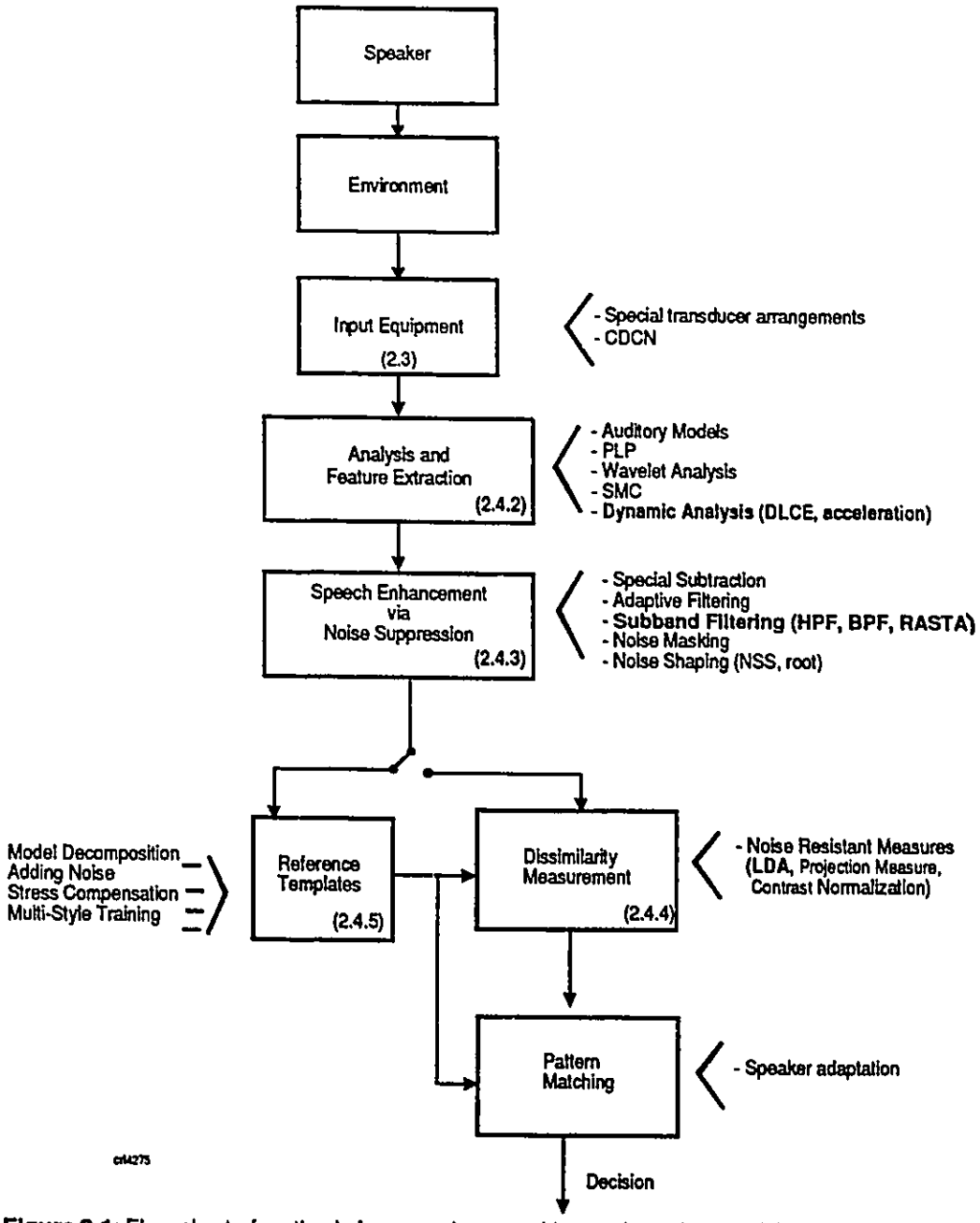


Figure 2.1: Flow chart of methods for speech recognition under noisy conditions (adapted from [25])

The fundamental issue concerning speech recognition in adverse conditions is whether training can be conducted under the same noise conditions as encountered

during test. If this is the case, performance is better than if training occurred solely in noise-free conditions. Conversely, performance is worse in quiet conditions if training took place in noisy environments. The mismatch between a clean reference template and a noisy test utterance can significantly influence recognition rates when the SNR goes below 24 dB [16].

The simplest and the most effective approach is to train the system (build models or templates) under the same conditions as expected in the test environment. This minimises the differences between the reference and test models. Multi-style reference templates can be used to represent the various test conditions. However, this is not always practical, nor safe in mobile test environments like avionics or automobiles. Difficulties in modelling the reference templates can arise in applications where the test conditions are changing. The automobile, for example, undergoes changes in speed, road conditions, positioning of the windows and the air conditioner. Also, the use of different microphones between training and testing adds to acoustic variability. Techniques that automatically adapt to different environments are the best solution.

2.2 Effects of noise on speech representations

Acoustic ambient noise is usually considered to be additive and it can be wideband (assummed to be gaussian distributed and white) or coloured, stationary or non-stationary. Referring back to Table I.1, we see that the sources of additive noise can be found in several environments, including the office, car and the aircraft cockpit. Distortion noise can be found in: i) room acoustics where there is reverberation; ii) microphone

mounting position relative to the sound source; iii) telephone channels. Telephone channels are characterised by a high-frequency pre-emphasis which significantly decreases the dynamic range of the speech signal and increases the signal correlation. There is a wide range of variation in attenuation in the frequency spectrum and this can cause spectral mismatch unless the telephone channel is measured and adapted to beforehand. Figure 2.2 compares the amplitude of frequency responses for two different telephone channels.

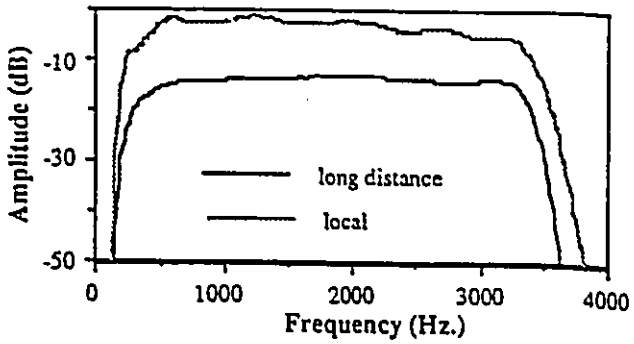


Figure 2.2: Frequency response for two different telephone channels (from [57])

Also, the distance between the microphone and the speaker's mouth can cause a poor SNR and the acoustics of the room or the microphone can differ from the orignal set-up that the reference patterns were created in. Figure 2.3 compares the frequency characteristics of two types of microphone transducers.

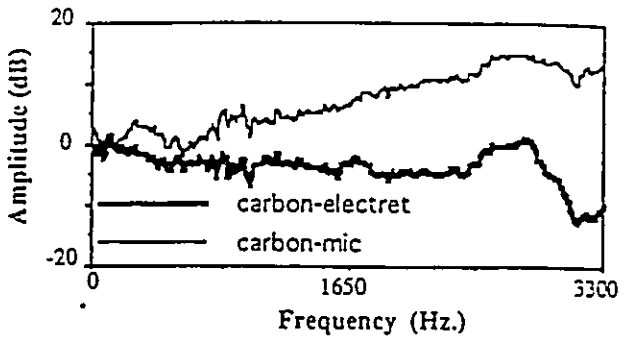


Figure 2.3: Frequency response for two different microphone transducers (from [57])

In loud ambient noise, talkers adjust their speech production, resulting in distortions referred to as the *Lombard effect*. The Lombard effect is manifested in such changes as [51]:

i) Louder speech (increased energy).

ii) Energy migrations in the frequency domain, F_1 increases while F_2 decreases.

iii) Changes in spectral balance (i.e. *tilt*).

Perceptual observations of Lombard speech indicate that it is more intelligible than normal speech when the words are not confusable but less intelligible otherwise.

A talker's articulation of words is also affected by noise [51]. Experimental results indicate that the indirect effects of noise can exert a greater influence on performance than the noise that actually enters the microphone [76]. Hansen has studied the effects of Lombard and stressed speech and has presented morphological based solutions in [32].

The variability of the noisy environment is typified by the helicopter cockpit. Figure 2.4 shows the noise floor PSD for four different conditons found in the NRC flight research helicopter. A brief description of the noise sources is tabulated in Table 2.2. For a description of how a real-time IMELDA based speech recogniser was deployed in this adverse environment the reader is referred to [67] and [82].

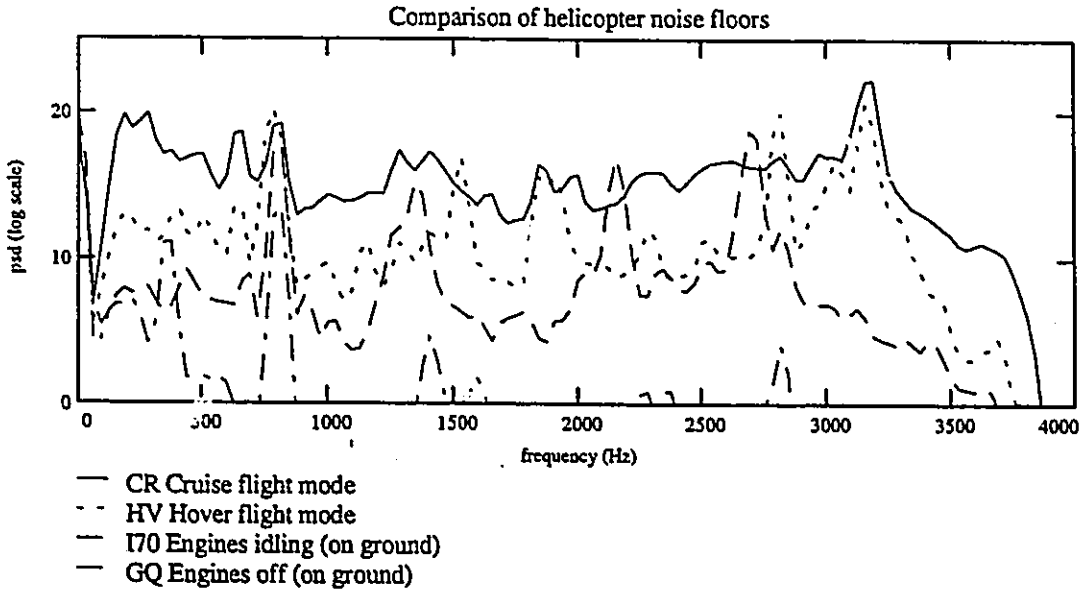


Figure 2.4: Four different acoustic conditions of the helicopter environment

Flight Mode		noise level	SNR	noise description
GQ	ground quiet, inside hangar engines off power systems on		25 dB	i) power-supply harmonics (most noticeable at 800 Hz).
I70	engines idling at 70% main rotor RPM	81 dBA 107 dBL	23 dB	i) power-supply harmonics ii) rotational machinery noise (gas engine turbine, main & tail rotors) frequency location about 70% that of HV and CR. Most noticeable at 500 Hz, 1.3 kHz, 2.1 kHz and 2.7 kHz. iii) body cavity resonances @ 1.5 Hz causing modulation of speech.
HV	hovering < 5m above ground	91 dBA 109 dBL	19 dB	i) power-supply harmonics ii) rotational machinery iii) Lombard effect iv) ground flap noise v) AM noise - rotor blade (5 Hz)
CR	cruising at 60-80 knots	94 dBA 114 dBL	17 dB	i) power-supply harmonics ii) rotational machinery iii) Lombard effect iv) wind noise v) AM noise - rotor blade (5Hz)

Table 2.2: Noise description for 4 different helicopter modes

We define "AM" noise in the helicopter case as frequency-domain periodic impulses which are caused by rotor blade passage. This component occurs every 5 Hz and it is additive. To illustrate this phenomenon, we conducted Fourier analysis on the outputs of a Mel-scale filterbank. Figure 2.5 illustrates the subband modulation (from output of 300 Hz spectra) of the helicopter noise floor during the cruise flight mode. The 5 Hz fundamental of the rotor blade passage is very clear with the 10 Hz harmonic being the strongest component.

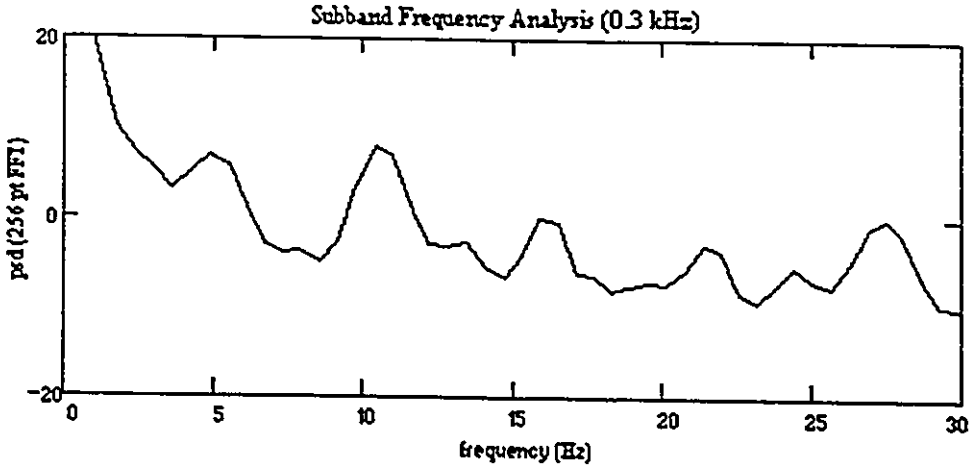


Figure 2.5: Frequency analysis of the 300 Hz subband depicting "AM" noise caused by helicopter rotor blade passage

In following sections, methods to overcome the above problems are described.

2.3 Input equipment - special transducer arrangements

To reduce the influence of low-frequency noise and room reverberation and to prevent the problem of microphone displacement, a head-mounted noise cancelling microphone can be used. These microphones are appropriate in applications where the talker position is fixed (e.g. automobiles and aircraft). The noise cancelling type of microphone is designed specially to cancel far-field sounds and receive only the near-field sound such that the received signal can be almost noise-free. The talker's mouth must be kept parallel and close to the transducer. Speech power can drop by as much as 15 dB with a corresponding performance drop if the microphone is displaced by as little as 10 cm and a 30-degree rotation. This impact has been seen in our helicopter flight trials at the NRC speech laboratory. Multi-sensor arrangements, including speech

beamforming and other types of gradient microphones are described in [50] and [87]. In an effort to locate the optimal mounting position for a fixed microphone in the automobile, Smolders et al conducted a series of experiments. Their conclusion was that a position on the ceiling in front of the speaker was best [80].

2.4 Methods for dealing with environmental noise

Noise compensation is utilized in both the acoustic front-end and the back-end components of the speech recognition process (see Figure 1.1). Figure 2.1 summarises the wide array of strategies that are available. This section outlines the principal strategies for combatting environmental noise.

2.4.1 Compensating for variation from different kinds of microphones

The differences in characteristics of microphones used for training and recognition can have delitorious consequences for recognition rates. Liu et al [59] have shown that it is possible to normalize the cepstral representation of the test utterance using estimates of the input speech. Additive noise and linear filtering distortion are jointly compensated. The algorithm is called CDCN (Codeword-Dependent Cepstral Normalization) and it is more effective than the spectral subtraction and spectral normalization methods (described below). A second advantage of this algorithm is that it adapts to the variation of the acoustical conditions without the need for apriori long-term statistical information

[1]. An alternative method of cepstral subtraction that has been applied to telephony is described in [57].

2.4.2 Analysis and speech feature extraction

The front-end stage has been the subject of considerable investigation into finding analysis techniques that produce speech features which are inherently robust to noise. This section briefly describes some of the more successfull approaches.

2.4.2.1 Auditory models

Models of human hearing are used in speech recognition because they are robust to additive noise and to the indirect results of speech variations. Two examples are [44] and [28]. Psychoacoustic characteristics of the cochlea that are represented in these auditory models include nonuniform frequency resolution, nonlinearities, masking and two-tone suppression (lateral inhibition).

Pereceptual linear predictive (PLP) analysis is a technique that estimates the auditory spectrum by incorporating properties of hearing. The properties are a critical-band Bark-scale spectral resolution (differs from the Mel-scale described above), frequency dependent spacing along the equal loudness curve and intensity-loudness power law. Filter spacing is much broader than the triangular functions typically associated with the Mel-scale and the spectrum is computed by an autoregressive all-pole model [35].

2.4.2.2 Wavelet analysis

Wavelet analysis is being used in place of conventional filterbanks because of the joint time-frequency properties. This approach may result in an optimal solution to the temporal analysis requirements listed in Table 1.2. The Sextant Avionique company in France has recently reported that signal analyis via a wavelet transform can permit a good and robust speech parameterization which is suited to SI recognition [70].

2.4.2.3 Short-time modified coherence analysis

The inherent coherence of adjacent speech segments can be used to increase the speech SNR in a technique called *Short-time Modified Coherence* (SMC) [63]. The approach is to all-pole model the autocorrelation sequence rather than the speech waveform directly. The SNR improvement is data dependent and a 10 to 12 dB improvement can be obtained for input speech signals with SNR's in the 0 to 20 dB range.

2.4.2.4 Dynamic feature analysis

Improvements to recognition rates in both Lombard and additive noise cases can be obtained with temporal dynamic features obtained via both first and second order regression [33]. Long window lengths in the dynamic analysis (1.4) result in integration across speech sounds and this makes the features more resistant (uncorrelated) to noise

and better able to extract relatively long-term spectral changes [33]. Low modulation frequencies are suppressed by the high-pass nature of these dynamic features [27].

The static acoustic feature may be more readily degraded by noise and Lombard effect than the derivative features, and it has been found to be advantageous to exclude this feature in high ambient noise. Considerable effort has been made into developing an algorithm that determines the optimal combination of features, but this remains an elusive goal [15]. In this thesis, we attempt to prove that a robust IMELDA transform can provide the optimal combination for cross condition testing (Chapter 3).

A drawback of the 2nd order regression feature, (also called *acceleration*) is the selection of the window length, because there exists some dependency on the speech corpus. Furthermore, long window lengths may be inappropriate for continuous word recognition systems [47]. This issue of integration across speech sounds arises in our thesis investigation into subband filtering (Chapter 4).

2.4.3 Speech enhancement through noise suppression

This section describes several techniques for enhancing noisy speech. A model of speech in noise is presented first. We will be referring to this model throughout this thesis.

2.4.3.1 Speech recording model

A model of the speech received at an input microphone can be described by:

$$y(k) = h(k) \otimes s(k) + n(k) \qquad (2.1)$$

Speech is represented by s(k), background noise by n(k) and the channel as h(k). This model assumes that the channel can be modelled by a linear filter and that the background noise and speech are additive. Nonlinear distortions such as room reverberation and nonlinearities due to switching in telephone networks, for example, are not included in the model. The Lombard effect is not easily modelled and is thus not shown in the simplified paradigm. The speech signal can be recovered from the recording y(k), if given knowledge of the channel and noise source(s).

In the event that both types of noise are not present simultaneously we can separate the speech from the received signal quite easily. We first assume the noise and speech are uncorrelated and then conduct Fourier analysis. In the first case we ignore the channel noise, reducing (2.1) to (2.2). Since the noise is assumed to be uncorrelated, there are no crossproducts.

$$|Y(f)|^2 = |S(f)|^2 + |N(f)|^2 \qquad (2.2)$$

If we know the magnitude of the additive noise component, it is then possible to remove this from the observed signal by subtraction in the frequency domain. Refer to Section 2.4.3.2 for further details on this approach. Alternatively, the reference model can be modified to reflect the noise level received by the microphone (Section 2.4.5). In the second case we assume n(k) in (2.1) is negligible. Then we can separate the channel

noise from the speech by using the property that the Fourier transform turned the convolution relationship in (2.1) into a multiplicative relationship. A log operation will separate the two (2.3). We call this deconvolution.

$$\log(|Y(f)|) = \log(|H(f)|) + \log(|S(f)|) \tag{2.3}$$

Channel distortion can be removed by subtracting or filtering this component. Section 2.4.3.5 describes this approach.

2.4.3.2 Spectral subtraction

One of the most common methods of handling wideband noise is based on subtracting an estimate of the spectrum of the noise from that of the noisy speech. The pioneering work of Boll [9] has been widely applied in the field of speech recognition and speech synthesis. For a formal review of the theory, the reader is referred to [9]. The key points are summarised below:

i) The acoustic noise is additive and uncorrelated from the speech.

ii) Noise does not have to be stationary but the estimate must be representative of the noise present during the subtraction.

iii) Secondary procedures are required to attenuate the residual noise after the subtraction

iv) Methods are required to compensate for the case where the estimated noise spectra is less than the noisy speech spectra, otherwise, the power is negative valued.

v) An FFT of size 256 provides sufficient spectral resolution.

vi) There is a tradeoff with the time resolution of the analysis and having a sufficient number of windows in the averaging to keep the noise variance at a minimum.

The speech spectrum is estimated by subtracting the estimated noise spectrum from the noisy speech, (2.2). This is referred to as power spectral subtraction (2.4).

$$|S(f)|^2 = |Y(f)|^2 - \langle |N(f)|^2 \rangle \qquad (2.4)$$

When the noise has non-zero variance it is possible that over-subtraction may occur. This results in negative numbers which are not allowed in the succeeding log transformation. To compensate for this, a threshold or an offset is added which is usually just above the noise floor. However, this may mask some useful speech features and we will keep this in mind thoughout this chapter.

2.4.3.3 Adaptive filtering

The conventional method of adaptive noise cancelation using the adaptive filtering algorithm to process two input signals is a natural choice for applications with changing noise conditions. An estimate of the differential path characteristics from a noise source to the primary and reference inputs is made. Two microphone noise cancellation using

the normalized LMS algorithm in a helicopter application is described by Kang et al [52]. Approximately 10 dB of noise suppression was obtained with this method. Figure 2.6 depicts the two microphone system, the principle of operation is summarised below:

i) The transfer characteristics from the reference noise source to the primary microphone (speech plus noise) are approximated by a FIR filter.

ii) The estimated noise from the reference (noise only) microphone is subtracted from the primary microphone.

iii) Noise must be uncorrelated from the speech.

iv) The reference noise must be highly correlated with the noise picked up by the primary microphone.

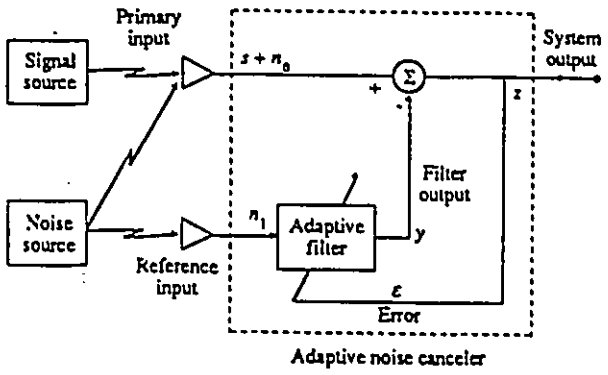


Figure 2.6: Two microphone noise cancellation (from [25])

The author of this thesis developed and evaluated a two microphone cancellation system that was proposed by Hunt and Lefebvre[1]. The novelty of this approach is that the Wiener filtering and cancellation operates on the subbands of a Mel-scale filterbank. This is similar in spirit to DFT based LMS cancellation systems. The chief advantages of a frequency domain approach are:

i) Both wideband and coloured noise components (that may be situated at multiple frequencies) can be easily isolated and removed.

ii) Channel distortions which are additive in the log domain can also be identified and subtracted.

We had found that substantial improvements could be made to a conventional cepstrum based speech recogniser but the IMELDA recogniser which was already quite robust did not gain from this effort[2].

Some of the key practical issues found in applying two-microphone cancellation are:

i) The distance between the microphones is critical. A small distance ensures the reference noise source and primary pickup are correlated but at the same time, sufficient distance is necessary to prevent the speech signal from reaching the

---

[1] The NRC has initiated a patent on this and the work remains unpublished.

[2] Unpublished.

reference microphone. Hence, there is a tradeoff between noise correlation without having speech correlation.

ii) Adaptation parameters are crucial since the rate of adaptation and the filter length itself will determine the effectivness of the cancellation.

Kang found a separation of 10 cm to be sufficient when using noise-cancelling type microphones in the helicopter. In an automobile application, it was found that the reference microphone placement which ensured no speech pickup meant that the only coherent noise source was that of the engine [50]. More sophisticated speech enhancement schemes which make use of speech characteristics are summarised in [50].

2.4.3.4 Noise shaping

The *log* spectrum domain and the perceptual sound intensity scale can be utilized in order to minimise the speech distortion inherent in the spectral subtraction. This approach involves modelling of joint noise and speech statistics [87] and is referred to as Nonlinear Spectral Subtraction (NSS). The main idea is to apply less subtraction in the higher SNR regions. This is comparable to the *noise shaping* techniques employed in speech and image coding. An effective implementation has been deployed in the automobile environment where NSS was shown to outperform standard power spectral subtraction in several scenarios [60], [61].

The root operator (1.3) proposed by Lim, [58] is an alternative to the *log* deconvolution. As stated earlier, r is less than 1. Published values for r which worked

well (depending on the spectral representation) are 1/2 and 2/3. The premise for using root homomorphic deconvolution schemes in place of the classical *log* scale is because logarithmic compression causes a whitening effect in the spectrum which makes cepstral analysis particularly sensitive to noise contamination [5]. Proper selection of the root operator can emphasise spectral peaks much more than the *log* operator. In addition, if this is combined with spectral pre-emphasis on the outputs of a linear-scale filterbank, recognition can be improved significantly. This was the case in the noisy car environment reported in [5].

As part of this thesis investigation, we evaluated the root operator with the static IMELDA representation and found the low frequency spectrum to be overemphasised and recognition rates degraded in spectral tilt and white noise experiments [84]. Alexandre [6] has confirmed that the root-cepstrum is sensitive to spectral tilt. We were advised to these critical points:

i) Frequency pre-emphasis (1.1) is particularly important for flattening the spectrum.

ii) Linear spacing may have an advantage over the Mel-scale for low-frequency noise because of its lower frequency resolution in this range.

iii) The low in frequency car noise benefits from the combination of i) and ii).

iv) Root cepstrum showed its usefullness when applied to dynamic acoustic parameters.

Some of the root operators that we investigated in the thesis are shown in Figure 2.7.

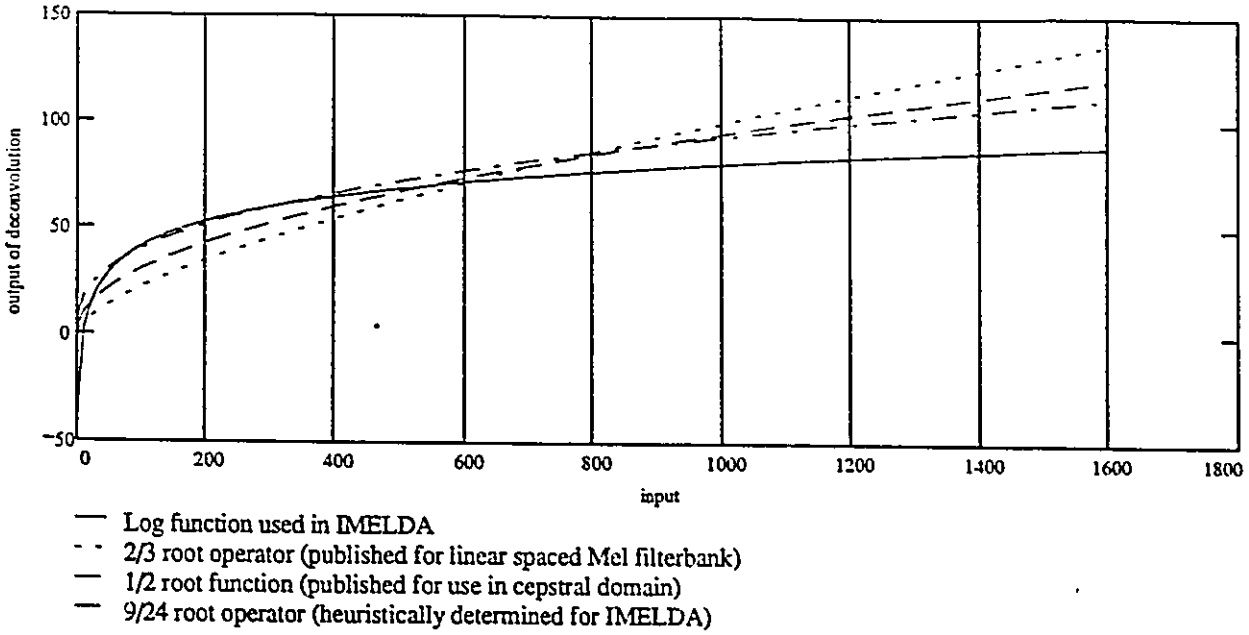


Figure 2.7: Comparison of standard log function (base 10) with various root operators that were recommended in the literature

2.4.3.5 Slow variation removal through subband filtering

A key thrust in this thesis is the principle of subband filtering. This section introduces the reader to the background theory and Chapter 4 describes our investigation into applying subband filtering to IMELDA. Research conducted by Hirsch [39] suggested that speech recognition in environments where room reverberation and stationary noise modify the speech signal can be improved by a technique called subband filtering. The idea is to employ high-pass filters (HPF) in the subbands (filterbank outputs) of the acoustic front-end (Figure 1.5).

It is worthwhile at this point to reiterate the significance of the analysis window, T_w (Section 1.4). This window slides along the input speech signal at a 40 Hz rate to capture the slowly varying physiological aspects of the vocal tract (refer to Table 1.2). This low frequency modulation is a separate issue from the frequency characteristics of the filterbank. Each subband of speech, regardless of the frequency band that it represents, contains temporal contours that fall within the 40 Hz windowing action. As stated earlier, the perceptually relevant portion of speech modulation is between 3 Hz and 25 Hz with speaker specific information dominating below 10 Hz. In this domain, noise appears as a very low-frequency component. Thus, the outputs of each subband can be filtered by a filter with a sharp spectral zero at zero frequency. Applying a log deconvolution prior to the subband filter reduces the distortion to an additive component (2.3) which is then removed by the differencing operation of the HPF.

As part of this thesis investigation, we analysed the temporal contours of subband speech. We did this by computing the long-term PSD of a single output from the filterbank. The filterbank analysis rate is 6.4 ms (f_a = 156 Hz) and with a size 256 FFT applied every 25.6 ms (f_w = 40 Hz). This produces a frequency resolution of 0.61 Hz. Figure 2.8 illustrates the subband modulation of speech for the multi-speaker case (10 male speakers averaged together) and two individual speakers. We also looked at helicopter rotor noise and this was shown in Figure 2.5.

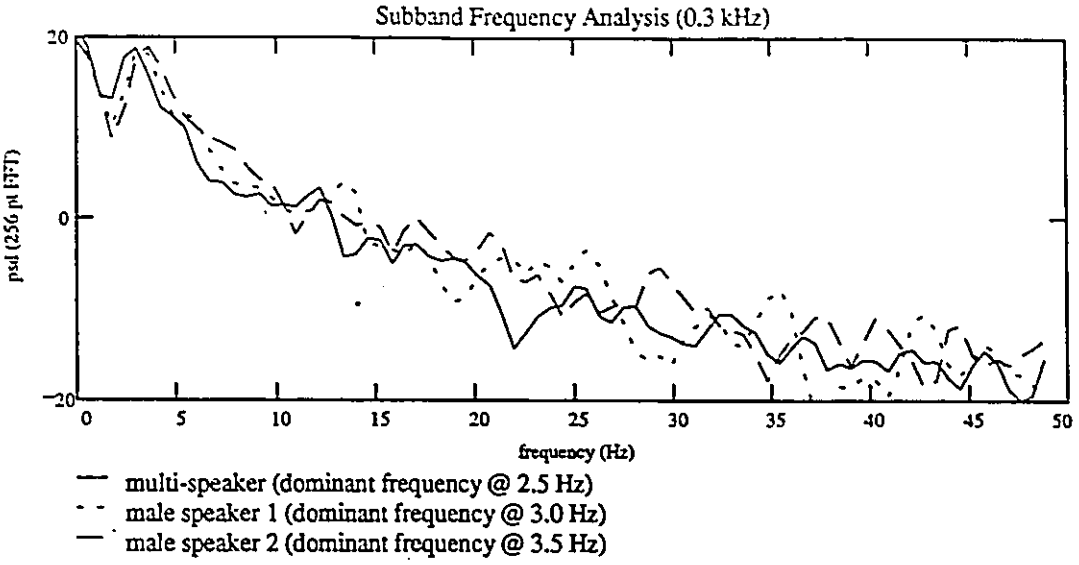


Figure 2.8: Speech modulation for two single speakers and an average of 10 speakers. The strong component at 0 Hz represents noise. The band of interest is 3 Hz to 25 Hz.

Experiments in [39] indicate that a HPF with a 3 dB cut-off situated below 10 Hz decreases the influence of speaker variability and also improved SI speech recognition. Dobler substantiated this conclusion and found that performance with inhomogeneous reference and test could also be improved [20]. One of the objectives of this thesis was to test this hypothesis to see whether any substantial gains could be made to IMELDA. This is detailed later in Chapter 4.

A band-pass form of filtering (BPF), called RelAtive SpecTrAl Approaches (RASTA) operates in the log domain on the outputs of a PLP filterbank [36]. A comparison of the BPF and HPF approaches indicates that the crucial frequency characteristics for a subband filter are [34]:

i) The filter must suppress frequencies below 2 Hz.

ii) Have a flat pass-band to at least 14 Hz.

iii) Either method can achieve these requirements.

Filtering of the subbands can take place either in the linear domain in order to remove additive noise or in the *log* domain to remove channel noise (and speaker specific information). In the case where both types of noise exist, we have the dilemma of which domain to do the filtering. Cascading a linear domain filter with a log domain filter has been shown to be ineffective. One solution to this dilemma is to choose a domain that balances the linear and log characteristics. Morgan [66] presented a solution for the linear and log domain choice by introducing the *J* domain (2.5). Parameter *J* is a signal dependent positive constant. This parameter is based on an estimate of the noise in such a way that small spectral values transform the input linearly, while large spectral values result in more of a logarithmic transformation. Subband filtering takes place in the *J* domain.

$$S(f) = \log(1 + J*Y(f)) \qquad (2.5)$$

While this method does not explicitly handle the situation where both type of noise are present simultaneously, it does offer a certain degree of flexibility. Trials reported in [37] indicate that optimal performance for signals corrupted by both types of noise can be obtained by selecting *J* such that most of the signal is in the logarithmic-like part of the nonlinearity and the noise in the linear-like part. Further details are available in [53], [38]. This is another example of noise shaping (e.g. NSS, root deconvolution).

Augmenting the static subband filtered features with its time derivatives leads to improved recognition rates although the difference is not as great if the features had not been filtered [34]. It was shown that the first order regression filter exhibited a similar suppression below 2 Hz and a flat pass-band up to 10 Hz, thus the low modulation frequencies are already removed by the dynamic features in the un-filtered case.

An alternative form of removing low frequency changes in the *log* domain spectrum is through spectrum normalization [27], [30]. In this method, low-pass filters are used to provide an estimate of the average value of the log-power spectrum in each subband. The normalized spectrum is then calculated by subtracting these estimates from the current spectrum of each subband. This is a HPF operation, and it will suppress noise that is approximately stationary. Furthermore, the HPF characteristic can be used to reduce the effects of a non-stationary channel as shown above. The difference from the subband HPF is the separation of the two contradictory requirements of the filter function: i) filtering to reduce channel variations; ii) time differentiation to include dynamic information. This is accomplished by replacing the HPF with spectrum normalization to achieve i) and using the conventional dynamic speech representation for ii). The estimation interval should not be too short, otherwise it acts as a differentiation. The interval for spectrum normalization suggested in [27] is one second. The HPF approach has a comparatively small time constant and so it is in-between a differentiation and spectrum normalization. The longer estimation interval then produces an average that is over many speech sounds and is thus independent of the actual utterance [27]. In practice, the channel influence should be removed in such a way that the speech is not

distorted and yet the estimation interval should still represent the channel transfer function. Further details are provided in Chapter 4.

2.4.4 Noise resistant similarity measurements

The idea behind robust distortion measures is to emphasise similarity measures from those parts of the spectrum that are less sensitive to noise. This definition covers similarity measures as well as methods that transform speech into stable representations (e.g. LDA). The reader is referred to [50], [25] and [46] for a description of spectral and cepstrally weighted distance measures.

The cepstral projection measure has been found to be more accurate than any other metric when there is discrepency between noise levels in test and reference [69]. This metric, called d_5 is based on the premise that additive white noise reduces the *norm* of cepstral vectors, thus modifying the standard Euclidean metric (1.12). This shrinkage is a function of the noise and the projection measure reflects that. Recognition of Lombard speech can be improved by combining this measure with PLP analysis [2].

Spectral contrast is defined as the spectral variance across a frame (also the norm) of log energy filterbank outputs and this is calculated by summing and squaring the elements across the frame and removing the frame mean [48]. White noise reduces the dynamic range and spectral contrast decreases but the location of the spectral peaks does not change. Frames of voiceless fricatives would have a low contrast measurement. In the d_5 metric, the reference and test are scaled to have norms of one, the Euclidean distance is computed and then scaled by the norm of the test frame, this gives a greater

weight to the more reliable high-norm frames. The angle is less sensitive than the magnitude of the norm and this has given rise to another metric, d_2 [12]. A simpler metric, d_3 is supplemented with a contrast normalization technique [48].

2.4.5 Reference model compensation

This subsection differs from the previous ones in that compensation of the reference models takes place in the so called back-end stage of the recognition system. In the preceding subsections the input speech signal was compensated in the acoustic front-end processing stage.

2.4.5.1 Model decomposition

The premise of the received speech being composed of clean speech and contaminating noise is once again used, but with one difference, each is modelled separately by a HMM and then combined during the search for the best pattern. This technique is referred to as model decomposition [85]. Without describing HMMs, it is sufficient to understand that temporal and frequency characteristics of speech are well modelled in a sort of state machine where the transitions between states are assigned a weight which is based on statistical observations. There are two advantages to model decomposition of speech and noise:

i) This method handles noise that is temporally structured and highly time varying.

ii) It is preferrable to spectral subtraction because the problem of masking useful speech features is avoided.

Using a filterbank, good recognition accuracy has been obtained in stationary pink noise with SNR's as low as -3 dB [85].

A Mel-cepstrum model combination was evaluated in [26], where model parameters are transformed into the linear spectral domain, adapted and then transformed back into the cepstral domain. The advantage of adapting in the linear spectral domain is that low energy speech events are masked in the presence of high energy noise. This avoids the risk of indiscriminantly masking some of the useful speech features as is the case in the spectral subtraction method.

One of the chief differences between DTW and HMM is the concept of state. In DTW, each frame of speech is represented by a single state and the transition to the next frame is constrained by the step functions described in Section 1.5.3. An HMM state, on the other hand, represents several frames, where the duration (i.e. length) is based on probability measures. Therefore, in order to apply HMM noise-modelling techniques to DTW, we must apply the state-based calculation (e.g. noise model update) on every reference frame. The resolution is finer in DTW, but this is at the expense of increased calculations.

The helicopter environment is ripe with periodic noise components, therefore, we considered utilizing the temporal properties inherent in the model decomposition method

to IMELDA[3]. By masking the high energy spectral events associated with the rotor noise only during the peaks (every 5 Hz, according to Figure 2.5) we could avoid the problem of inappropriately masking speech events that occur between the peaks.

Wiener filtering in the cepstral domain has been applied to the model decomposition approach in order to find a clean signal estimate from the noisy observation [7]. The Wiener filter represents the relationship between the desired speech signal and the observed spectrum. In the log domain the desired speech can be extracted from the observation by adding an estimate of the Wiener filter frequency response to the observed spectrum.

2.4.5.2 Adding noise during training phase

Rather than reducing the noise in the test utterance, it is possible to add estimates of the noise to the reference templates. Mismatch between training and test environments totally disappears, provided the noise can be estimated accurately. This simple approach is reportedly better than the more sophisticated techniques of Kalman filtering and spectral transformations [68]. An advantage of the noise-adding approach is that the negative power problem resulting from over-estimation of the noise during spectral subtraction methods is avoided.

In time-varying environments, the noise estimate would be updated regularly as well as the reference templates. However, the time consuming nature of transforming the

---

[3] Described in an internal company memorandum but remains unpublished.

templates from features into the linear domain for the addition, and then back to the feature representation is a real-time disadvantage. In an adaptive noise masking application, it was suggested that time could be saved by only updating the template frames as they are used, and only when the noise changes significantly [86].

Mokbel [68] showed the possibility of transforming clean speech feature vectors to noisy vectors by applying a translation, **B** for the additive noise and a rotation, **A** for the Lombard effect. Letting **X** and **Y** represent log domain filterbank outputs for clean and noisy speech respectively, the two can then be related by the approximation:

$$Y = A*X + B \tag{2.6}$$

Additive noise can be simulated with a translation, by transforming speech and noise vectors from the logarithmic, to the linear scale, adding them, and transforming back to the logarithmic scale. The rotation part of the transformation is more difficult to determine and Mokbel was not able to show this.

We have taken a similar approach in this thesis but for the purposes of artificially modifying the speech classes with additive noise and spectral tilt in order to compute a noise robust IMELDA transform. We ignored the Lombard effect and treated each noise separately, thus simplifying (2.6). Convolutional noise (i.e. tilt) was modelled as being additive in the logarithmic domain and $\mathbf{B}_t$ represents the translation vector (2.7). An anti-log operation returns the speech spectrum to the linear domain where additive noise, represented by translation vector $\mathbf{B}_n$ can be added to it (2.8).

$$Y = X + B_t \tag{2.7}$$

$$Y = 20\log\left(\left(\frac{X}{20}\right)^{10} + B_n\right) \tag{2.8}$$

2.4.6 Stress compensation

The spectral changes that occur in a talker's speech when subjected to noise (Lombard effect) can be compensated with spectral mapping techniques [78]. Mapping rules are obtained by analysing samples of the same words spoken in stress-free and noisy conditions. In the test condition, stressed speech is mapped back to the stress-free state and then conventional pattern matching can take place with the stress free reference templates. This approach was employed in the study of Mokbel [68] in order to confirm their noise additvity assumption.

An alternative stress compensation technique that is suitable to HMMs is based on a linear transformation [13]. The idea is to shift the means and scale the variances of the states in the reference model, according to measured observations in stressed speech. Both of these methods have led to improved recognition. The reader is referred to the studies by Hansen et al [32] for a summary of current approaches to the robust recognition of stressed speech.

2.5 Coping with speaker variation

There are several factors other than noise that contribute towards making speech recognition outside of the laboratory difficult. These factors include:

i) Spontaneous speech that falls outside of the library of reference words. Examples are coughing, restarting, additional and unexpected words or the wrong words and the repetition of utterances. Methods to deal with spontaneous speech are referenced in [25].

ii) Style shifting (change in delivery) can result from the psychological awareness that the talker is communicating with a machine.

The difference in a speaker's style is evident when we compare how a speaker reads from a list (often the case in laboratory experiments) and when they are trying to accomplish a task. In the NRC helicopter experiments, we have noticed the following:

i) Rate of delivery is faster when reading from a list and this can bother a recogniser that is trained this way, but operated in a more controlled, and hence slower delivery style. This is another example of mismatch between training and testing.

ii) Articulation of words can be more precise when controlling a device by voice. This was particularly noticeable when the talker is busy with other tasks (e.g. flying).

iii) Speech amplitude decreases towards the end of a string of spoken words.

iv) The linguistic style changes when switching from reading to commanding.

v) Feedback of the state of the recogniser to the talker is important and so is the time it takes to respond.

Solutions to these problems involve multi-style training. One example is to use reference templates that represent words spoken in isolation and those with co-articulation effects.

In Chapter 1, we indicated some of the between-speaker variations. In a SI recogniser, the performance can be greatly improved by training with a large database and incorporating statistical models of speech variation. Since listeners can adapt to an unfamiliar speaker's voice after hearing just a few syllables, it is possible to develop systems that adapt to the speaker. The major speaker-adaptation methods are referenced in [25].

2.6 Conclusions

In this chapter, a review of the problems that occur in speech recognition under adverse conditions was provided. The major methods for dealing with this problem are all described in detail. The ideal circumstance, which is to train the recogniser with the same conditions as encountered in the field, was contrasted with the more realistic cases of mismatch that can take place. Input equipment (e.g. microphone, telephone channel), environment noise (e.g. room reverberation, additive noise), and the speaker (e.g. stress and the Lombard effect) all contribute to the mismatch between reference and test. The helicopter environment was depicted as an application with changing acoustic variability. Two microphone adaptive filters were shown to be effective in environments where the

noise is changing. It was pointed out that microphone placement is crucial in this method.

The concept of noise shaping was introduced by an example of root deconvolution (Section 2.4.3.4). Spectral peaks can be emphasised more than the conventional log operator. Although the preliminary investigation outlined in this thesis was not successful, reasons are explained and alternate solutions were presented. It was theorised that IMELDA also does a form of noise shaping and this was left to Chapter 3 to explore.

Dynamic feature analysis was shown to have noise resistant properties. Long window lengths allow integration over many speech sounds that make dynamic features uncorrelated to noise. Stated another way, slowly stationary noise is low in frequency content compared to speech and this can be removed by the differencing operation inherent in dynamic feature analysis. This principle is also utilized in the technique of subband filtering. In Section 2.4.3.5 we described two approaches to removing slowly varying low-pass noise:

i) High-pass and band-pass filtering on the outputs of the filterbank analysis.

ii) Channel noise removal through spectrum normalization.

Theoretically, subband filtering can reduce acoustic variability. This leads to:

i) Removal of stationary additive noise.

ii) Reduction of channel distortions (convolution noise).

iii) Improved recognition of inhomogeneous reference and test speech.

iv) Reduction of speaker variability and thus enhance SI recognition by removing low-frequency inter-speaker differences.

Chapter 4 considers which subband filtering techniques are appropriate for integration within the IMELDA system. Experiments are designed to draw conclusions about the cross conditional performance of subband filtering.

Reference model compensation (Section 2.4.5) differs from other noise compensation techniques because the back-end rather than the front-end of the recognition system is used. A proposal to utilize the temporal properties of the model decomposition technique in IMELDA was described in Section 2.4.5.1. It is recommended that this be pursued further. Adding noise during the training phase (Section 2.4.5.2) reduces the mismatch problem, if the noise can be modelled. It was shown that clean speech features could be transformed to noisy vectors by a translation (vector addition) for additive noise and a rotation (matrix multiply) for the Lombard effect. A similar approach was taken in this thesis for the robust IMELDA transform method used in Chapter 3.

Problems with speaker variation (e.g. spontaneous speech and style shifting) and between-speaker variations are discussed in Section 2.5. In this thesis we only address the latter issue. This was taken in context of SI recognition where we study LDA in Chapter 3 and subband filtering in Chapter 4.

CHAPTER 3

EXAMINATION OF IMELDA

In this chapter we focus our attention on IMELDA. We illustrate the reasons why this technique works well in adverse conditions and ways that it can be improved upon. First, the signal processing and pattern matching techniques that are employed in IMELDA is outlined. This section adds to the background material of Chapter 1. In Section 3.2 we undertake a rigorous examination of how LDA is applied in IMELDA. Attention is drawn to those details that are relevant to real-world speech recognition systems and are often not answered in the literature. These include:

i) The application of noise in the LDA procedure.

ii) Sample size requirements.

iii) Determining the optimal number of transform parameters.

iv) Initialisation.

We answer these questions by applying the fundamental principles of discriminant analysis (outlined in Section 1.5) along with statistical measurements to speech data that is drawn from computer simulations. Analytical tools that predict performance of IMELDA are described and compared with results from speech recognition experiments.

3.1 The IMELDA system

This section contains a brief description of the speech recognition system that we used in our experiments. With respect to Figure 1.1 the front-end is an FFT-based filterbank followed by a linear discriminant transformation while the back-end is DTW pattern matching. Figure 1.5 provides details of the front-end signal processing. The input sampling rate is 8 kHz with a passband of 150 Hz to 3.6 kHz. We do not pre-emphasise the audio input because it has recently been shown that this has a deleterious effect in noise [22]. Instead we flatten the spectrum representation directly by adding a fixed value to each subband resulting in a 6 dB/octave lift from 1 kHz to 4 kHz. The speech is analysed at a 6.4 ms frame rate with 51 samples per frame. This serves as a compromise between the different rates required for female and male speech. The standard spectral analysis described in Section 1.5.1 is used to extract the static LCE representation. DLCE parameters are derived from a seven frame, 45 ms interval of the static parameters (1.4).

We use a threshold vector such that each element corresponds to one subband in the LCE vector. The threshold vector is designed to lie below the minimum speech level and above the noise floor. Traditional systems use a single value for the thresholding (a uniform vector), whereas we use the entire vector. This yields a more accurate spectral representation of the noise. Another key difference in our approach is that we utilize spectral characteristics of the speaker's reference templates (in LCE format) as boundary conditions for the threshold vector. These conditions are computed by constructing histograms of LCE data. A 30% histogram point is considered to be the

maximum value of thresholding and this acts as an upper bound to the threshold vector. This process is repeated for each subband. In this way the vector is allowed to be greater in high energy regions (e.g. between 300 Hz and 1 kHz) and lower in low energy regions (e.g. above 3 kHz). The result is referred to as the speaker's *spectral envelope*. We believe this is a form of noise shaping. This work was reported in [55]. In Figure 3.1 we depict a threshold vector that is designed to lie above the noise floor for the helicopter PSD of Figure 2.4 but below the speaker's spectral envelope (see subbands 17 to 19).

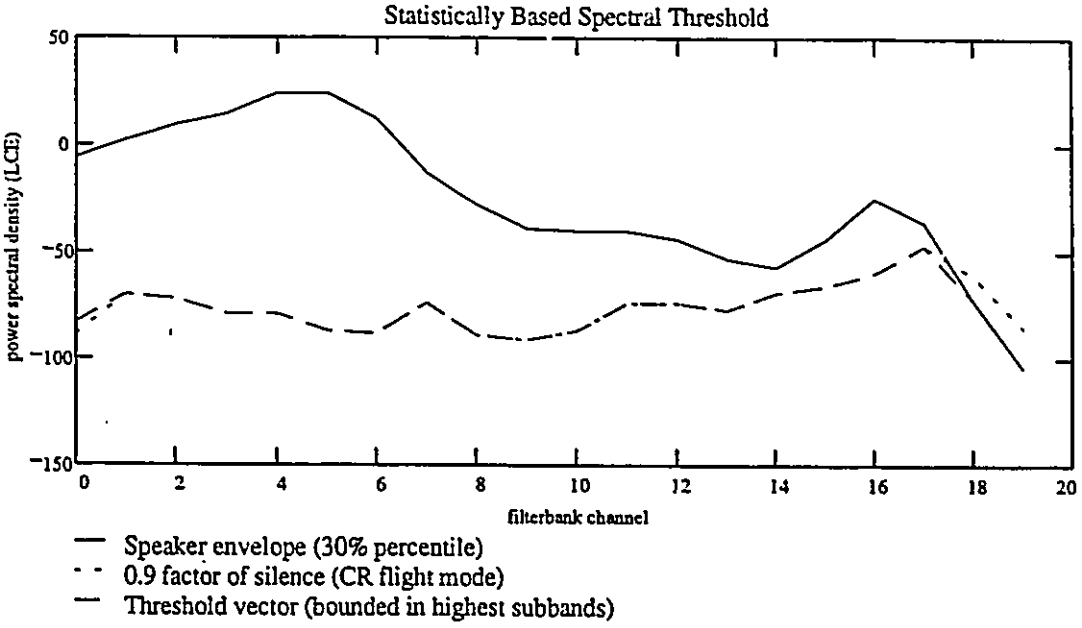


Figure 3.1: Threshold vector designed to lie above the helicopter noise PSD of Figure 2.4 but it is constrained by the speaker's spectral envelope.

In our experience, it is best to apply a modest threshold vector during the training process [56]. This allows the LDA to fully 'appreciate' all speech sounds including the low level sounds. After the reference templates and the LDA transform have been computed, a threshold vector that is appropriate for the noisy test condition is applied to the

templates. It has been reported elsewhere [48] and confirmed with our findings that the noise level does not have to be constant and accurately known.

In the back-end, we use the one-pass DTW continuous word recognition algorithm pioneered by Bridle [11]. Section 1.5.2 provides the reader with the basic idea behind DTW pattern matching. Three adjacent feature vectors are averaged in order to decimate the frame rate from 6.4 ms (156 Hz) to 19.2 ms (52 Hz). This three to one reduction provides a factor of nine savings in the pattern matching. Reducing the resolution has a small impact on the performance of the system [45]. Experiments have been conducted with a four to one decimation but this degraded noise performance.

3.2 Examination of linear discriminant analysis

3.2.1 Overview

The purpose of this section is twofold, first we undertake a thorough investigation into those properties of LDA that are relevant for application in a practical speech recognition system. Secondly, we provide analytical methods that enhance our understanding of LDA and give us a level of confidence before undertaking time-consuming recognition tests. To achieve these goals we:

i) Investigate the importance of the template alignment.

ii) Analyse statistical variability of the **B** and **W** covariance matrices for various pool sizes and acoustic representations.

iii) Assess the usefullness of separability measures, J_1 and J_4.

iv) Compare the standard LDA transform where only quiet data is used with that which is augmented by samples of noise.

v) Introduce the *confusion* matrix as a tool for assessing the separability of speech classes.

vi) Inspect the F-ratios of the LDA transform in order to determine the optimum number of coefficients for a given application.

vii) Evaluate cross-conditional performance (i.e. matched versus mismatched conditions).

Throughout this section we state which aspects of LDA contribute to performance in noise.

3.2.2 Applying LDA to speech recognition

LDA has been used in speaker recognition tasks by Bricker et al [10], first in 1971 and later by Atal [4] and Hunt et all [41]. In 1979, Hunt [42] described how LDA could be used in speech recognition, and in 1988 he employed this technique in an auditory model-based system [44]. The objective of this application was to combine two disparate sets of parameters, 64 in total and to reduce this to eight features. When LDA was used to combine the static and dynamic outputs of IMELDA, it was found that the features were noise-resistant [45], [46]. The key assumptions that Hunt made are:

i) A word can be modelled as a sequence of (generally continuous) speech sounds.

ii) The acoustic representation at any point in time is multi-variate and normally distributed about the mean representation for that sound.

iii) The distributions about the means are the same for all speech sounds.

iv) Dynamic time warping can sufficiently align corresponding speech sounds in different samples of the same word.

v) The templates (mean representations) of a word is estimated by aligning and averaging together many samples of the word. This is repeated for all words in the vocabulary.

vi) The within-class covariance matrix, **W** is estimated by aligning each individual sample to its template and then subtracting the corresponding values of the acoustic representation in each aligned frame pair. This is repeated for all words in the vocabulary and the results are pooled togther.

vii) The between-class covariance matrix, **B** is estimated by computing the covariance of all frames from all templates about their overall mean.

3.2.3 Alignment of the speech sounds

Reference templates are constructed by *warping* (i.e. DTW aligning) pairs of samples together. This is repeated in a pyramidal fashion where the process starts at the bottom of the pyramid with N/2 pairs, where N is the number of samples. The next higher row is (N/2)/2 and so on until there is a single entity, the reference template. The

entries for **W** are computed by warping each sample with the reference template and computing the difference.

Although acoustic representations are used in computing **W** and **B**, the warping is done on the feature representation. This is because of the need for orthogonality in the Euclidean distance calculation. We use the cosine transformation to produce the features used in this step of the training process. The result is an LDA transform and a set of reference templates. The cosine transform (1.17) was selected because it is easily created and can be applied to any quiet speech database. An LDA derived transform is best suited to the type of data that it was derived for (i.e. it is not generic). A second iteration of computing the transform takes place by using the transform result from the first pass. We do this because the discriminant functions may be better at distinguishing dissimilar sounds than the cosine basis functions. Figure 3.2 compares coefficients from the cosine transform with an LDA transform. It can be seen that the LDA functions are not nearly as smooth as the cosine functions and are thus more discriminating. The smooth functions account for much of the variance but less of the discriminating information in log power spectra according to Hunt [46].

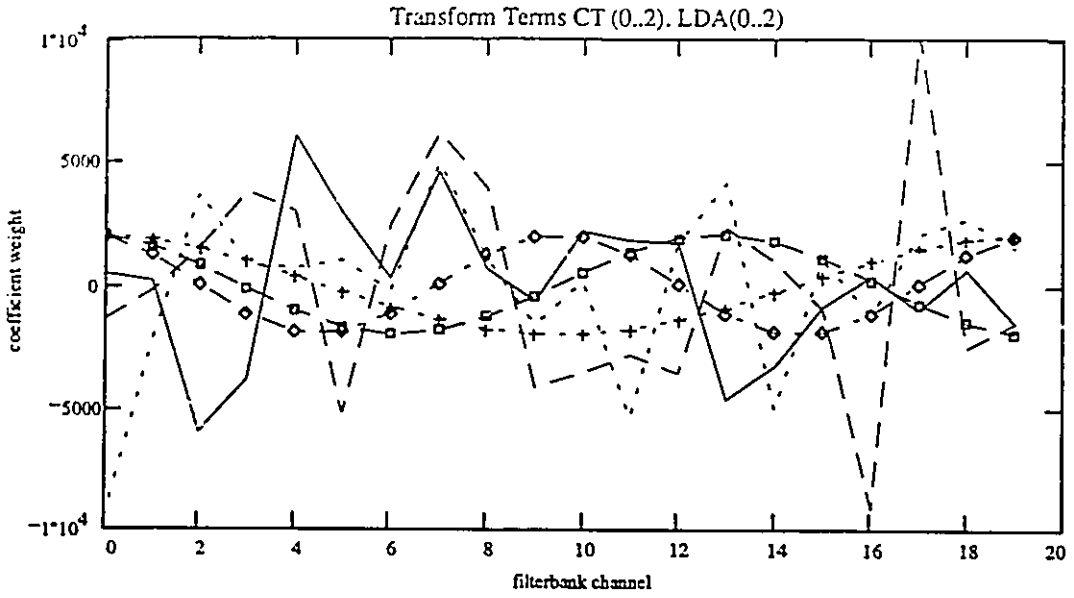


Figure 3.2: Comparison of LDA eigenvectors and cosine basis functions. The LDA functions provide more discriminating power.

Referring to Table 3.1, column 3, the 1-pass and 2-pass template alignment methods are compared for a single speaker pool (i.e. rows a1 and a2) and for the multi-speaker pool (i.e. rows b1 and b2). The conclusions that we can draw from this are:

i) 1-pass alignment to 2-pass alignment improves separability. The outcome is a decrease in tr(**W**) and an increase tr(**B**), consequently improving the ratio of between to within class scatter. This is reflected in both J_1 and J_4. Both single-speaker and multi-speaker pools are improved by the 2-pass alignment, but this is much more noticeable for the latter.

ii) Further evidence of the crucial nature of the template alignment is shown by comparing b1 and b2 with and b2 and $b2^2$. Here we used a transform for the

alignment that was derived from LCE and DLCE, and we observe the outcome on the variance of the LCE parameters only. In the latter case, tr(**W**) improves slightly by becoming lower in value, but more importantly tr(**B**) increased significantly. Therefore, DLCE helps in separating dissimilar speech sounds.

	pool	align	LDA data	acoustic reps		tr(W)	tr(B)	J_1	J_4
a1	SD 150	1-pass	Q-W,Q-B	LCE		5112	29440	53.03	5.76
a2	SD 150	2-pass	Q-W,Q-B	LCE		5078	30270	55.97	5.96
a3	SD 150	2-pass	QNT-W,Q-B	LCE		12760	30270	38.61	2.37
a4	SD 150	2-pass	Q-W,Q-B	LCE&DLCE		24820	94090	84.11	3.79
a5	SD 150	2-pass	QNT-W,Q-B	LCE&DLCE		36280	94090	61.88	2.59
a6[1]	SD 150	2-pass	QNT-W,QNT-B	LCE&DLCE		36160	85380	53.86	2.36
b1	SI 1200	1-pass	Q-W,Q-B	LCE		12400	23390	15.87	1.89
b2	SI 1200	2-pass	Q-W,Q-B	LCE		10810	24820	18.07	2.30
b2[2]	SI 1200	2-pass	Q-W,Q-B	LCE		11110	26680	19.04	2.40
b3	SI 1200	2-pass	QNT-W,Q-B	LCE		17490	24820	15.38	1.42
b4	SI 1200	2-pass	Q-W,Q-B	LCE&DLCE		36970	65550	29.52	1.77
b5	SI 1200	2-pass	QNT-W,Q-B	LCE&DLCE		44530	65550	26.51	1.47

Table 3.1: Table of covariance and separability measures. This table is a compilation of measurements for various hypothesis that are investigated throughout the chapter. Rows designated by 'a' indicate a single speaker pool while rows with a 'b' denote a multi-speaker pool of covariance data.

[1] same as QNT but the reference (centroid) used in **W** and **B** is average of the 3 conditions

[2] two-pass time-alignment of LCE with transform from LCE&DLCE

In Figure 3.3 covariance of the individual subbands (diagonal of covariance matrix) is plotted from left to right with filter spacing along the Mel-scale. We see that the within-class and between-class variance is improved in a consistent fashion for all subbands. Figures 3.3.1 and 3.3.2 indicate SI multi-speaker pooling while Figures 3.3.3 and 3.3.4 are for a single speaker (SD) pool. We have also observed a decrease in within-class correlation and an increase in between-class correlation. We conclude that the 2-pass alignment method is a useful addition to the preparation of a linear discriminant transform.

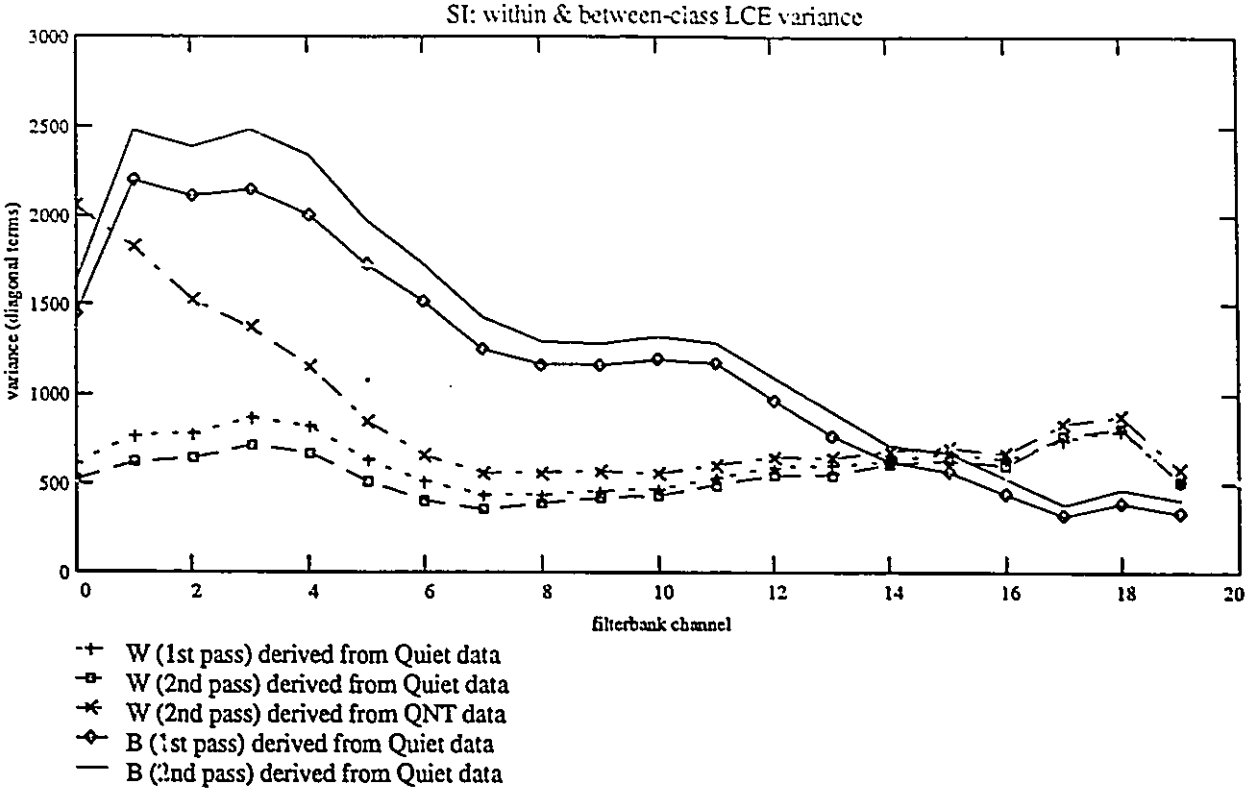


Figure 3.3.1: Increase in B and decrease in W variance from 1 to 2 pass is good. QNT increases W however. Subbands 15 to 19 are poor in discrimination power because B<W.

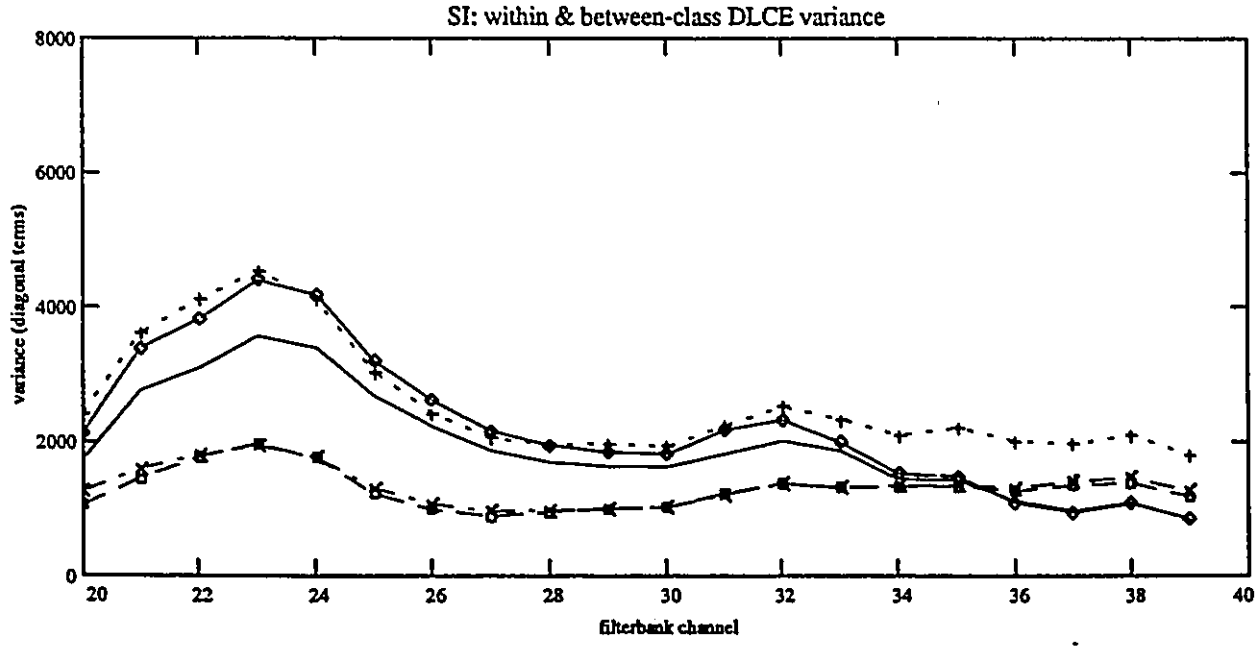


Figure 3.3.2: Dynamic representation shows significant decrease in W variance from 1 to 2 pass. QNT does not change (good) but B decreases. Subbands 15 to 19 are poor in discrimination.

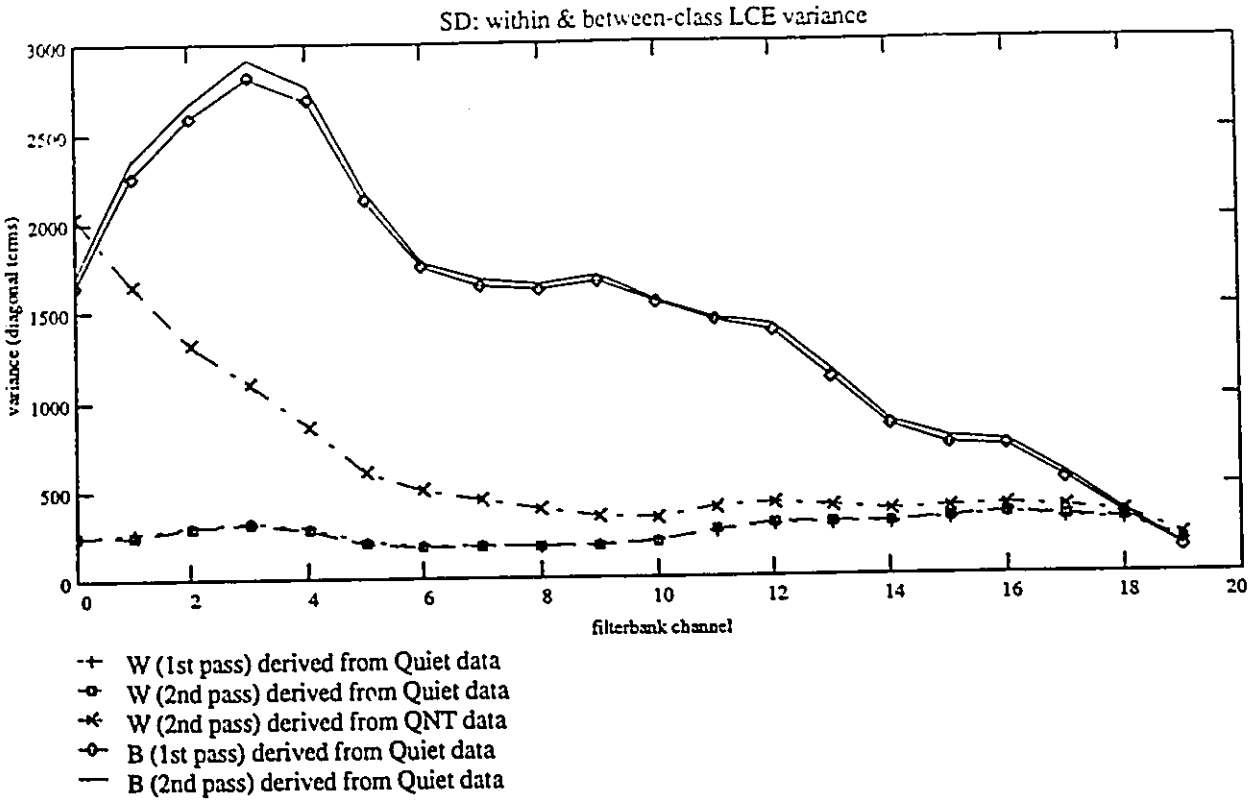


Figure 3.3.3: For SD LCE the increase in B is slight, no change in W variance from 1 to 2 pass. QNT increases W however. Only subbands 18 to 19 are poor in discrimination.

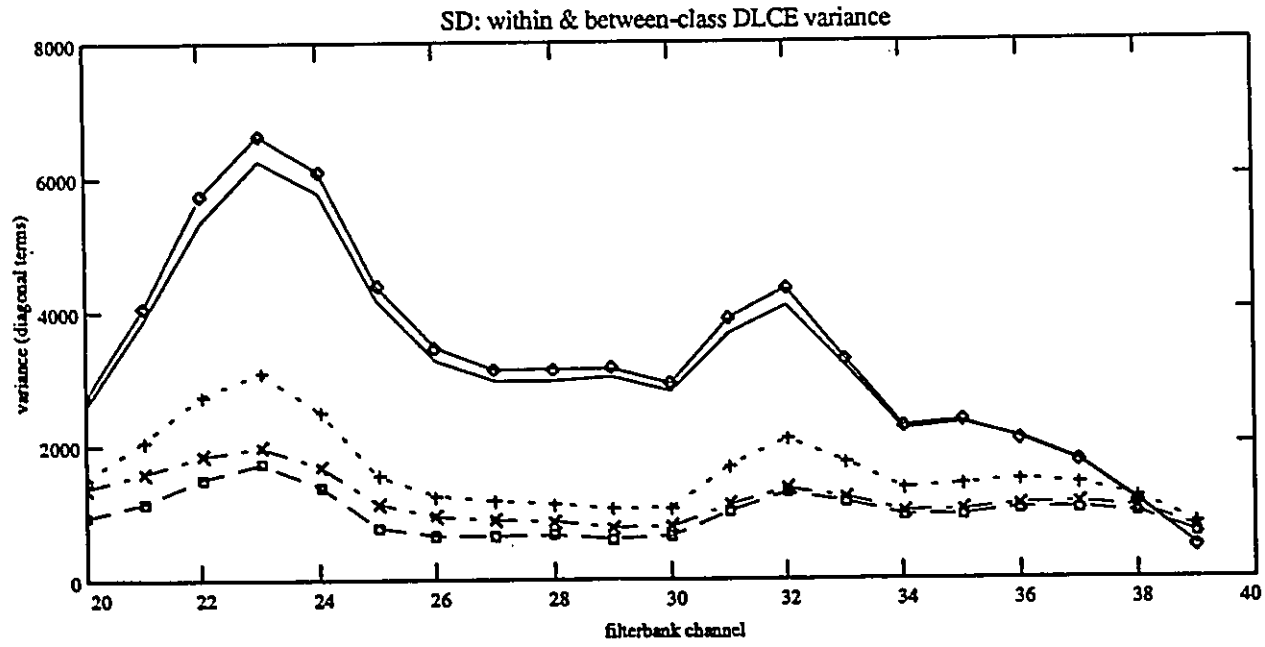


Figure 3.3.4: For SD DLCE there is a decrease in W variance from 1 to 2 pass. QNT increases only slightly (good) but B decreases. Only subbands 18 to 19 are poor in discrimination.

3.2.4 Adding degradations to the within-class covariance estimate

Speech recognition for noisy test conditions can be improved by adding samples of noisy speech to **W** before the discriminant analysis. We use the term *QNT* to describe the process of applying *N*oisy and *T*ilted degradations to the *Q*uiet covariance matrix. The concept was published by Hunt [45], but no theoretical explanation has been given to why it works as well as it does in noise. In this thesis we attempt to answer this question. The first step is to apply general LDA theory, outlined in section 1.5.2, to the assumptions that have been made by Hunt. We believe that the desirable properties for **W** are:

i) Many samples of the words are to be provided to represent as much of the total variance as possible.

ii) All variations of the words that are encountered in the test condition should be made available for the covariance analysis. Noise influences the distribution about the centroid (the mean gets shifted) and so noisy speech samples should be included in the covariance analysis. This includes degradations like additive and convolution type noise.

iii) The alignment process must accurately order together the phonetically similar sounds in the sample with the mean sample (template), so that the difference and therefore the variance is not misrepresented. Consequently, the alignment path should be from speech that is not degraded by noise.

Requirements for the templates used in **B** are:

i) The acoustic representation should provide cues that distinguish between the speech sounds of a word. The words chosen for the vocabulary should have sufficient speech sounds to make them distinguishable. If the templates are quite different, then the covariance will be large between the classes and discrimination will improve. If the words are too similar then the scatter is going to be small and the discrimination will be poor.

ii) The templates should be from quiet speech and not be exposed to any degradations that will diminish the desired property outlined in i). This is in contrast to property ii) of **W**.

Clean speech is required in constructing **B** because we are interested in the scatter between the classes while ignoring the scatter (variations) within the class. Since we are representing each class by its centroid it is vital that we do this carefully as possible, therefore we use clean speech. Figure 3.4 illustrates this process in block diagram form and this is described below:

i) Templates are built in the normal fashion by averaging together quiet speech samples. The resulting templates are used as the mean representations for **B**.

ii) The within-class covariance matrix, **W** is estimated in the normal fashion with samples of quiet speech. The alignment paths are retained for the next step.

iii) **W** is then augmented with samples of noisy speech. The time-warping paths from quiet speech samples are used in the alignment process.

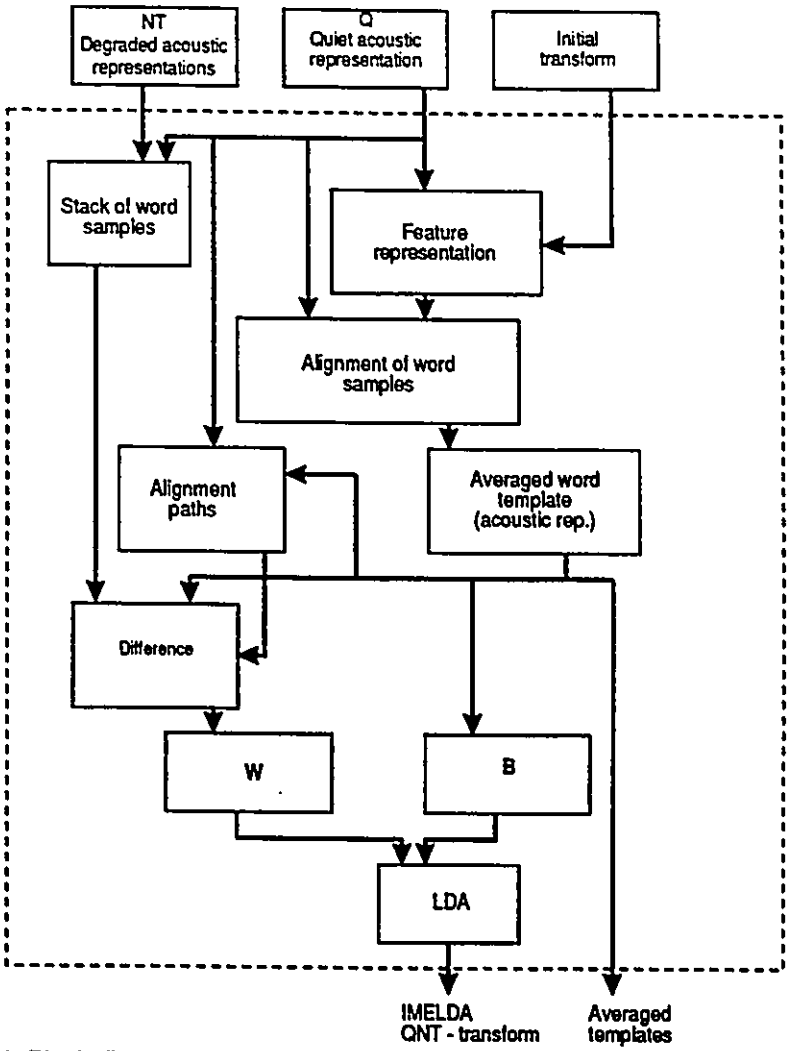


Figure 3.4: Block diagram of Q and QNT transform preparation

The noisy speech samples used for our experiments were:

i) Broadband additive noise that is simulated by adding artifical white noise to the audio waveform for a 15 dB SNR (N).

ii) Linear distortion (spectral tilt) which simulates audio channel noise with a 6 dB/octave filter (T). This is obtained via (1.1) where α=0.95.

In the QNT process, both noise sources are treated separately. In [45], the degradations are applied to the audio waveform before generating the LCE and DLCE spectral representations used in augmenting **W**. In this thesis, we applied the degradations directly to the spectrum of the quiet samples. By making modifications directly to the spectral representations the QNT process is greatly simplified. This represents a significant memory saving for the training process since it would no longer require the audio waveform representation but just the filterbank acoustic representation. Linear distortions (i.e. convolutional noise) are handled by employing the fact that convolution in the time domain (audio waveform) becomes multiplicative in the frequency domain (post FFT) and then additive in the log domain (LCE filterbank). This was shown in (2.1) and (2.3). Thus, linear distortions can be simulated by adding a pre-determined offset (i.e. $\mathbf{B}_t$) to each subband in the LCE acoustic representation (2.7). When the distortion is stationary the following procedure is used to compute $\mathbf{B}_t$:

i) Treat the channel and recognition system as two systems in cascade. Determine the frequency response of the recognition system with and without the channel.

This is done by applying a time domain impulse and measuring the spectrum via the LCE.

ii) Compute the difference between the two frequency responses, the result is the frequency response of the unknown channel. The result is a vector, $\mathbf{B}_t$ where each element corresponds to the offset of each subband.

Time-varying channels require an alternative method for determining a sequence of offset values. For additive white noise that is relatively stationary we used the axiom described in (2.8). The LCE acoustic representation of the noise source is collected and stored in a buffer that is long enough to encompass the duration of the stationarity. With respect to (2.8) the noise vector $\mathbf{B}_n$ represents each sample in the noise buffer. Noise and speech are converted to the linear domain through an anti-log operation and added together. The result is converted back into the log domain and is used in the QNT process.

We would like to compare the effectiveness of various transforms for their ability to discriminate or separate words once they have been corrupted by noise. Scatter diagrams (e.g. Figure 1.6) are a good way of showing the separation between classes and the dispersion about the class centroid (template). However, this is only suitable for two features because it is impossible to display multidimensional features. Therefore, as part of the thesis, we conducted an experiment that used DTW to compute the dissimilarity (i.e. *score* (1.21)) between quiet reference digits and noisy test digits. The test data was from the two noise conditions described above. Results are stored in a

table called a *confusion matrix*. The lowest relative score is along the diagonal, while the off-diagonal entries show the relative difference between each pair of words. We have extracted a single row from the matrix corresponding to digit 'one' and presented this in bar chart form in Figure 3.5.

Dissimilarity measures are normalized to the digit 'one'. Digits 'four', 'five' and 'nine' have the smallest score and are thus the most easily confused with 'one'. Digits 'zero', 'three' and 'six' are the least likely to be confused with 'one'. Because the score gets smaller as we go from the quiet test case to the noise and tilt cases we present them on the same bar-graph. The topmost line corresponds to quiet (highest score), the second highest is the score in noise and the final mark is for tilt (lowest score).

The data used in this experiment is from the multi-speaker LCE&DLCE acoustic representation. Details for the Q and QNT transforms are in rows 'b4' and 'b5' of Table 3.1 respectively. The first graph of Figure 3.5 is the results from the Q transform and the second graph the QNT transform. In quiet and noise conditions the Q transform appears to be marginally better than the QNT by virtue of the slightly higher dissimilarity measures. The advantage shifts to the QNT transform for the tilt test condition. Scores drop considerably for the Q transformed digits while the drop is not nearly as significant for the QNT digits! We conclude from this experiment that the QNT transform is superior overall. We may also conclude that it is the dynamic acoustic representation, DLCE which is imparting robustness against the additive noise in this experiment. Thus, any expected advantage that the QNT transform may offer in additive noise is not noticeable.

This conclusion is substantiated in Section 3.2.8 where recognition experiments compare performance with Q and QNT transforms with and without the dynamic representation.

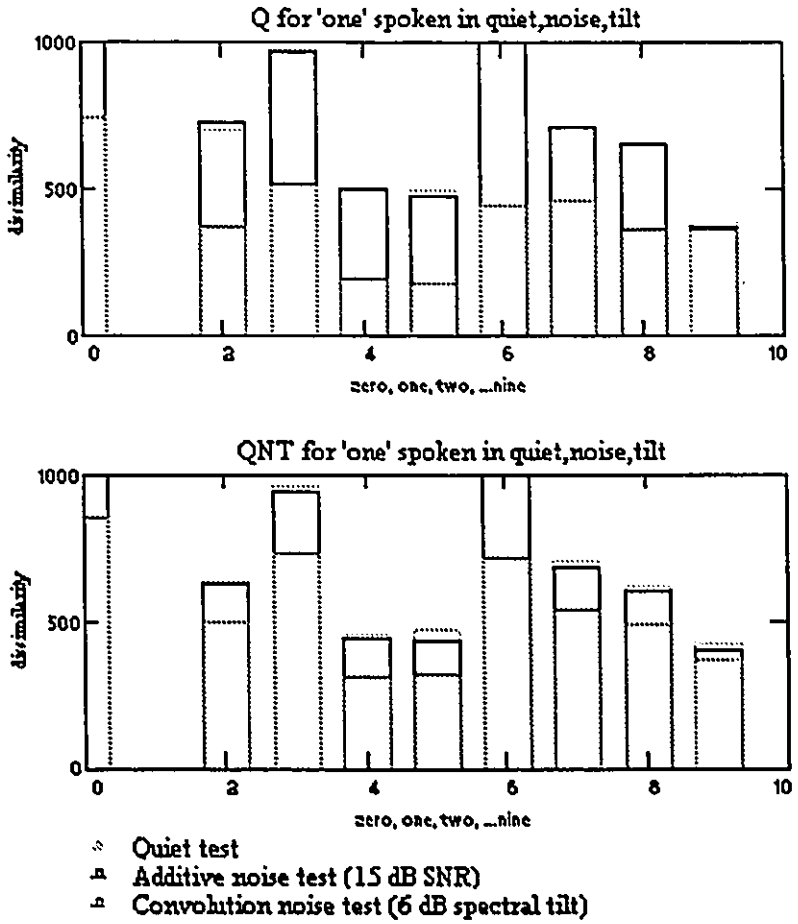


Figure 3.5: Confusability for the digit 'one' spoken in quiet, noise and tilt. QNT imparts better dissimilarity than Q in the tilt condition.

Column four in Table 3.1 identifies the experiments with covariance matrices constructed solely from quiet samples (Q) from those that were augmented with noise (QNT). The following conclusions have been drawn:

i) Augmenting the **W** covariance matrix with noise, results in much higher variance for **W** and this is expressed in metrics that are lower.

ii) If we consider **W** as representative of the test condition(s) and **B** the reference condition then we can see that the Q transform represents the matched case. The separability is much lower for QNT because it signifies the mismatched conditions. We need to keep this in mind when we compare the two, in other words Q is optimum for quiet test conditions, whereas QNT represents quiet, noise and tilt together.

iii) In experiment a6 we allowed the centroid of the two scatter matrices (covariance) to represent the three test conditions (the test templates were not changed). We expect the within-class scatter to be shifted closer to the modified centroid, and the trace did decrease but only slightly. By introducing noise to **B** we anticipated that the variance would increase, but it decreased. Consequently the metrics both decreased (compared to test a5).

3.2.5 Significance of statistical variability in the within-class covariance

This section focuses on the question of how much data is necessary for a practical recognition system and yet satisfies the covariance properties outlined in Section 1.5. Intuitively, we can say that we must have enough speech samples to cover the variability encountered in the test conditions. Variability includes: i) spectral characteristics (frequency and power) of the speaker's vocal tract resonants; ii) temporal modulations

(rate of delivery); iii) environmental (additive and convolution noise); iv) vocal effort (shouting, cold, stress). We cannot cover all of these possibilities, so we focus our attention on speaker and environment variability. This is achieved by analysing covariance matrices from multiple speakers pooled together, full and reduced samples from an individual speaker and speech spoken in quiet conditions and with noise.

Markel found that the Fisher linear discriminant ratio (J_1, J_4) increased significantly by performing long-term averaging on the acoustic parameters [64]. Without repeating the derivation it can be shown that the variance of a long-term averaged parameter (feature) is inversely proportional to the number of samples. Thus, as the number of samples increases the within-class variability decreases. This is evident in Figure 3.6. Two-dimensional scatter plots for the first three LDA transformed features, Y_1 to Y_3 from acoustic parameters used in a speaker verification task [64] are shown. The four classes are A, B, C and D. In Figure 3.6 there are 10 times the number of samples, L_v going from left to right. The variability is illustrated by the dispersion of samples around the class mean and this is enclosed by an ellipse for each of the four classes. Classes B and C have an undesirable overlap for all three combinations of features when L_v = 100. Increasing the number of samples to 1000 decreases the variance. The scatter is now tightly dispersed around the class mean and now just two features Y_1 and Y_2 are needed for perfect discrimination.

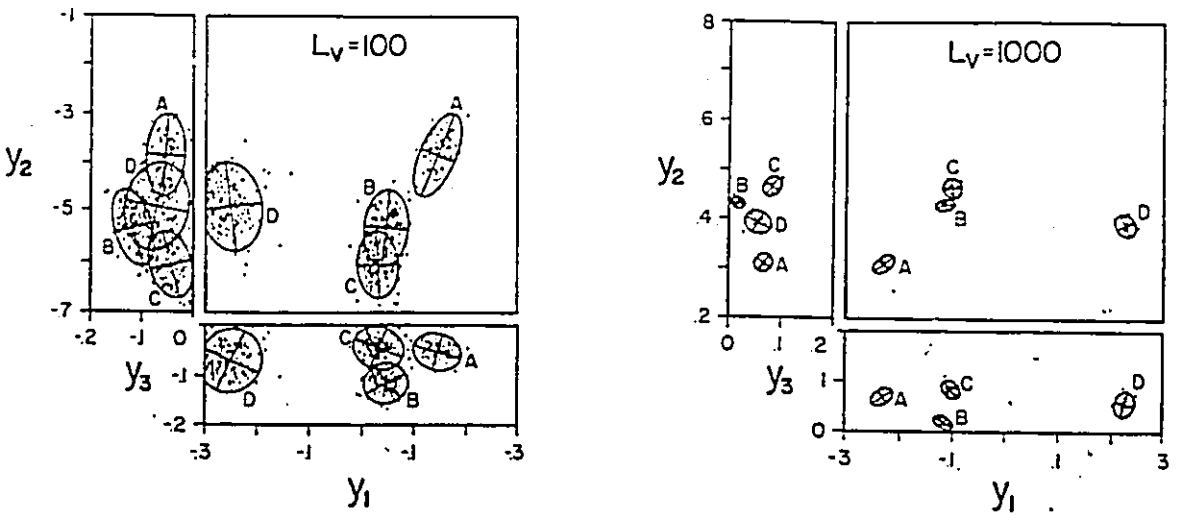


Figure 3.6: Relationship between sample size and discrimination (from [64]). Sample size is 100 on left, 1000 on right. Variance decreases and class spearability increases.

In a speaker-verification task, Doddington [19] reported that the large dimension (20) of word-level covariance matrices require large amounts of input data. This conclusion was drawn from open-set experiments where test data that was not included in the training data proved to be inferior to closed-set experiments in which the training and test data were the same. This is the same experience that we have in speech recognition and it is the reason that we try to reflect the test conditions in the reference templates. We do this by augmenting the **W** matrix with data from the test condition in the QNT process.

Hunt reported that recognition rate decreased when the amount of training data (from quiet and test conditions) was split in half and isolated from that used for testing [45]. The conclusion was that decreasing the pool size for the covariance matrix was the reason for the drop in performance rather than any bias that may have resulted from

including test data in the training set. Interestingly, the additive noise tests were worse off than the quiet and tilt cases. Thus, the number of samples is a critical element of LDA and this affects noise performance.

Doddington's experiments also showed that augmenting the covariance matrix with test data improved open-set performance but closed-set performance decreased! This has been our experience as well. When augmenting quiet speech with white noise and spectral tilt the corresponding test cases improve, but at the expense of the quiet test condition. Therefore, cross-conditional performance will not necessarily be improved by the QNT process. We will substantiate this statement with recognition experiments in Section 3.2.8.

We examine the issue of speaker variability by comparing covariance matrices with data from a single speaker and covariance data when the pool has been augmented with nine more speakers. The pool size increases from 150 to 1200. Notations "a" and "b" in Table 3.1 denote the single speaker and multi-speaker pools respectively. The results confirm our expectations. These are listed below:

i) **W** becomes significantly higher.

ii) The scatter in **B** is now much lower. The scatter of **B** decreases because the reference templates that are now an average of many different speakers have become 'generic' and therefore less distinguishable.

iii) Both metrics are significantly lower for the multi-speaker case.

pool	Change in J_1 from Q to QNT	
	LCE	LCE&DLCE
single speaker pool	-30%	-26%
multi-speaker pool	-15%	-10%

Table 3.2: Decrease in separability when augmenting **W** covariance data with QNT. The LCE&DLCE acoustic representation degrades less than the LCE and the multi-speaker pool is more resilient than the smaller single speaker pool.

Referring to Table 3.2 we see that the decrease in value for metric, J_1 is less for the multi-speaker pool, than the single-speaker pool when going from a Q to QNT transform. Thus, any drop in performance for the quiet test condition should be less for the multi-speaker test case. We can infer from this that larger pool measurements makes the transform less susceptible to distortions and provides better cross-condition performance. However, this is difficult to prove from these tests because the separability is much higher for the single speaker pool. An appropriate test would have a large set of samples for one speaker which could be split in half.

The next question to answer is regarding the minimum amount of data for LDA. Atal [4] and Bricker et al [10] say data limitations, when there are fewer samples than dimensions, frequently result in a singular (noninvertible) covariance matrix. They prefer to pool speakers to increase the sample size. Markel and Davis [65] suggest that an adequate sample proportion is roughly ten times the number of dimensions and in the context of speaker identification, they used individual covariance matrices for each speaker because this outweighed any inadequacy caused by under sampling in the speaker statistics. In SD speech recognition, a minimum of two samples per word is

recommended. On average there are 50 acoustic vectors per word. With dimension 40 (for LCE&DLCE) we would need to pool the covariance matrices of four words (of two samples each) together to meet the minimum requirements of 400 samples. Therefore, any practical speech recognition vocabulary will be adequate for avoiding a singular matrix.

3.2.6 Contribution of the individual acoustic parameters

Column five of Table 3.1 denotes which experiments used the LCE accustic representation by itself and those that included the DLCE. The variance increases for both **W** and **B** but much more for **W**. This is true for all cases of transform type and pool size. From Table 3.2 we see that the LCE&DLCE combination is more resilient to the multi-condition pooling (QNT-**W**) than the LCE by itself (Q-**W**). Table 3.3 shows the relative change in the separability metrics (the trend is the same for the single-speaker pool). Clearly the addition of DLCE improves both Q and QNT but QNT gains the most (72%). Metric J_1 is a better indicator of separability then J_4 because it is consistent for all conditions.

LDA data (multi-speaker pool)	%Change from LCE to LCE&DLCE	
	J1	J4
Q-W, Q-B	63%	-23.0%
QNT-W, Q-B	72%	3.5%

Table 3.3: Increase in separability when increasing the acoustic representations from 20 LCE to 40 with DLCE.

Tables 3.1 to 3.3 give us the complete picture of the variance while Figure 3.3 illustrates the variance across the frequency spectrum. Referring back to Figure 1.7 we see that the correlation of adjacent subbands is high. This is because the filters are overlapping and triangular in shape. In Figure 3.3 we see that below subband 12 (1.5 KHz) the variance is distributed differently for the multi-speaker pool than the single-speaker pool. This is due to the difference in fundamental frequency (pitch) between speakers. High covariance between non-adjacent subbands occurs because of harmonics of the pitch frequency. Bandwidth limitations of the audio recording will reduce the variance for the lower subbands and the highest subband.

The following observations have been made:

i) Within-class variance increases in all subbands when the covariance matrix is augmented with noise in the QNT process. The resulting variance is significantly higher for the lower frequency subbands (0 to 5) which span 100 Hz to 500 Hz. Subband zero now has the highest variance, subband three held this distinction in the Q covariance case. This shift occurrs in both the multi-speaker and single-speaker pool.

ii) The DLCE is less affected by adding noise to the within-class variance than is the LCE. This is more obvious in the multi-speaker case. This explains why DLCE is robust to noise.

iii) Despite the much higher **W** variance that DLCE introduces, the increase in the **B** variance results in better class separability.

iv) Between-class variance is highest in the region of 100Hz to 500Hz (600 Hz for multi-speaker). Between-class variance is relatively flat from 600Hz to 1500Hz and then falls off with increasing frequency. This corresponds to the long-term PSD of Figure 1.3.

v) In the multi-speaker case the within-class variance exceeds the between-class scatter above subband 14 (2 kHz). Consequently, the scatter ratio expressed in J_1 and J_4 is much less than unity and discrimination between words is expected to be low for the upper five subbands. Features that fall into this category will not be useful for speech recognition. In the single-speaker case the within-class variance does not exceed the between-class and the scatter ratio only approaches unity in the last two subbands (above 3.5 kHz).

vi) The 2-pass alignment helps all subbands by decreasing **W** and increasing **B**.

Observation v) is important because it implies that the upper subbands do not contribute to discrimination and are thus irrelevant. We expect that the eigenvectors will assign minimal weights to these subbands. Thus, a form of noise shaping takes place.

3.2.7 Dimensionality reduction

A chief asset of LDA is the dimensionality reduction that it provides when transforming the acoustic space to the feature space. Recall from Section 1.5 that the

rows of the IMELDA transformation matrix are constructed by selecting the eigenvectors with the largest eigenvalues (i.e. the F-ratio, **B/W**) and discarding the remainder. Importantly, this action of dropping the eigenvectors of the covariance matrix with the smallest eigenvalues succeeds in suppressing statistical noise, thus improving speech recognition. Now we are faced with resolving the question of how many eigenvectors to retain. The following procedure is recommended.

i) Rank the eigenvectors by variance and then eliminate those eigenvectors that have variance less than 1.0 since they represent parameters where the transformed **B** is less than **W** [18].

ii) Retain those eigenvectors whose total variance satisfies a minimum criterion. Bocchieri obtained optimal recognition performance with a 90% to 95% criterion [8].

iii) Consider the tradeoff between performance and real-time requirements (processing and memory) when lowering the number of eigenvectors.

iv) If possible, conduct recognition experiments to ensure the best choice or use the confusion matrix method.

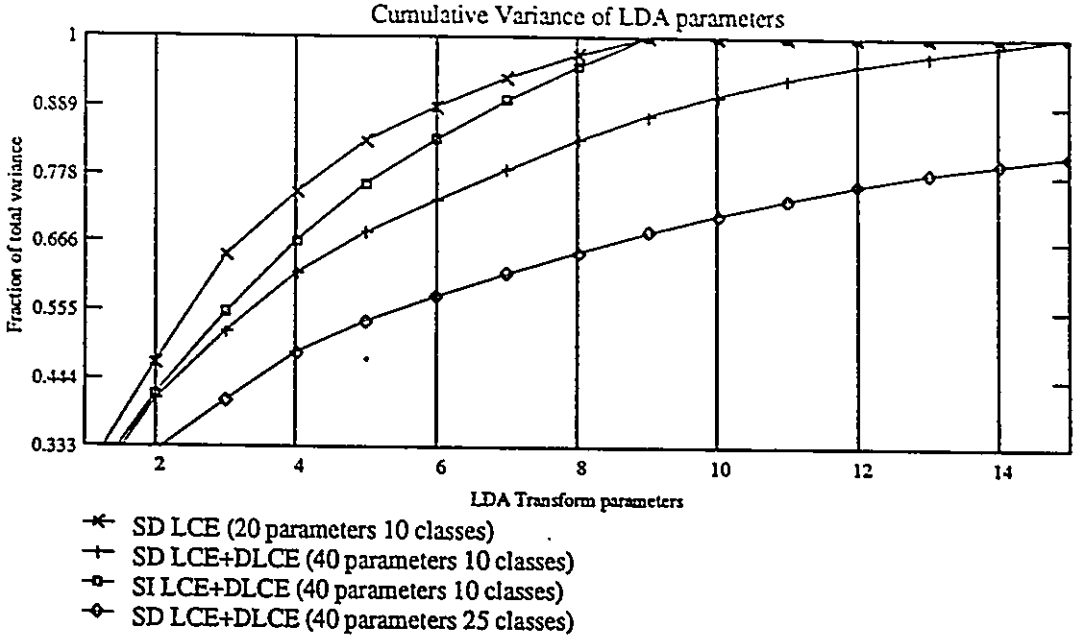


Figure 3.7 Cumulative fraction of total statistical variance as a function of the number of selected LDA transform parameters (eigenvectors).

No. of classes	No. of LDA input parameters	No. of LDA output parameters			
		90%	95%	100%	used
10 (SD)	20 (LCE)	6	8	9	10
10 (SD)	40 (LCE&DLCE)	10	12	15	12
10 (SI)	40 (LCE&DLCE)	7	8	9	12
25 (SD)	40 (LCE&DLCE)	23	29	39	12

Table 3.4: Number of LDA output parameters as a percentage of total variance. Rightmost column reveals that in 3 of 4 cases we are not using the optimal number of parameters (90% to 95% coverage). As the number of classes increases more LDA transform parameters are required.

The fraction of total statistical variance as a function of the number of selected IMELDA features is graphed in Figure 3.7. Table 3.4 contains values from selected points on the graph. When the vector of LDA input parameters is comprised solely of the

LCE representation (dimension 20), the selection of only eight eigenvectors represents 95% of the total variance. Thus, a 60% reduction can be obtained. When the number of input parameters is doubled in the LCE&DLCE case, 12 eigenvectors (only eight for the SI case) are required to retain the same degree of variance. As the number of classes increases from 10 to 25, we find that 29 eigenvectors would have to be retained to maintain the same level of statistical variance. It would appear that the number of LDA outputs is proportional to the number of classes. However, we reserve judgment on making this conclusion for the following reason. This case only used two samples per class while the other used 15 samples and thus the classes may have been under represented leading to inaccurate covariance estimates.

In the IMELDA system we currently retain 10 LDA outputs to represent LCE and 12 LDA outputs for LCE&DLCE input parameter combinations. According to Table 3.4 we should be using fewer LDA outputs for the 10 class case and more for the 25 class case. Therefore, we recommend that the procedure outlined above be employed in the training phase of the IMELDA speech recognition system.

3.2.8 Experimental set-up and baseline results

Several speech recognition experiments were conducted using the various alignment techniques, pool sizes and LDA data types that are described above. For the sake of brevity only a subset of these experiments is presented in Table 3.5. To make comparisons with the analytical expectations easier this table uses the same format as Table 3.1. The speech was recorded in quiet with a studio quality microphone. Test data

consists of four types, the same three used in the QNT covariance matrix also broadband noise with a 9 dB SNR. Results are tabulated separately and the cumulative results for quiet, 15 dB SNR and tilt are shown in a separate column.

In the SI test there are 10 male speakers with a vocabulary of digits spoken in triplets. The speakers were averaged together for the reference material with 12 samples per digit per speaker, 1200 samples in total. There are 30 test digits per speaker, 300 test digits in total. The test samples were not included in the reference material. Results are reported as percentage error for the accumulated results. For the SD tests, two male speakers were selected from the pool of 10. Each speaker had the LDA transform and reference templates computed from 15 samples per digit, 150 in total. The test data were 75 digits that were from the reference material.

The QNT within-class covariance data was created from artificially generated 15 dB SNR white noise, 6 dB/octave tilt and the quiet data. The covariance data did not include the 9dB SNR white noise. The LCE&DLCE transform is composed of the eigenvectors with the 12 highest F-ratio terms while the LCE transform has 10 terms.

	pool	align	LDA data	acous. reps		J_1	Test Results (% error)				
							quiet Q	15 dB N	tilt T	tot. QNT	9 dB
a5	SD 150	2-pass	QNT-W,Q-B	LCE&DLCE		62	0.00	2.00	0.67	0.89	27.0
a6	SD 150	2-pass	QNT[1]-W,QNT[1]-B	LCE&DLCE		54	0.00	5.00	0.00	1.67	18.0
b1	SI 1200	1-pass	Q-W,Q-B	LCE		24	3.33	14.33	67.00	28.22	48.7
b2	SI 1200	2-pass	Q-W,Q-B	LCE		30	2.00	10.67	70.33	27.67	34.0
b3	SI 1200	2-pass	QNT-W,Q-B	LCE		13	3.66	7.00	4.67	5.11	31.3
b4	SI 1200	2-pass	Q-W,Q-B	LCE&DLCE		29	0.66	1.33	20.33	7.44	13.0
b5	SI 1200	2-pass	QNT-W,Q-B	LCE&DLCE		26	0.33	2.00	1.67	1.33	15.7

Table 3.5: Comparison of recognition performance for various LDA transforms. Rows designated by 'a' are SD with 150 training samples and 75 test samples and the results are averaged for 2 speakers. Rows marked by 'b' are SI with 1200 training samples from 10 male speakers, test with 30 separate samples from these speakers and the results averaged for the test speakers. [3] [4]

Conclusions from these tests are summarised below:

i) From rows b1 and b2 we see that a 2-pass alignment procedure helps recognition performance. This is in accordance with metric J_1.

ii) Comparing rows b3 and b5 it is clear that increasing the dimension of the acoustic representation from 20 to 40 with the addition of the DLCE, helps recognition in all conditions. This advantage was predicted by studying metric J_1.

---

[3] Results for the tilt condition in tests b1 and b2 are so bad that comparisons are not justifable.

[4] The LCE and QNT-W used in experiments b1, b2, b3 differ from those reported elsewhere in this report, therefore they should be considered for their relative performance.

iii) Augmenting the **W** matrix with samples of additive noise and convolutional noise (the QNT process) helps dramatically for the tilt test case and quite significantly for the noise case, but the quiet rate decreases (as per row b2 versus b3). In tests which are not shown in the table, we found that as we sought to improve the tilt test case the noise results decreased (see b5 and b4).

iv) The large overall increase in performance for the QNT process contradicts the decrease in separability reflected by both J_1 and J_4. This can be explained by the fact that while the trace of **W** increased to reflect the added variance, **B** did not follow suit because it is computed solely from quiet. Consequently, the ratio of **B** to **W** decreases. Experiments with adding noise to **B** (row a6 of Table 3.1) were unsuccessful, because the ratio become even lower. Therefore, J_1 and J_4 can only be considered relative to the quiet reference condition, consequently they are incapable of predicting the perfomance of a QNT transform.

Observation iii) is a typical problem in this field. It is difficult to obtain the same level of performance across all conditions. There is usually a tradeoff involved. In overall performance the QNT methodology is superior.

3.3 Conclusions

In this chapter we described why IMELDA works well in noise and we showed areas where improvements could be made. Four factors were shown to have a significant bearing on the robustness of an IMELDA transform:

i) Providing samples of the distortions encountered under test can dramatically improve performance under these conditions. This is particularly noticeable with channel type noise. The process for creating the QNT transform was described in detail.

ii) The dynamic acoustic representation imparts a resilience to additive type noise. This representation was also shown to help in distinguishing speech sounds and it was particularly effective in the multi-speaker pool. It was demonstrated to be more competent than it's static counterpart for both cases.

iii) Solving one type of noise problem is usually done at the expense of one of the other conditions.

v) Large pool measurements make the transform less susceptible to distortions and the result is better cross-conditional performance.

Our conclusion from this is that the 2-pass QNT type IMELDA transform of LCE and DLCE acoustic representations provides the best overall recognition rate when the noise is known before hand.

Different methods of predicting the performance of an IMELDA transform were compared with actual results from speech recognition experiments. Metric J_1 was shown to be a more reliable indicator than J_4 but neither is capable of predicting the performance of a QNT transform. A new method for predicting recognition was proposed and this was based on a confusion matrix of dissimilarity scores. Interpretation of the confusion matrix proved to be consistent with the experimental results.

Some practical concerns of applying LDA in a speech recognition system were addressed. In summary:

i) The cosine transform is used in the first step of the LDA and the resulting transform is re-applied in a second pass. This 2-pass alignment method improves separability.

ii) There will not be any problems with singular matrices if at least eight samples are provided in the covariance matrix.

iii) Eliminating low variance eigenvectors reduces statistical noise and improves recognition. A procedure was derived for determining the optimal number of LDA output eigenvectors.

iv) Spectral thresholding was shown to be a form of noise-shaping. Low threshold values are used in the training process and much larger values can be used under test conditions.

CHAPTER 4

AN INVESTIGATION INTO SUBBAND FILTERING

4.1 Introduction

We have previously shown that slowly stationary noise is low frequency in content, compared to speech, and that this noise can be suppressed by high-pass filtering. In this chapter, we evaluate the effectiveness of subband filtering the outputs of the Mel-scale filterbank used in IMELDA (refer to Figure 1.3). Several different forms of subband filtering were described in Chapter 2, the options are repeated below:

i) Filtering type: high-pass and band-pass versus spectrum normalization.

iii) Choice of domains: linear for removal of additive type noise or the log domain for convolution type noise.

Section 4.2 compares different filtering formats and presents several practical issues that were encountered during the course of this investigation. Conclusions drawn in this section are followed through in the remaining sections. A technique that is similar to high-pass filtering but with a long time constant is called spectrum normalization or *mean removal*. Analysis and discussion of this technique are presented in Section 4.3.

A series of speech recognition evaluations on subband filtered speech is covered in Section 4.4. The objectives are outlined below:

i) **Reducing the need for apriori information:** Unlike the QNT IMELDA transform, subband filtering does not require samples of the acoustic noise. This removes or at least reduces the need for apriori information during the training stage. As shown in Section 3.2, augmenting the transform with samples of spectral tilt (i.e. QNT) significantly lowers the number of errors in the SI test. To decide whether subband filtered speech can help, we use the quiet-data only transform (i.e. Q) and compare results. We test on quiet, tilt and noise. Theoretically, log domain subband filtering should not help in additive noise, therefore, we rely on the dynamic acoustic embodiment to look after this component. Our main concern is the tilt test, we wish to have a robust acoustic representation that does not require apriori measurements. This could eliminate the need for the QNT transform.

ii) **Reducing between-speaker variability:** The normalization property of subband filtering is reportedly beneficial for reducing between-speaker differences. This is essential for real-world SI applications. Since the number of errors was so low in the quiet test reported in Chapter 3 we introduced a set of speakers from outside the training pool. We then compare performance on quiet data, with and without the subband filtering.

iii) **Inhomogeneous reference and test.** We test this hypothesis by using a microphone and recording channel for a reference environment that differs significantly from the test condition. Helicopter speech is used for this in the SD

mode. Finally, we conduct SI recognition experiments in the helicopter. This last experiment is the most challenging case of mismatched reference and test.

4.2 Analysis of subband filters

High-pass filtering has been shown to alleviate the effects of convolutional noise when applied in the log domain and additive noise when applied in the linear domain. Band-pass filtering is subject to the same constraints, but it does offer an advantage in another area. The low-pass effect of the BPF can smooth some of the analysis artifacts of the windowing operation. Hermansky [38] describes these artifacts as being caused by the frame-to-frame spectral changes in the short-term spectral estimate (see Section 1.4). In this chapter we evaluate both types of filtering.

The transfer function for the IIR-type HPF recommended in the literature is shown in (4.1). The filters are characterised by a sharp spectral zero at 1.0 and a single pole, ρ.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1 - z^{-1}}{1 - \rho z^{-1}} \qquad (4.1)$$

In our thesis investigation we simulated the two published filters with poles at 0.7 and 0.95 and we compared the spectral response with higher order IIR filters (Butterworth, Chebyschev, Elliptic) with different rolloff and passband characteristics [84]. A summary of our findings is provided at the conclusion of this section.

High-pass FIR filters of order N are expressed in equation (4.2). Factor β was set to 0.94 for a 2.0 Hz cutoff and we chose N=16 according to the literature.

$$H(z) = \frac{Y(z)}{X(z)} = 1 - \frac{\sum_{n=1}^{N} \beta^n z^{-n}}{\sum_{n=1}^{N} \beta^n} \qquad (4.2)$$

BPF IIR filters of order N were implemented as shown in equation (4.3). The pole position is set by ρ, and in two experiments, we used value 0.90 as advocated by Hanson [34] and 0.94 used by Hermansky [38]. Zero locations are specified in the numerator of the BPF and we used N=5 and α=0.1, in agreement with both Hanson and Hermansky.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\alpha \sum_{n=0}^{N-1} [n - \frac{N-1}{2}] z^{-n}}{1 - \rho z^{-1}} \qquad (4.3)$$

The numerator of the BPF is a regression filter, like the one used in the dynamic acoustic representation (1.4). Thus, it is not difficult to see the relationship between subband filtering and dynamic acoustic analysis. By applying dynamic analysis to the subband filtered representation, we have in essence a second order dynamic acoustic representation, much like the 'acceleration' feature discussed in Section 2.4.2.4.

In our simulations with speech signals, we found two issues that must be dealt with carefully. First, the high-pass filtering operation highlights differences in the spectrum and this can cause negative valued results. Secondly, the filters have a finite response time where artifacts from previous speech or noise remain in the filter. We call this the transient effect, the duration of which can be found by inspecting the filter's impulse

response. A summary of the issues encountered in the design and implementation of subband filters is given below.

i) **Transient response**: The longer the transient response of the filter the better the integrity of the speech, but there must be a sufficiently long interval between words. A filter with a short transient response time can accommodate smaller gaps between words, but there is the possibility of removing too much speech information. This tradeoff is a major concern for connected word speech recognition.

ii) **Linear domain:** This domain is suitable for the removal of additive type environmental noise (2.2). Negative numbers resulting from the filtering operation must be dealt with in a careful fashion prior to the log deconvolution. Methods such as squaring or full-wave rectification of the signal are adequate solutions.

iii) **Log domain**: Convolution type environmental noise (2.3) can be suppressed in this domain. The HPF operation is sensitive to small perturbations caused by variance in the noise floor. Therefore, spectral thresholding must be applied prior to filtering.

iv) **IIR filters:** The recursive nature of IIR filters requires a careful analysis of the transient response. Filters with poles close to the unit circle require a longer time for the filter to stabilize. Poles further from the unit circle stabilise faster.

v) **FIR filters:** Since these filters are non-recursive, the transient response is simple to figure out (finite duration of length N). High orders are required to achieve a flat passband. However, as filter order increases, the transient increases in duration and this is undesirable. Stop-band attenuation is not as important as the transient response.

vi) **Filter initialisation:** In order to ensure that speech and non-speech artifacts are removed from the filter, it is necessary to carefully initialise these filters. We initialise the filter by applying a sample of the environment and iterating this for the duration of the filter's impulse response.

vii) **Reference models:** Unlike the conventional static acoustic representation, subband filtered speech is influenced by preceding samples. This is a crucial requirement for reducing channel distortion, because the high-pass filtering essentially removes the average that was taken over several speech samples. A side effect is that the reference models are now dependent on the sequence of speech in the training condition. This issue is not described in the literature because the published systems were tested on words spoken in isolation. For

connected-word recognition systems like IMELDA, we recommend multi-style training, such that reference models reflect both isolated speech and connected speech styles.

The duration of the impulse response was the deciding factor when comparing filters that had similar frequency characteristics. For our speech recognition experiments, we decided to proceed with simulations for two BPF type filters with poles at 0.90 and 0.94 (4.3) and the 16th order FIR type HPF (4.2). These three filters are reported in the literature to have the best results (at least for PLP subbands at 10.0 ms sampling rate). Table 4.1 summarises the filter characteristics. Frequency response and impulse response curves for the three filters are shown in Figures 4.1 and 4.2 with our 6.4 ms (156 Hz) sampling rate. All three subband filtered representations were augmented by their dynamic embodiment for the speech experiments.

Filter	Type	3 dB high-pass (Hz)	Transient response (samples)
HPF (N=16)	FIR	2.6	16
BPF (ρ=0.90)	IIR	2.35	42
BPF (ρ=0.94)	IIR	1.45	52

Table 4.1: Table of filter characteristics for the three filters studied

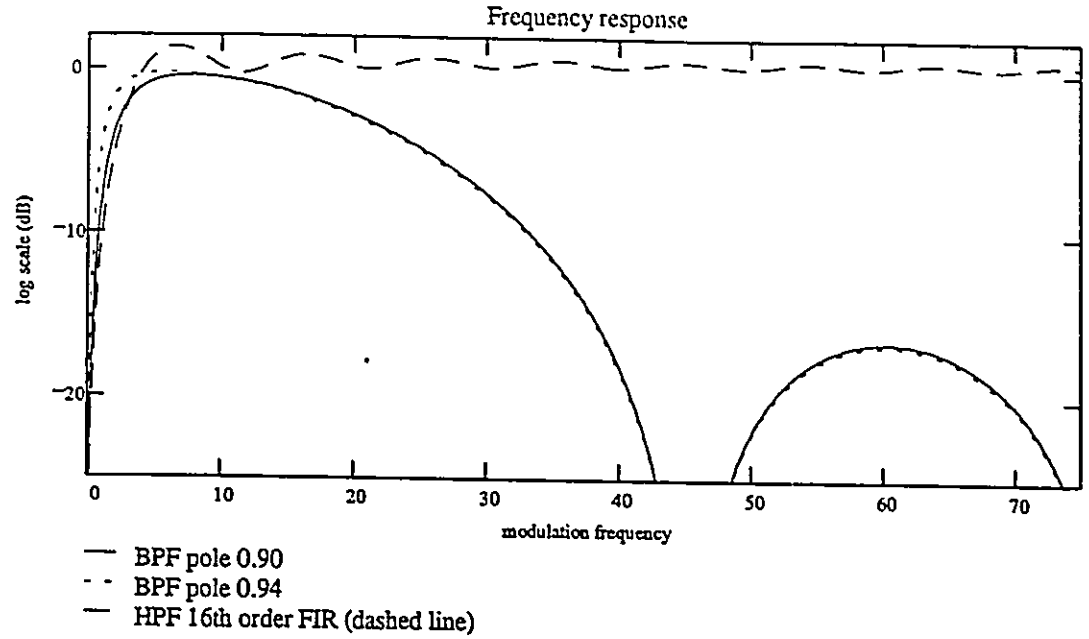


Figure 4.1.1 The two BPF have similar passbands while the HPF exhibits ripple in passband. All three have their cutoff frequency below 3 Hz as per the literature.

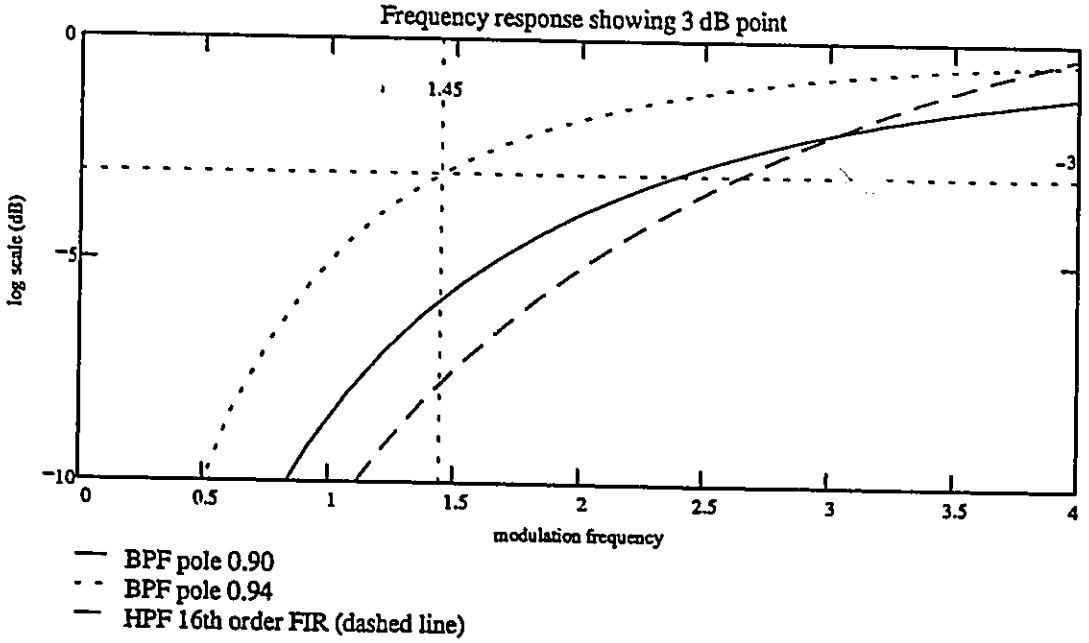


Figure 4.1.2 The 3dB point for the filter with the best recognition performance (BPF with pole at 0.94) is at 1.45 Hz.

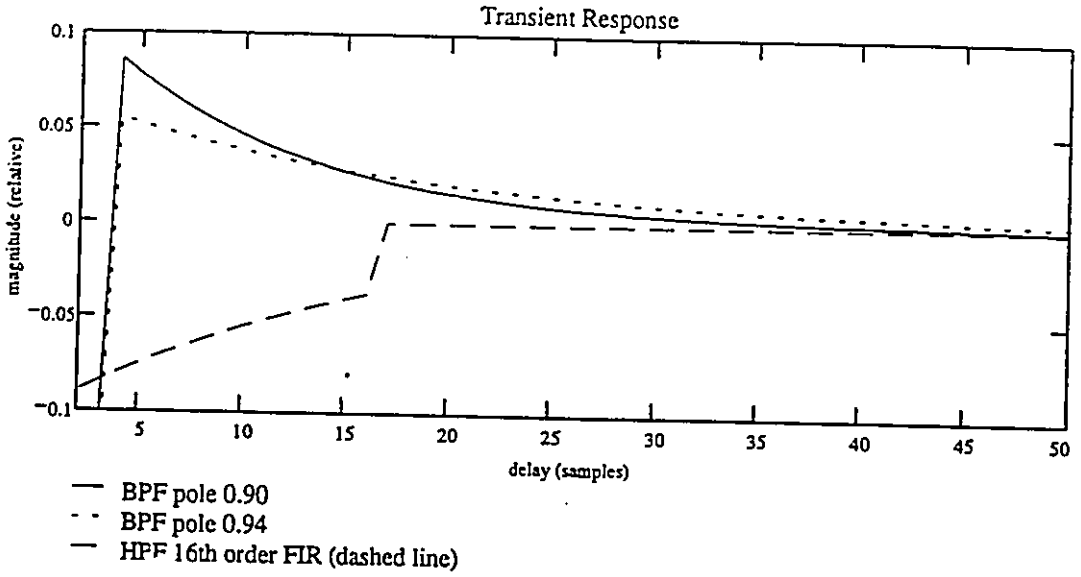


Figure 4.2: The FIR filter has the shortest duration transient. Of the two BPF type filters the one with pole at 0.90 has the best (shortest).

4.3 Spectrum normalization

The spectrum normalization technique described in Section 2.4.3.5 was implemented as a one second LPF according to the description in [27]. The filtering operation is shown in equation (4.4). This is computed by subtracting a mean estimate, <y> of previous samples, from the current sample. This is a high-pass filtering operation. The mean estimate is obtained via LPF filtering, where β and the number of samples, N decides the weighting function and time constant respectively. With a sampling rate of 6.4 ms, N is set to 150 and β to 0.95.

$$y(k) = y(k) - \langle y \rangle \qquad (4.4.1)$$

$$\langle y \rangle = \frac{\sum_{n=1}^{ij} \beta^n y(k-n)}{\sum_{n=1}^{N} \beta^n} \qquad (4.4.2)$$

Our simulations revealed a speech spectrum that is similar to the static LCE representation, however, the filter leaves a residue that affects the next speech sample. This is a consequence of the gliding nature of this method. The original researcher, Haeb [30] now advocates computing the sample mean over the entire utterance instead of using an LPF for the estimate (i.e. mean removal). Thus, N would be the length of the detected utterance and β would be set to 1.0, turning equation 4.4.2 into the arithmetic mean. The mean is subtracted from each frame of the utterance. An improvement in error rate of 20% for a SI evaluation was reported by Haeb. In a personal communication with Haeb [31], it was said that this method performs better than the one second moving estimate. However, the drawbacks of this technique are:

i) It is not suitable for real-time because it introduces a delay that is the length of the utterance.

ii) The technique does not discriminate between speech and silence and this is a problem. Haeb suggested that a reliable speech detector for marking the utterance boundaries would help with this dilemma.

iii) It remains to be seen how well this works with additive noise, since the performance gains were shown for a quiet database and subtraction in the log domain is valid only for convolution type noise.

Recent publications have shown mean removal techniques to be better than high-pass filtering when it comes to suppressing channel noise [3], [77]. Although, in terms of all-around performance, band-pass filters are reportedly better than mean-removal [38]. Therefore, we did not proceed with mean removal.

4.4 Evaluating subband filtering

4.4.1 Removing the need for apriori knowledge of test conditions

In this section, we investigate whether subband filtering can eliminate the requirement for apriori knowledge of the test environment. The three subband filter types described above were evaluated, first on a small-scale SD test and then on a SI database. The experiments are summarised in Table 4.2.

	pool	align	LDA data	acoustic reps		J_1	Test Results (% error)				
							quiet Q	15 dB N	tilt T	total QNT	9 dB
a5	SD 150	2 pass	QNT-W,Q-B	LCE&DLCE		53	0.00	2.00	0.67	0.89	27.0
a7	SD 150	2 pass	Q-W,Q-B	LCE&DLCE HPF 2 Hz		38	0.00	18.00	0.00	6.00	52.7
a8	SD 150	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		67	0.00	1.33	3.33	1.55	11.3
a9	SD 150	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.90		65	0.00	2.00	4.00	2.00	11.3
b4	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE		29	0.66	1.33	20.33	7.44	13.0
b5	SI 1200	2 pass	QNT-W,Q-B	LCE&DLCE		26	0.33	2.00	1.67	1.33	15.7
b7	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE HPF 2 Hz		16	3.33	17.00	3.66	8.00	40.0
b8	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		26	1.33	2.00	5.66	3.00	12.3

Table 4.2: Comparison of subband filtering recognition performance. Rows denoted by 'a' are SD with 150 training samples and testing on 75 of these samples, results are averaged for 2 speakers. Rows denoted by 'b' are SI with 1200 training samples from 10 male speakers and tested on 30 separate samples from each speaker, results are averaged for the 10 speakers. [1] [2] [3]

BPF type subband filters are superior overall to the HPF in both SD and SI recognition. This is shown in test results a7 versus a8 and a9 for SD and row b7 versus b8 for SI. The BPF with pole position at 0.94 performed the best of the three and so we continued to use this filter in later experiments. We attribute the success of this particular filter over the others to it's stop-band characteristics. Table 4.1 indicates that this filter

---

[1] Inappropriate templates for 'eight' caused insertion errors that were not included in the Q and T test of experiments a7 and b7.

[2] Rows a5, b4 and b5 are repeated from Table 3.5 for comparison purposes.

[3] The same microphone and recording conditions existed for training and test.

had the lowest 3 dB cut-off frequency at 1.45 Hz. In terms of transient response, it was not the best (Figure 4.2). However, by increasing the level of spectral thresholding we were able to reduce the input variance and thus decrease the transient effect from 52 to 45 samples. We did not optimise the pole position in terms of recognition rate because this requires testing on several databases, but we do recommend this course of action.

Under homogeneous test conditions (i.e. reference and test have same speaker pool, and microphone) we do not expect subband filters to attain the same level of performance as the QNT. The reason for this is that the QNT IMELDA transform contains samples from the conditions under test. Thus, we set the QNT as the standard when comparing subband filtering implementations. Row a5 then represents the goal for SD and row b5 for SI. Note that a quiet-data only transform (Q) was used for the subband filter tests in order to separate the contributions of filtering from the LDA. Previous results obtained from non-filtered data is included for comparison purposes (row b4).

The SI test on spectrally tilted speech (b4 @ 20.33% error) had the most to gain from the QNT transform (b5 @ 1.67% error). Subband filtering produced significantly better results (b8 @ 5.66% error) than Q. We attribute this result to the property of tilt being additive in the log domain and it is removed by the filter's stop-band suppression below 2 Hz. Even though subband results were not quite as good as the QNT, it did this without any measurement from the environment.

For the same cross-conditional performance, subband-filtering of the acoustical representations, is better at separating the classes than using the QNT transform (see

J_1 in Table 4.2). Subband filtering with the Q transform (row a8, a9) is better than QNT (row a5) for SD and equal in SI (row b5 and row b8). Comparing this with a non-subband filtered Q-transform, the separability measure is lower though (26 versus 29).

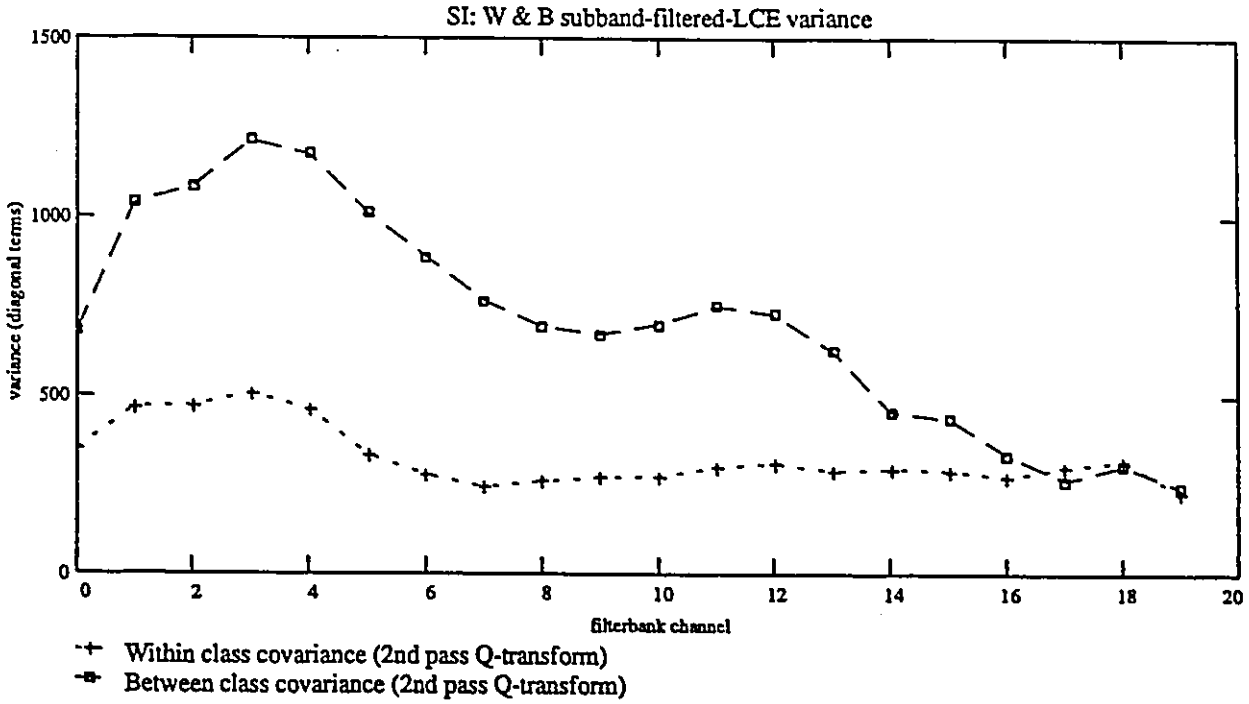


Figure 4.3 The ratio of **B** to **W** is positive for all of the subband filtered channels except 17 through 19. This is an improvement for 15 and 16 compare to Figure 3.3.1

Figure 4.3 depicts **W** versus **B** for the filtered subbands. The overall variance is lower than the conventional method (Figure 3.3.1) but one advantage is that the positive ratio (i.e. $J_1 > 0$) of **B** to **W** is extended from channel 14 to 16. We then expect the two additional filters to be contributing more to the discrimination.

The most encouraging news from these tests is that the cross-conditional performance (column labelled "tot. QNT") of subband filtered acoustic representations is

quite good. It appears that the dynamic acoustic representation has helped the additive noise case and so we can successfully use the log domain for the subband filter implementation. Interestingly, other researchers have reported the chimera between linear and log domain implementations for additive and convolution type noise respectively. The only remark that we can make to this point is that we use LDA which finds the most effective combination of subband filtered static (LCE) and dynamic (DLCE) acoustic representations.

In conclusion, we can employ subband filtering in the log domain to obtain significant improvement in convolution type noise (spectral tilt) and still maintain the good additive type noise performance which we believe has been afforded by the dynamic acoustic representation.

4.4.2 Reducing between-speaker variability

In this section, we investigate the claim that subband filtering is able to reduce between-speaker variability. Researchers believe (refer to Section 1.4) that for speech perception, the relevant parts of the modulation frequency spectrum are between 10 and 25 Hz, while speaker specific information dominates for frequencies below 10 Hz [39]. We made an attempt to verify this by computing the subband frequency content for two male speakers and compared this with a conglomerate of 10 male speakers as reported in Section 2.4.3.5. In Figure 2.8 we saw that the dominant peaks are between 2.5 Hz to 3.5 Hz for the three cases. It is not obvious that filtering this frequency range would

provide any gain, but we did procede with a set of experiments in order to confirm or deny the claims made in the literature.

We extend the tests reported in Table 4.2 by introducing a set of speakers from outside the pool to create a true SI test. We performed two cases: i) female speakers ii) male speakers. We then compare performance on quiet data with, and without the subband filtering.

	pool	align	LDA data	acoustic reps		J_1	Test Results (% error)
							quiet Q
g1	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE		29	23.67
g2	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		26	23.33
h1	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE		29	4.84
h2	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		26	4.92

Table 4.3: Between-speaker variability experimental results. Rows designated by 'g' are SI and trained with 1200 samples from 10 male speakers and tested on 60 samples per 5 female speakers, results averaged for the test speakers. Rows designated by 'h' use the same reference as 'g' but testing on 156 samples per 9 male speakers. [4] [5]

It appears that subband filtering offers no improvement over the conventional method. Of the five female speakers in experiment 'g' the error rate was uniformly bad

---

[4] In tests g1 and g2 the test speakers were from outside the training pool and they were female. The same microphone was used for the training and test sets.

[5] Tests h1 and h2 used male test speakers from outside the training pool. The recording conditions were similar (anechoic chamber) but the microphone was different, a close fitting head-mounted Shure SM-10 for test and a boom microphone for training.

(a sixth one was removed from the pool since errors were greater than half). We may attribute this to the fact that the difference in pitch for females and males (refer to Section 1.3) results in word patterns that differ so much in formant location (Table 1.1) that we could not expect good rates whatever the analysis method. In test 'h' we see that speech recognition drops by more than 4% (compared to row b4 in Table 4.2) when speakers from outside the training pool are used. Although the errors were different between the two methodologies, the average results are nearly identical. The male SI tests contained a component of inhomogenous testing in that the microphony used in tests h1 and h2 differed between reference and test. If this was contributing to the error rate, then subband filtering was not helping to reduce this.

In conclusion, we are unable to claim any improvement for subband filtering in dealing with difference between test and reference that is attributed to speaker differences.

4.4.3 Evaluations with inhomogenous reference and test

Previous sections have shown that subband filtering works well for channel noise that has been artificially generated (i.e. the spectral tilt test). In this section, we examine a practical real-world problem: different microphones between reference and test. In a broader sense we carried the idea of inhomogeneous or mismatched reference and test to a SI study in the helicopter. The SI helicopter study is extremely challenging because of the mismatches listed below:

i) Like Section 4.4.2 this is a true examination of SI.

ii) Channel difference between the boom microphone and the helmet-mounted noise-cancelling type. A greater difference than that used in Section 4.4.2.

iii) Recording channel difference between the anechoic chamber and the bandwidth limited (and noisy) avionics communication channel.

iv) Several additive noise components exist in the helicopter. These are listed in Table 2.2 and illustrated in Figure 2.4. Segmental SNR's ranged from 25 dB to 17 dB and these are much lower than the reference SNR.

The QNT transform used in the experiment was not optimised to the helicopter noise, rather it was the same transform described earlier. We felt that it was too difficult to model the channel noise and there are several additive noise cases to consider. Furthermore, the two noise types occur simultaneously whereas in the earlier tests we separated the two. This makes subband filtering more desirable because the channel and noise do not have to be modelled.

Results from helicopter experiments are tabulated in Table 4.4. Subband filtering makes a significant contribution to speech recognition in the helicopter for the SI mismatched case. An overall improvement of 9% is obtained by subband filtering (e3 versus e2). To further prove these results we extended the worst case scenario (HV and CR) to two more speakers and the advantage became even greater 38% (e5 versus e4). Interestingly the QNT test (e2) was 4% better overall than the standard transform (e1).

	pool	align	LDA data	acoustic reps		J_1	Test Results (% error)				
							25 dB GQ	24dB I70	19 dB HV	17 dB CR	total
d1	SD 50	2 pass	Q-W,Q-B	LCE&DLCE		259	3.63	2.38	5.44	10.24	5.4
d2	SD 50	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		299	6.60	3.40	5.78	8.53	6.1
e1	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE		29	12.82	23.07	19.87	27.33	20.8
e2	SI 1200	2 pass	QNT-W,Q-B	LCE&DLCE		26	3.84	19.87	17.95	25.36	16.8
e3	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		26	3.20	5.77	10.90	11.59	7.9
e4	SI 1200	2 pass	QNT-W,Q-B	LCE&DLCE		26	n/a	n/a	50	50	50
e5	SI 1200	2 pass	Q-W,Q-B	LCE&DLCE BPF 0.94		26	n/a	n/a	10.70	15.00	12.8

Table 4.4: Comparison of helicopter recognition performance. Rows denoted by 'd' are SD with 2 samples per word (25 words) in quiet (GQ)[6]. Test on 294 connected-word command strings for 1 helicopter pilot. Rows denoted by 'e' are SI with 1200 training samples from 10 male speakers and tested on 150 digits (in triplets) for 1 helicopter pilot[7]. Tests e4[8] and e5[9] were conducted on two more helicopter pilots (averaged) where only the HV and CR test conditions were available.

SD recognition test results in the helicopter (rows d1 and d2) did not show a clear advantage for subband filtering. In fact, performance was worse in three of four

---

[6] In test d2 one half of the errors were substitutions of the word 'feet' by 'eight'. The reference pattern included a 't' release sound which was not found during test conditions. Subband filtering highlights this event more than regular processing.

[7] In tests e1 through e5, the test speakers were from outside the training pool. The recording conditions were quite different, an anechoic chamber with a boom microphone was used for training providing an SNR of about 23 dB. Test recordings were from the bandwidth limited helicopter avionics communication system with a close fitting helmet-mounted M-1/DC amplified dynamic microphone.

[8] Errors for the e4 test exceded 50% and were too difficult to properly tabulate.

[9] One half of the errors in test e5 were the deletion of digit 'four'. This may be due to the mismatch in microphone bandwidths which affects the 'f' fricative.

conditions. The following explains this dilemma and highlights two inherent characteristics of subband filtering:

i) Although there is a significant mismatch (see noise levels in Figure 2.4) between training and test, it is confined to additive noise only. In the SD test, training was conducted in the helicopter with the avionics communications channel, and so there are no channel differences to report. This precludes any advantage that log domain subband filtering may offer.

ii) More than one-half of the recognition mis-classifications were due to words with final *t-release* sounds (see Section 1.2.2). For example 'feet' was often mistaken for 'eight' because of the difference between speaking style during training and test. This is a common obstacle in speech recognition and was previously described in Section 1.5.3. We believe subband filtering exasperates the problem because of its sensitivity to short-term perturbations as manifested in the plosive *t-release* sound.

By placing more attention on the training of subband-filtered speech patterns, we believe that subband filtering is worthwhile utilising even in the SD case described above. After all, the separability measure, J_1 is higher for subband-filtered speech (299 versus 254) and the error rate is lower in the worse case condition (CR).

4.5 Conclusions

In this chapter, high-pass and band-pass type subband filters were evaluated in the subbands of the IMELDA speech recognition system. Several practical issues were described, the most serious of which are summarised below:

i) Careful attention must be given to filter response time. Artifacts from both speech and nonspeech activity can exist within a filter for several hundred milliseconds. This can cause have a deleterious outcome on a connected word speech recogniser. An initialisation procedure was described.

ii) High-pass filtering is sensitive to differences in the spectrum and this can lead to negative valued results. For linear domain applications, 'adhoc' procedures were recommended. Spectral thresholding alleviates the problem in log domain implementations.

iii) Subband filters were shown to be more sensitive to short-term perturbations than the static acoustic representation. Reference words with plosive speech sounds (e.g. 't' in 'feet' and 'eight') require special care. We suggest multi-style training be used to accommodate these type of words.

iv) Mean removal is an alternate method for dealing with channel noise but it was shown to be impractical for real-time systems such as IMELDA because of a large delay incurred before applying the normalization.

We did not search for the optimal filter since this would have taken extensive testing, but we do recommend BPF over HPF, and of the two published BPF filters, the one with pole position at 0.94 was better than that at 0.90. For the speech recognition experiments we augmented the subband acoustic representation with its dynamic counterpart. The subband filters were utilized in the log domain with the purpose of removing channel noise and reducing low-frequency speaker differences. We relied on the dynamic realisation to suppress additive noise. Linear discriminant analysis was employed to combine the two subband filtered representations. This approach was taken to alleviate the convolution/additive noise dilemma pointed out in Section 2.4.3.5.

In Chapter 3 we had shown that for SI covariance pooling, the upper four subbands do not contribute to discrimination. In this chapter we have shown that subband filtering reduced this to the upper two subbands. For SD pooling, subband filtered representations have higher J_1 metrics and consequently, improved class separability. Theoretically, we expect better speech recognition performance. For the mismatched channel case, we have shown that subband filtering significantly reduces the error rate (3.66%) over the conventional method (20.33%). Significantly, subband filtering achieved these rates without apriori knowledge of the channel. This was close to the best expected results (1.67%) that the QNT method produces. The worst mismatch with

additive noise is the SD test in 9 dB SNR. Subband filtering halved the error rate (27.0% to 11.3%).

To evaluate the influence of between speaker differences a pool of speakers from outside the training pool was used as test material. Subband filtering did not help. In an experiment to see whether gender related speaking differences could be reduced, we tested female speech (in quiet) against male reference templates. Both subband filtering and the standard Q transform produced error rates exceeding 23%. This is quite poor compared to the 4% reported for the mismatched male test speech. We believe the pitch related differences (e.g. formant locations) result in far greater mismatch than that related to low frequency modulation.

The first known assessment of subband filtering on helicopter speech was described. This test exhibited the worst case scenario of mismatch: channel, microphone, and speaker differences that can be expected. Furthermore, the four helicopter conditions tested exhibit several forms of additive noise (Table 2.2). Subband filtering stands out under these terrible conditions, with an overall error rate that is half that (7.9%) of the conventional method (16.8%). While we would not advocate flying the helicopter in SI mode, with these high error rates, we have clearly shown that subband filtering makes a tremendous difference under the most benign case of mismatch. In SD tests the mismatch is reduced to the additive noise only, consequently subband filtering was slightly worse overall. This result was partially explained by the reference model preparation.

The mismatch between microphones used in training and application is not uncommon in avionics. A speech recogniser that operates independently of the microphone is necessary. We were not able to prove this. However, we suggest an experiment where the speaker trains with a different microphone than that used in test. Here we would expect subband filtering to show a distinct advantage over the conventional acoustic representation.

CHAPTER 5

CONCLUSIONS

Noise shaping was shown to be a common theme in several different noise suppression algorithms. A preliminary investigation was conducted into root deconvolution where spectral peaks are emphasised more than the conventional log operator. We were not successful in improving IMELDA and we showed why. We do believe that this method can be employed in IMELDA if the filterbank is modified. The concept of noise shaping was also shown to exist in IMELDA in two areas: i) spectral thresholding reduces the variance in the subbands by restricting the signal to within an envelope. The envelope is dictated by both noise and speaker characteristics. ii) the QNT process lends itself to noise shaping because the subbands that are influenced by this act of adding noise to the representation are ultimately given less weighting.

In this thesis we substantiated the assertion made by others that the static acoustic feature is more readily degraded by noise than the dynamic features. Rather than excluding this feature in high noise as others would suggest, we have proven that a robust QNT IMELDA transform can provide the optimal combination for cross condition testing.

In Chapter 3 we portrayed the reasons why IMELDA works well in noise. Our conclusions and recommendations stemming from this study are:

i) Apriori knowledge of the distortions encountered under test can dramatically improve performance under these conditions. This is the basis of the QNT transform which was shown to be robust to channel type noise.

ii) The dynamic acoustic representation is the most significant contributor to improving recognition in additive type noise. This representation also helps in distinguishing speech sounds and thus improves SI recognition.

iii) It is difficult to attain high cross-conditional performance. The QNT type IMELDA transform of both static and dynamic acoustic representations provide the best overall recognition rates.

iv) Increasing the pool size of the within-class covariance reduces susceptibility to distortions and this results in better cross-conditional performance. Minimal requirements were also indicated.

v) The reason for eliminating some LDA output eigenvectors and obtaining higher recognition rates was proven to be due to a reduction in statistical noise. A procedure for establishing the optimal eigenvectors was given.

vi) Metric J_1 is better than J_4 for predicting the performance of an IMELDA transform. Neither is suited to the QNT transform and so a new algorithm, based on the confusion matrix of dissimilarity scores was developed. This proved to be a dependable alternative.

vii) For multi-speaker pools the 2-pass alignment method improves separability over the 1-pass method, and this leads to better recognition rates. The optimal number of iterations was not shown, but this is a recommended course of action.

Adding noise during the training phase reduces the mismatch problem, if the noise can be modelled. We used this algorithm in the thesis to artificially modify the quiet speech samples used in the QNT process. Previously, audio recordings had to be made in the condition of interest. This was not used in the helicopter study, because we felt that it was too difficult to model the channel, and the time varying noise.

This thesis has contributed in many ways to increase the understanding of how to use subband filters from a practical standpoint and we have shown conclusively that it can be integrated successfully into IMELDA speech recognition systems. Our conclusions are summed up as follows:

i) Filtered subbands are not like any conventional static representation. In this new paradigm, designers must be mindful of filter response time. In the case of recursive filters, non-speech and speech artifacts can exist within a filter for several hundred milliseconds. This can harm a connected word speech recogniser. Careful initialization is necessary. In connected word recognisers, multi-style reference templates are very important because subband filtering exasperates the problems of plosive sounds (e.g. 'feet' and 'eight').

ii) High-pass filtering is sensitive to differences in the input, consequently, procedures were recommended to alleviate this problem. Spectral thresholding is appropriate for IMELDA .

iii) Mean removal is reportedly good for channel noise but it is not suited to real-time systems.

iv) We did not optimise the filter design, but our test results suggest BPF are better than HPF. It is recommended that more effort be placed in this area.

v) The convolution/additive noise dilemma was resolved by augmenting the subband acoustic representation with its dynamic counterpart. The subband filters were utilized in the log domain with the purpose of removing channel noise and reducing low-frequency speaker differences. We relied on the dynamic realisation to suppress additive noise. Linear discriminant analysis was employed to combine the two subband filtered representations. This is the first known case that used IMELDA in subband filtering. Subband filtering significantly reduces the error rate due to channel noise over the conventional method and it does this without apriori knowledge of the channel.

vi) SI helicopter experiments greatly benefit from subband filtering. The overall error rate is cut in half. SD tests were less successful and this was attributed to the reference model preparation. We recommend further study.

vii) It was pointed out that microphone mismatch is a common occurrence and we believe that subband filtering can help.

viii) Small speaker pools are attractive for real-world SI applications because they are more readily obtained. Algorithms that can decrease speaker differences based on accent or gender should be pursued because small pool sizes are incapable of covering this variability. We were not able to show that subband filtering could help here.

REFERENCES

[1] A. Acero & R.M. Stern, "Cepstral Normalization for Robust Speech Recognition," Proc. ESCA Workshop on Speech Proc. in Adverse Conditions, Nov. 92, pp. 89-92.

[2] T. H. Applebaum & B.A. Hanson, "Regression Features for Recognition of Speech in Quiet and in Noise," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-91, 1991, pp. 985-988.

[3] K.T. Assaleh & R.J. Mammone, "New LP-Derived Features for Speaker Identification," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, Oct. 1994, pp.630-638.

[4] B.S. Atal, "Automatic Recognition of Speakers from their Voices," Proc. IEEE Vol. 64, April 1976, pp. 460-475.

[5] P. Alexandre et al, "Root Homomorphic Deconvolution Schemes for Speech Processing in Car Noise Environments," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-93, 1993, pp. 99-102.

[6] P. Alexandre. Personal communication, January 1994.

[7] V.L. Beattie & S.J. Young, "Hidden Markov Model State-Based Cepstral Noise Compensation," Proc. Int. Conf. on Spoken Lang. Processing, ICSLP-92, 1992, pp. 519-522.

[8] E.L. Bocchieri & G.R. Doddington, "Frame Specific Statistical Features for Speaker Independent Speech Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, No. 4, Aug. 1986, pp. 755-764.

[9] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 2, April 1979, pp. 113-120.

[10] Bricker et al, "Statistical Techniques for Talker Identification," Bell Syst. Tech. J. Vol. 50, No. 4, April 1971, pp. 1427-1454.

[11] J.S Bridle et al, "An Algorithm for Connected Word Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-82, 1982, pp. 889-902.

[12] B.A. Carlson & M.A. Clements, "Speech Recognition in Noise Using a Projection-

Based Likeli-hood Measures for Mixture-Density HMM's," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992.

[13] Y. Chen, "Cepstral Domain Stress Compensation for Robust Speech Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-87, 1987, pp. 717-720.

[14] W.W. Cooley and P.R. Lohnes, "Multivariate Data Analysis," John Wiley and Sons, Inc. 1971.

[15] J. Cupples, Rome Air Force Base, Rome, N.Y. Personal communication. May 1994.

[16] B. A. Dautrich, L. R. Rabiner & T.B. Martin, "The Effects of Selected Signal Processing Techniques on the Performance of a Filter-Bank-Based Isolated Word Recognizer," Bell Syst. Tech. J. Vol. 62, No. 5, May 1983, pp. 1311-1336.

[17] S.B. Davis & P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-28, Aug. 1980, pp. 357-366.

[18] G.R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-89, 1989, pp. 556-559.

[19] L.P. Netsch & G.R. Doddington, "Speaker Verification Using Temporal Decorrelation Post-Processing," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992, pp. II-181 - II-184.

[20] S. Dobler, P. Meyer & H.W. Ruehl, "A Robust Connected-Words Recognizer," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992, pp. 245-248.

[21] R.O. Duda and P.E. Hart, "Pattern Classification and Scene Analysis," Wiley, New York, 1973.

[22] A. Erell & M. Weintraub, "Filterbank-Energy Estimation Using Mixture and Markov Models Recognition of Noisy Speech," IEEE Trans. on Speech and Audio Processing, SAP-1, No. 1, Jan 1993.

[23] K. Fukunaga, "Introduction to Statistical Pattern Recognition," 2nd. ed., Academic Press Inc., 1990.

[24] S. Furui, "Speaker-Independent Isolated word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-34, No. 1, Feb. 1986, pp. 52-59.

[25] S. Furui, "Towards Robust Speech Recognition Under Adverse Conditions", Proc. ESCA Workshop on Speech Proc. in Adverse Conditions, Nov. 92, pp.31-42.

[26] M.J.F Gales & S. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992.

[27] D. Geller, R. Haeb-Umbach, H.Ney, "Improvements in Speech Recognition for Voice Dialing in the Car Environment", Proc. ESCA Workshop on Speech Proc. in Adverse Conditions, Nov. 92, pp.203-206.

[28] O. Ghitza, "Auditory Neural feedback As A Basis For Speech Processing," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-88, 1988, pp. 91-94.

[29] Y. Gong & W.C. Treurniet, "Speech Recognition in Noisy Environments: A Survey," Communications Research Centre, Department of Communications Ottawa, CRC-TN 93-002, 1993.

[30] H. Haeb-Umbach, D. Geller, H. Ney, "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-93, 1993, pp. 239-242.

[31] R. Haeb-Umbach. Personal communication, September 1993.

[32] J.H.L Hansen, "Morphological Constrained Feature Enhancement With Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, Oct. 1994, pp. 598-614.

[33] B.A. Hanson, T.H. Applebaum, "Robust Speaker-Independent Word Recognition using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-90, 1990, pp. 857-860.

[34] B.A. Hanson, T.H. Applebaum, "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-93, 1993, pp. 79-82.

[35] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," J. Acoust. Soc. America., 1979 pp. 1738-1752.

[36] H. Hermansky et al, "Compensation For the Effect of the Communication Channel in Perceptual Linear Predictive (PLP) Analysis of Speech", Proc. European Conf. on Speech Commun. and Technology, EUROSPEECH-91, 1991, pp. 1367-1370.

[37] H. Hermansky, N. Morgan, H.G. Hirsch, "Recognition of speech in Additive and Convolutional Noise Based On RASTA Spectral Processing," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-93, 1993, pp. 83-86.

[38] H. Hermansky, N. Morgan, "RASTA Processing of Speech," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, Oct. 1994, pp. 578-589.

[39] H.G. Hirsch, P. Meyer, H.W. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes," Proc. European Conf. on Speech Commun. and technology, EUROSPEECH-91, 1991, pp. 413-416.

[40] W.J. Holmes & D.J.B. Pearce, "Sub-Word Units for Automatic Speech Recognition of any Vocabulary," GEC Journal of Research, Vol. II, No. 1, 1993, pp. 49-59.

[41] M.J. Hunt, J.W. Yates, J.S. Bridle, "Automatic Speaker Recognition For Use Over Communication Channels," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-77, 1977, pp. 764-767.

[42] M.J. Hunt, "A Statistical Approach to Metrics for Word and Syllable Recognition," J. Acoust. Soc. America. Vol 66, 1979.

[43] M.J. Hunt, "The Speech Signal," Proc. NATO AGARD Lecture Series No. 129, Speech Processing, May 1983. pp. 2.1-2.12.

[44] M.J. Hunt and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-88, 1988.

[45] M.J. Hunt and C. Lefebvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech, Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-89, 1989, pp. 262-265.

[46] M.J. Hunt and C. Lefebvre, "Distance Measures for Speech Recognition", Aeronautical Note, NAE-AN-57, Ottawa, March, 1989.

[47] M.J.Hunt, D.C. Bateman, S.M Richardson and A. Piau, "An Investigation of PLP

and IMELDA Acoustic Representations and of Their Potential for Combination," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-91, 1991, pp.881-884.

[48] D.C. Bateman, D.K. Bye and M.J.Hunt, "Spectral Contrast Normalization and Other Techniques for Speech Recognition in Noise," Proc. IEEE Trans. on Acoustics Speech and Signal Processing, ICASSP-92, 1992, pp. 241-244.

[49] N.S. Jayant & P. Noll, "Digital Coding of Waveforms: Principles and Applications to Speech and Video," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1984.

[50] B.H. Juang, "Speech Recognition in Adverse Environments," Computer Speech and Language (1991) 5, Academic Press Ltd., pp. 275-294.

[51] J.C. Junqua & Y. Anglade, "Acoustic and Perceptual Studies of Lombard Speech: Application to Isolated Words Automatic Speech Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-90,1990, pp.841-844.

[52] G.S. Kang & L.J. Fransen, "Experimentation with an Adaptive Noise-Cancellation Filter," IEEE Trans. on Circuits and Systems, Vol. CAS-34, No. 7, July 1987, pp. 753-758.

[53] Koehler et al, "Integrating RASTA-PLP into Speech Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-94, 1994, pp. I-421 - I-424.

[54] D.H. Klatt, "A Digital Filter Bank for Spectral Matching," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-79,1979, pp.573-576.

[55] C. Lefebvre et al, "Further Optimisation of a Robust IMELDA Speech Recogniser for Applications with Severely Degraded Speech," Proc. Int. Conf. on Spoken Lang. Processing, ICSLP-92, 1992, pp. 691-694.

[56] C. Lefebvre. Personal communication, October 1992.

[57] S. Lerner & B. Mazor, "Telephone Channel Normalization for Automatic Speech Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992, pp. I-261-I-264.

[58] J.S. Lim, "Spectral Root Homomorphic Deconvolution System," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 3, June 1979, pp. 223-233.

[59] F.H Liu, A. Acero & R.M. Stern, "Efficient Joint Compensation of Speech for the

Effects of Additive Noise and Linear Filtering," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992.

[60] P. Lockwood, J. Boudy , M. Blanchet, "Non-Linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environments," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-92, 1992, pp.265-268.

[61] P. Lockwood, J. Boudy, "Experiments With A Non-Linear Spectral Subtractor (NSS), Hidden Markov Models And The Projection, For Robust Speech Recognition In Cars," Speech Communication, vol. 11, Elsevir Science Pub., 1992, pp. 215-228.

[62] D. Mansour, B.H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-37, No. 11, Nov. 1989, pp. 1659-1671.

[63] D. Mansour, B.H. Juang, "The Short-Time Modified Coherence Representation and Noisy Speech Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-37, No. 6, June 1989, pp. 795-804.

[64] J.D. Markel et al, "Long-Term Feature Averaging for Speaker Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 4, Aug. 1977, pp. 330-337.

[65] J.D. Markel & S.B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 1, Feb. 1979, pp. 74-82.

[66] N. Morgan & H. Hermansky, "RASTA Extensions: Robustness to Additive and Convolutional Noise," Proc. ESCA Workshop on Speech Proc. in Adverse Conditions, Nov. 92, pp. 115-118.

[67] M. Morgan, D.R. Starks, "Fly-by-voice: A Technology Demonstration", Proc. 80th Symp. AGARD Flight Mech. Panel on Flight Testing, Greece, 1992.

[68] C. Mokbel & G. Chollet, "Word Recognition in the Car: Speech Enhancement / Spectral Transformations," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-91, 1991, pp. 925-928.

[69] R.K. Moore, "Systems for Isolated and Connected Word Recognition," Proc. NATO ASI series Vol. F16, published in "New Systems and Architectures for Automatic Speech Recognition and Synthesis, Edited by R. De Mori & C.Y. Yuen, Springer-Verlag Berlin 1985. pp 73-142.

[70] French National Update, NATO DRG/AC243 Panel III/RSG10 meeting, Oct. 1994, Royal Military Acedemy, Brussels.

[71] D. O'Shaugnessy, "Speech Communication Human and Machine," Addison-Wesley Publishing Co., 1987.

[72] T. Parsons, "Voice and Speech Processing," McGraw-Hill, Inc., 1986.

[73] J. Picone, "Continuous Speech Recognition using Hidden Markov Models," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-7, July 1990, pp. 26-41.

[74] G. Quenot et al, "A Dynamic Time-Warp VLSI Processor for Continuous Speech Recognition," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-86, 1986.

[75] Rabiner et al, "An Improved Training Procedure for Connected-Digit Recognition," Bell Syst. Tech. J. Vol. 61, No. 6, July-Aug. 1982, pp. 981-1001.

[76] P.K. Rajasekaran, G. R. Doddington, J.W. Picone, "Recognition of Speech Under Stress and in Noise," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-86, 1986, pp. 733-736.

[77] D.A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, Oct. 1994, pp. 639-643.

[78] D. B. Roe, "Speech Recognition With A Noise-Adapting Codebook," Proc. IEEE Trans. on Acoustics. Speech and Signal Processing, ICASSP-87, 1987, pp. 1139-1142.

[79] H. Silverman & D.P. Morgan, "The Application of Dynamic Programming to Connect Speech Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP 7, July 1990, pp. 6-25.

[80] Smolders et al, "On the Importance of the Microphone Position for Speech Recognition in the Car," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-94, 1994, pp. I-429 - I-432.

[81] D. Starks, "An Investigation into Systolic Arrays for the DTW Algorithm as applied to Speech Recognition," report submitted to Dr. W. Steenaart, University of Ottawa, April 1991.

[82] D.R. Starks, J.M. Morgan, "Integrating Speech Recognition Into a Helicopter",

Proc. ESCA Tutorial and Workshop on Speech Processing in Adverse Conditions, Cannes-Mendelieu, France, 1992.

[83] D. Starks, "Parallel Processing with DSP: A look at the TMS320C40 in a Large Vocabulary Speech Recognition Application," report submitted to Dr. S. Panchanathan, University of Ottawa, Nov. 1993.

[84] D. Starks, "Robust Speech Recognition: An Investigation into Reducing the Dependency on Apriori Knowledge of the Test Environment," DIR Project No. 22071EC05, Canadian Marconi Company Doc. No. MS3074, Oct. 1993.

[85] A.P. Varga & R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-90, 1990, pp. 845-848.

[86] B.A. Mellor & A.P. Varga, "Noise Masking in a Transform Domain," Proc. IEEE Trans. on Acoustics, Speech and Signal Processing, ICASSP-93, 1993, pp. 87-90.

[87] D. Van Compernolle, "DSP Techniques for Speech Enhancement," Proc. ESCA Workshop on Speech Proc. in Adverse Conditions, Nov. 92, pp. 21-30.