CELP and Speech Enhancement

Ian McLoughlin

PhD Thesis

School of Electronic and Electrical Engineering
Faculty of Engineering
University of Birmingham

October 1997

Synopsis

This thesis addresses the intelligibility enhancement of speech that is heard within an acoustically noisy environment. In particular, a realistic target situation of a police vehicle interior, with speech generated from a CELP (codebook-excited linear prediction) speech compression-based communication system, is adopted.

The research has centred on the role of the CELP speech compression algorithm, and its transmission parameters. In particular, novel methods of LSP-based (line spectral pair) speech analysis and speech modification are developed and described. CELP parameters have been utilised in the analysis and processing stages of a speech intelligibility enhancement system to minimise additional computational complexity over existing CELP coder requirements.

Details are given of the CELP analysis process and its effects on speech, the development of speech analysis and alteration algorithms coexisting with a CELP system, their effects and performance.

Both objective and subjective tests have been used to characterize the effectiveness of the analysis and processing methods. Subjective testing of a complete simulation enhancement system indicates its effectiveness under the tested conditions, and is extrapolated to predict real-life performance.

The developed system presents a novel integrated solution to the intelligibility enhancement of speech, and can provide a doubling, on average, of intelligibility under the tested conditions of very low intelligibility.

Acknowledgements

I would like to express thanks to the following people who have assisted me greatly in my work: Firstly, my supervisor, R. J. Chance who for the past three years has rendered invaluable help and assistance, has continually expressed his interest in my work, and has provided much of the encouragement and guidance that has resulted in this thesis.

Many other past and present members of the University of Birmingham School of Electronic and Electrical Engineering have also assisted me in my research, and although too numerous to name, I would like to especially thank the following: Dr J.A.Edwards who was instrumental in establishing the initial industrial link with Simoco (formerly Philips PMR) and has remained interested in this work throughout. Dr S.I.Woolley has provided ideas and encouragement, and in particular has lent her time to comment on my draft thesis. Dr. H. Ghafouri-Shiraz, Dr. I. Gu, Dr. D.G. Checketts, Dr. S.M. Bozic and D. Pycock also deserve mention for their help.

I wish to acknowledge the encouragement and support of Dr. M. Rayne of Simoco International (Cambridge), who accepted the industrial link from the former Philips Telecom PMR, who has assisted in the funding of the work, and who has expressed much interest in the research.

Further acknowledgements must go to my ever patient and helpful wife, my parents, fellow researchers and friends, and also to K. Worrell and D. Pinfold of Her Majesty's Government Communications Centre (HMGCC, Hanslope, Buckinghamshire) who have remained in contact with me, and have kept me up to date with the latest industrial developments.

I would like to thank Chief Inspector Anderson and Sgt. Bassett of West Midlands Police who arranged for my police vehicle noise measurement tests, and for Sgt. Higgins and Sgt. Simmonite who patiently drove me around North Birmingham for half a day, attached to recording equipment.

Finally, my thanks extend to those people who participated in the listening tests, for their patience and willingness to help.

Contents

List of figures and tables

Figures

Tables

Glossary of terms

ACELP	Algebraic Code Excited Linear Prediction
AMDF	Average magnitude Difference Function
ANSI	American National Standards Institute
audibility	the degree to which a sound can be distinguished by a listener
Bark	a frequency scale with units of equal perceptual relevance
CCITT	International Telegraph and Telephone Consultative Committee (now known as the ITU)
CELP	Code Excited Linear Prediction
cepstrum	the signal obtained by inverse Fourier transforming the log-magnitude of a Fourier transformed signal (where the logarithmic action is to convert a convolution of signals to an addition)
DAM	Diagnostic Acceptability Measure
DCC	Digital Compact Cassette (© Philips)
DMCT	Diagnostic Medial Consonant Test
DMOS	Degradation Mean Opinion Score
DRT	Diagnostic Rhyme Test
DSP	Digital Signal Processing
ETSI	European Telecommunications Standards Institute
FIR	Finite Impulse Response
formant	the spectral peaks in speech due to vocal tract resonances
fricative	a consonant formed through a constriction in the vocal tract
FS1015	US Department of Defense standard LPC speech coder
FS1016	US Department of Defense standard CELP speech coder
F*x*	Formant number *x*
G.728	CCITT standard CELP coder
glissando	musical term for a continuous slide of musical notes upwards or downwards in frequency
GSM	Groupe Speciale Mobile (renamed in 1994 to Global Standard for Mobile communications)
IIR	Infinite Impulse Response
IPA	International Phonetic Alphabet
ITU	International Telecommunication Union
LPC	Linear Predictive Coding
LPC	Linear Prediction Coefficients
LSP	Line Spectral Pair(s)
LTP	Long Term Predictor

masking	the process whereby a sound may be inaudible in the presence of a second sound of similar frequency and amplitude
MiniDisc	miniature magnetic/optical audio recording format (© Sony)
MFLOPS	Million FLoating point Operations Per Second
MIPS	Million Instructions Per Second
MOPS	Million Operations Per Second
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
MSE	Mean-Squared Error
PARCOR	Partial Correlation
PB	Phonetically balanced
PEWF	Perceptual Error Weighting Filter
phoneme	a single unit of speech sound
POW	abbreviation for frame power measure
psychoacoustic	difference between a measured and a subjective audio experience
RPE	Residual Pulse Excitation
SEGSNR	Segmental Signal-to-Noise Ratio
slur	musical term to frequency slide between different musical notes
SNR	Signal-to-Noise Ratio
SpAT	Spelling Alphabet Test
SPL	Sound Pressure Level
TETRA	ETSI standard: Trans-European Trunked radio system
voiced	to add a vocal chord contribution to a sound
unvoiced	a sound produced without a pitch element
TIMIT	a standard database of American English speech recordings
VSELP	Vector Sum Excited Linear Predictor
ZCR	Zero-Crossing Rate

1 Introduction

This thesis describes a system for automatically adjusting speech, transmitted through a CELP codec, into an environment containing high levels of interfering acoustic background noise. The aim of the enhancement is to improve the intelligibility of the transmitted speech for a listener located in that environment.

The CELP speech codec, is a method for compressing the information present in a speech signal for storage or transmission, and later reconstruction. CELP relies upon an internal structure similar to the human vocal apparatus to analyse speech, and thus compresses the speech signal into parameters, each of which encodes a separate aspect of the speech content, resulting from one of several vocal production stages.

Speech reconstructed from CELP analysis will, like unprocessed speech, be unintelligible when listened to in the presence of high levels of acoustic background noise. This situation commonly arises due to the widespread use of CELP coding for mobile communication systems such as public service vehicle radios and mobile telephones.

When both the listener and speaker are located in high levels of acoustic background noise, the speaker will naturally adjust his or her voice to enable the listener to understand. In a situation where the listener is located in a noisy environment but the speaker is in quiet, the speaker's voice is unsuited to the listener's environment. Speech enhancement can be considered to be a method of automatically adjusting the speakers speech to suit the conditions of the listener.

It is useful to define bounds on the speech enhancement system, and this is done with the premise that the speech enhancements developed in this thesis should integrate closely with the CELP coder, and should be tested to enhance speech intelligibility in a particular environment. This *target environment* is a vehicular communication system where speech is transmitted from a quiet base station to the acoustically noisy interior of a police vehicle. When considering the target environment in this thesis, the person located in the vehicle is termed the *listener*, and the person in the quiet environment, the *speaker*, after their usual roles for speech enhancement purposes. In fact, the listener may also speak. This is important in that his speech alters the characteristics of the acoustic noise in the target vehicle.

One of the latest CELP variants designed for public service use, the TETRA codec, is considered to be the speech coder in use in the target environment. Methods described in this thesis are matched against the target environment, as are the tests against which the performance of those methods are judged.

Part I of this thesis describes relevant background information relating to methods of enhancing speech, situations requiring enhancement, background acoustic noise, the speech production and understanding mechanisms and speech compression using CELP. Chapter 4 then ends part I by discussing further methods of enhancing speech, both existing and proposed.

Part II discusses the basic processes required to implement the chosen enhancement methods of part I, in particular introducing the novel LSP processing method. Chapter 6 proposes methods of analysis that are required to enable automated speech enhancement to occur, and chapter 7 gathers this information together to create and describe a speech enhancing CELP system.

Part III of this thesis tests and analyses the processes developed in part II, and describes experimental procedures used to characterise the performance of a fully integrated enhancing CELP structure. The final outcome of the testing procedure, given in chapter 10, is a prediction of speech enhancement performance, and a description of the final speech enhancing CELP structure.

Chapter 11 concludes by relating the speech enhancing CELP coder to the target environment, outlining the novel aspects of the system, and its potential for application elsewhere.

PART I: Investigation and Context

This section begins by laying the foundations upon which the speech enhancement algorithms will later be based. In the search for a method of automatically adjusting speech output from a CELP coder in order to improve its intelligibility to a listener in an acoustically noisy environment, these chapters introduce CELP; the noisy environment, and methods of enhancing speech.

In chapter 2, existing methods of speech enhancement, and their relationship to the human vocal or speech comprehension systems are described, and further suggestions made. The acoustic background noise of the target environment is then considered.

Chapter 3 introduces the CELP algorithm, with particular reference to those parts of the CELP algorithm upon which speech enhancements will rely, or can be integrated with.

Chapter 4 then summarises the methods of speech enhancement discussed in chapter 2, relating these to the target environment and the CELP algorithm, in order to choose a subset for further investigation. The analysis and control requirements of each of the chosen speech enhancement methods are discussed, and will be considered further in part II of this thesis.

2 Candidate speech enhancement methods

2.1 Definitions and background

It is useful before embarking upon a description of speech enhancement methods to define speech intelligibility. Intelligibility, the ability of the speech to impart information, must be clearly distinguished from quality, a measure of how pleasant a sound appears to a panel of listeners [104]. The ANSI definition of intelligibility is followed in this thesis:

> "That property which allows units of speech to be identified. Intelligibility over a speech communication system is that property which allows trained listeners to receive and to identify speech spoken by trained talkers or by a speech coder when the talkers or that coder and the listeners are connected by a speech communication system."[3].

Considering the above definition of intelligibility, speech enhancement may be defined as any process which causes intelligibility to increase in a given situation. This chapter is mainly concerned with methods of speech intelligibility enhancement, discussing existing methods and proposing alternatives.

2.2 Existing methods of enhancement

In recent years, there has been much research conducted in the field of noise reduction, however this tends to apply to the different requirement of reduction or attempted removal of noise from noise-contaminated speech. Where an acoustically noisy listening environment is encountered, research has centred on adaptive noise cancellation techniques [22] rather than the converse: speech intelligibility strengthening techniques.

Very little research has been conducted on the adjustment of speech in order to improve intelligibility. It is still possible however to consider research in related fields such as adaptive noise reduction and noise removal, and in the field of CELP coding itself, where many individual techniques will apply to the speech-alteration case.

The outcome of much of the research relevant to speech enhancement is discussed in the following subsections.

In normal speech, vowels are spoken with an average amplitude of approximately 12dB louder than consonants [113], and in fact, without exception, vowels are spoken with more power than consonants. For average speech, the range of intensity for phonemes is around 28dB, as shown in table A1.1 in appendix A1.1.

The fact that vowels are louder than consonants is rather surprising considering that consonants convey more intelligibility than vowels. The relative information carrying content of vowels compared to consonants may be demonstrated by speaking a sentence such as “The yellow dog had fleas”, firstly with all consonants spoken the same (something like “Tte tettot tot tat tteat”) and secondly with all vowels spoken the same (“Tha yallaw dag had flaas”)! Both sentences sound a little unusual, but the second is more understandable than the first - hence consonants can be demonstrated as carrying a greater proportion of information (this idea was taken from a similar sentence presented by Her Majesty’s Government Communications Centre audio group in an internal demonstration tape).

Of course this does not mean that only consonants should be considered as candidates for enhancement: words like ‘bait’, ‘boat’, ‘bite’ and ‘beat’ can only be differentiated by their vowels. In this example, vowels convey all of the information regarding the difference between words, so although consonants usually convey more information, this is not always so.

However we can say that the more information-rich class of phonemes is 12dB lower in amplitude than the less information-rich class. The clipping process normalizes all classes of speech, effectively improving intelligibility in a similar manner to amplifying consonants by 12dB. Unfortunately the side effect of clipping is substantial distortion, which tends to reduce the intelligibility gain slightly.

Later research extended the clipping idea by high-pass filtering the speech signal prior to the clipping process [74][116][117]. For a 1.1kHz cutoff HPF with 12dB/octave response, 0dB signal-to-noise ratio and 90dB$_{SPL}$ total sound amplitude, around 7dB improvement is noted over non-high-passed results [116] (dB$_{SPL}$ being the ratio of the sound pressure to that of the average threshold of hearing for 1kHz tones of 0.00002Pa=0dB$_{SPL}$ [53]).

The effect of the high-pass filtering (combined with the subsequent renormalization) is to reduce the amplitude of the first formant (formants are discussed in section A1.2) with respect to the second and higher formants. As the second and higher formants together convey more intelligibility than the first formant, there is thus a greater intelligibility transmitted on average per unit energy if the power of the first formant is attenuated.

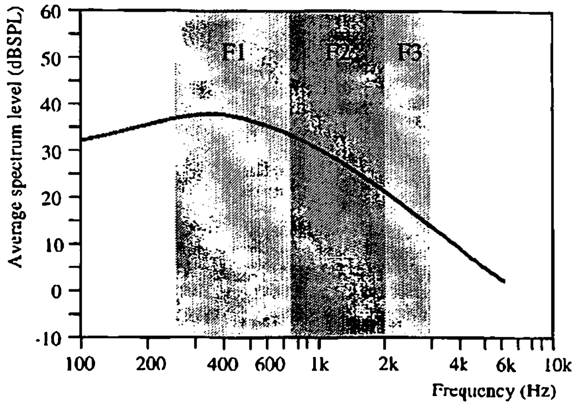


Figure 2.2: Long-time averaged speech power distribution constructed from examination of figures in (83) and showing ranges of average formant location from data in (127).

Fig2.2 illustrates the average energy distribution of speech, with the frequency ranges of the first three formants (F1, F2 and F3) overlaid. In fact 84% of speech energy is located below 1kHz [127], and this peak energy corresponds closely with the first formant frequency range.

Next consider the graph of fig2.3 which demonstrates the relative information carrying content of the speech signal. An analysis of fig2.3 reveals that if a speech signal were low-pass filtered at 1kHz, around 25% of speech syllables would be recognisable. If it were high pass filtered at 1kHz, around 90% would be recognisable.

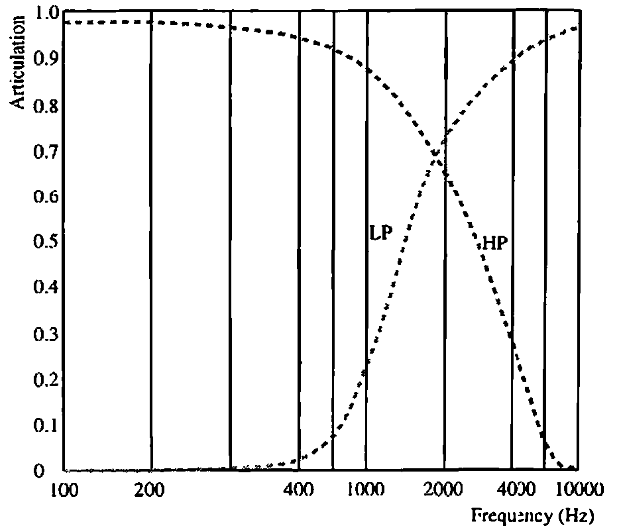


Figure 2.3: Effect of frequency range on the syllable articulation of speech. Graph reproduced from notes presented in a seminar given by HMGCC audio group.

Fig2.2 and fig2.3 together illustrate that much of the speech signal energy is conveyed by F1 (the first formant) but that if these frequencies were removed, approximately 90% of speech would still be intelligible. Thus the high-pass filtering process and subsequent renormalization effectively provides more intelligibility per unit energy, and thus higher overall intelligibility.

So combining the two methods in filter-clipping, improves speech both by redressing the amplitude-intelligence imbalance of vowels and consonants, and by redressing the amplitude-intelligence imbalance of F1 and higher formants. Many modifications have been made to these basic techniques by different authors, but are mostly limited to improving the clipping process [60][74] or the filtering step [117].

2.2.2 Enhancing CELP

Since the CELP speech compression system was introduced in 1985 [98], there have been many additions and alterations for enhancement. CELP will be described further in chapter 3, but it can be noted here that CELP usually relies upon a perceptual error weighting filter (PEWF: see section A2.2.3). The PEWF works by concentrating energy around the formant regions, effectively emphasising this important part of the speech signal [119][35].

The PEWF, although usually used within the CELP structure has also been applied successfully to the speech output from the CELP decoder in order to improve quality [16].

Although perceptual weighting of the CELP output has been shown to improve quality, there is no corresponding increase in intelligibility [16]. More advantageous in terms of intelligibility would be a perceptual weighting filter that attenuated F1 and amplified higher formants, as described in section 2.2.1. However the importance of the PEWF is its ability to modify the speech spectrum in relation to the formants.

2.2.3 Noise removal techniques

The most commonly applied method for removal of noise from noise-corrupted speech is spectral subtraction [11]. In this method, an estimate is generated of the frequency spectrum of the interfering noise, which is subtracted from the noisy speech spectrum to give an estimate of the clean speech spectrum. The noise frequency estimate may be generated from an analysis of the noise signal during non-speech activity - assuming of course that the noise spectrum remains fairly stationary throughout the speech signal.

Although the spectral subtraction process is not directly relevant to speech modification, both the speech activity detection and the noise spectrum estimation are useful. For a CELP-based speech-modification system working in a vehicle, it is essential to measure the ambient noise within the vehicle (to determine audibility and how to alter the speech), and this must be performed when the vehicle occupants are not talking. Thus attention must be paid to the development of speech detection and spectral estimation routines such as those used for spectral subtraction.

Further consideration will be given to speech detection and noise spectrum estimation in chapter 6.

2.2.4 Techniques developed for hearing-impaired listeners

Although any general speech intelligibility enhancement mechanism would be designed to improve intelligibility for average-hearing listeners, there has been little direct research to date in this field. However much research has been conducted, by the medical community, on modification of speech for the benefit of hearing-impaired listeners, and on the enhancement of the intelligibility of that speech. Whilst the type of speech adjustment may be inappropriate for normal-hearing listeners, the methods and assumptions used are still valid.

There are two classes of enhancements for hearing-impaired listeners: those that are designed to reduce the perception of noise in noisy speech signals, and those that adjust speech itself to improve intelligibility. The former are usually targetted at processing systems located in hearing-aids, and may include spectral subtraction [11] or other methods of noise reduction [45]. The latter may reside in hearing aids, telephony systems or in cochlear implants, and adjust speech through modification of formants [10][2][950][109] or by altering amplitude scales [26].

Schaub and Straub [95] reported the use of a perceptual weighting filter (as mentioned in section 2.2.2) using linear prediction parameters to filter speech in order to 'sharpen' the formants by increasing energy around them, and reducing energy in the spectral valleys between formants - *spectral sharpening.* Fig2.4 shows the results of a simulation of their perceptual weighting filter on a period of voiced speech.

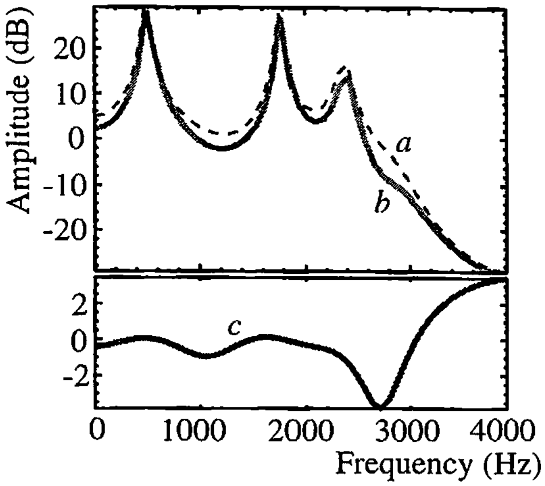


Figure 2.4: Spectrum of a) unfiltered and b) filtered test speech frame as operated on by c) perceptual weighting adaptive postfilter function.

In fig2.4, lines a and b, the unfiltered and filtered speech spectra, have both been normalized with respect to amplitude, whilst c, the alteration induced by perceptual weighting, has been to concentrate more power in the formant regions, and incidentally to amplify the high-frequency region of the spectrum. It must be noted that the simulation leading to fig2.4 did not replicate the complicated adaptive normalization process used in [95]: the effect of which was to ensure that all filtered speech was of higher peak amplitude than unfiltered speech, a factor which must be considered alongside their claims of speech enhancement. It is likely that some degree of enhancement would be noted by just using their amplitude normalization scheme and omitting their enhancement scheme.

The perceptual weighting technique outlined above was designed for commercial use in hearing aids, and particularly for people with reduced frequency selectivity. A similar scheme has been attempted by Alcántara et. al. [2] who additionally tested listeners with normal hearing. The scheme adopted was similar in intent to [95], however the formant adjustment was made by means of bandpass filters located at the formant centre frequencies.

Previous experimental results reported in [2] indicate that spectral sharpening of speech for normal hearing listeners generally reduces intelligibility unless the level of noise added to the processed signal is increased, to give signal-to-noise ratios of 6dB and below. Other results indicated that both spectral broadening (the inverse of sharpening) and spectral sharpening reduce speech intelligibility when that speech is listened to in a noise-free environment.

Alcántara et. al. [2] reported using two bandpass filters dynamically located at the frequencies of F1 and F2 (the first and second formants), and adjusted to attenuate the signals outside of the two formant regions. The filter Q settings were identical, resulting in a larger bandwidth for F2, than F1. This is unfortunate since in general F1 has larger bandwidth than F2 [115], and is probably unintentional considering the following note in their paper:

> Naturally occurring formants are typically $\frac{1}{3}$ to $\frac{1}{2}$ octave bandwidth for f_1 and $\frac{1}{10}$ to $\frac{1}{5}$ octave for f_2.[2]

The results presented in [2] for consonant-vowel-consonant and vowel-consonant-vowel tests (described later in section 9.2.3) are mixed, but indicate that formant adjustment can improve the recognition of vowel sounds when listened to in noise (and also in quiet conditions for hearing-impaired listeners). The case for consonant intelligibility improvement is less clear: a statistically-irrelevant reduction in intelligibility was noted.

2.3 Enhancement methods

Section 2.2 discussed known techniques for speech intelligibility enhancement. These are extended further in this section where speech, hearing, psychoacoustic features and the CELP coder are investigated in order to discover opportunities for enhancement. Some methods of achieving these enhancements are also discussed here.

2.3.1 Amplification

The most obvious method of negating the effects of background noise is that of amplifying the speech so that the speech-to-noise level is increased. Whilst this approach is valid (after all this is similar to the natural response of humans talking in the presence of background noise to raise their voice), it must be noted that it is more important to compare the level of the information carrying part of the signal over the corresponding noise frequencies, rather than the average level. For non-white noise, measurement of power could disguise the fact that it may consist of large components at certain frequencies, combined with a low noise floor. Thus, analysis of noise amplitude in the frequency band between about 800Hz and 3kHz, and suitable amplification to give a reasonable value of weighted signal-to-noise, would be advantageous [99].

Preferable would be a determination of actual formant frequencies and then an analysis of the speech-to-noise ratio in small frequency bands centred on those formant frequencies. Suitable filtering would then ensure that speech-to-noise levels at those important frequencies were maintained.

Amplification of speech must always operate with consideration to the variation in intelligibility with absolute intensity. This is because as the amplitude of a speech signal is increased above a certain limit (whilst maintaining constant speech to noise ratio), intelligibility does not continue to increase [53].

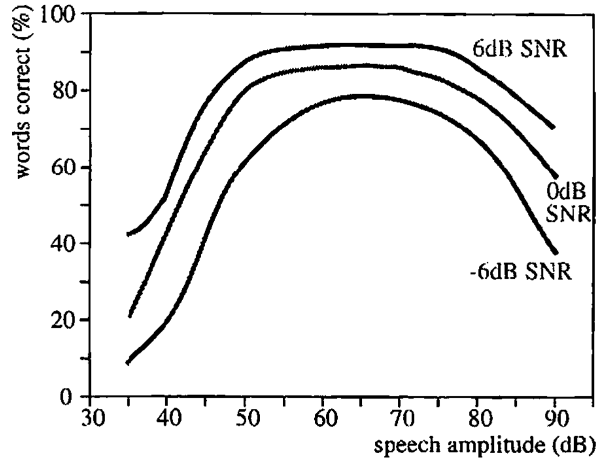


Figure 2.5: Variation of intelligibility (measured as percentage of spoken words correctly identified) with amplitude. Data obtained from examination of figures in (53) and (55).

The graph of fig2.5 indicates that for the three example levels of speech SNR, as the speech amplitude is increased above about 75dB, intelligibility begins to decrease: In fact, when SNR is 6dB, and speech amplitude is 80dB, a further increase of 6dB in speech amplitude to combat a 6dB rise in noise level results in lower intelligibility (78%, from 6dB SNR curve at 86dB speech amplitude) than if the speech amplitude was not increased (80%, from 0dB SNR curve at 80dB speech amplitude).

Any process must strive to ensure that the absolute amplitude of processed speech lies within the area of maximum intelligibility shown in fig2.5. As soon as amplification is required to give a certain speech-to-noise level that places the speech within the region of reducing intelligibility, it may be preferable to lower the speech-to-noise level target, and thus the amount of amplification, to improve intelligibility.

2.3.2 Spectral modification

Background noise having a steeply distributed shape may be negated in part by inducing a corresponding amplification-filtering operation on the speech such as those described in section 2.2.4. Limitations are that too much shaping of the speech spectrum will reduce quality and intelligibility, and perhaps even cause the speech to alter so much that the brain cannot recognise the sounds as speech.

A more general approach is to filter the speech so that the power of the information carrying part of the speech is increased in relation to other less essential frequencies. This may be accomplished by using methods to strengthen the formant frequencies (section 2.2.2), or may involve simple non-adaptive filtering to increase the amplitude of the all-important 800Hz to 3kHz frequency band within which most speech intelligibility resides (section 2.2.1).

We have noted that much of the intelligibility in speech is conveyed by formants, and thus an algorithm to detect formants [64][76][134], and to filter the signal so as to strengthen them, is likely to improve intelligibility in the presence of noise. Such algorithms are in general use within many CELP coders as the perceptual error weighting filter as covered in sections 2.2.2 and 2.2.4, and described in section A2.2.3.

Spectral modification of formants described here, includes both formant sharpening and the converse, broadening or widening. However one further class exists: that of adjusting the frequency of formants, developed for voice-changing applications [105].

Some further considerations follow:

2.3.2.1 Masking

It is important here to introduce the concept of masking. Masking in general is defined by the American standards agency as:

> The process by which the threshold of audibility for one sound is raised by the presence of another sound.
>
> The amount by which the threshold of audibility of sound is raised by the presence of another sound.

The frequency selective function of the basilar membrane [49][127] within the inner ear may be considered similar to a bank of bandpass filters with the threshold of audibility in each filter being dependent upon the noise energy falling within its passband [10]. The filters each have similarly shaped responses with bandwidths around 100Hz up to frequencies of approximately 1kHz. Above this frequency, bandwidth increases in a linear fashion with frequency up to a 3kHz bandwidth at 10kHz. Each logical filter is termed a critical band [99].

For a given tone having fixed amplitude and frequency, the sensitivity of the ear to other coincident tones of similar frequency is reduced. It has been possible for many authors to derive logical models of this masking process, of differing degrees of complexity and ability [18][19][33][63][99][101][121][124][135].

For sounds whose bandwidth falls entirely within one critical band, the intensity of that sound is independent of its bandwidth. However, for sounds with bandwidth greater than one critical band, the intensity depends very strongly on the proportion of the sounds bandwidth falling within one critical band. Thus, in general, a complex sound having components in more critical bands sounds louder than a complex sound with components in fewer critical bands [70]. Such a sound will thus be more intelligible in a given degree of acoustic background noise.

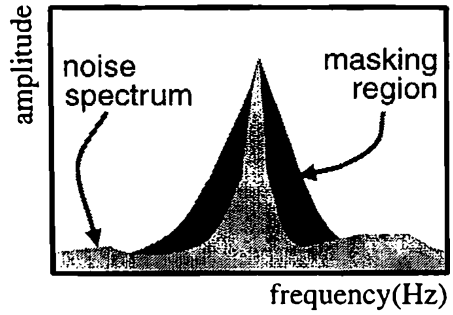


Figure 2.6: Illustration of the masking effect generated by a tonal noise.

Fig2.6 illustrates that a tonal noise of given amplitude generates a masking effect in the immediate frequency region of the tone, and extending beyond it with the effect reducing as the distance from the generating tone increases. If another tone were introduced whilst an average listener was hearing the tone shown, then the peak of the second tone would have to be outside of the masking region to be heard.

The effect of tonal masking, as described above, is to cause a sound to be hidden by another sound of similar frequency and greater amplitude [7]. As formants are extremely important to vocal communications, it is reasonable to expect that slightly altering the position of a formant that is to be masked by noise will improve its audibility, and thus its intelligibility.

Apart from the obvious methods of fixed or adaptive filtering, targetted spectral modification can be accomplished by applying well-established methods such as LPC-pole adjustment [105], or by developing alternative methods (see chapter 5).

2.3.3 Clipping and selective amplification

As discussed in section 2.2.1, consonants convey more information than vowels, and yet are 12dB lower in amplitude on average. If it were possible to amplify consonants to equalise their amplitude with respect to vowels, an improvement in intelligibility should result.

The clipping process developed during World War II, and described in section 2.2.1 is such a system, however the effects of this intelligibility enhancement include severe quality degradation and a reduction in the ability of listeners to identify speakers [59][60][74][116].

Amplifying consonants whilst not altering vowel amplitude would be expected to increase intelligibility [81] in a similar manner to clipping, but would not suffer the same degree of quality degradation. Such a proposed system would require a speech classifier to identify consonant and vowel periods, and an amplitude tracker to provide an adaptive level at which to normalize consonant amplitudes. Periods of non-speech need to be identified to prevent amplification of noise, and a smoothly varying attack and decay to amplification periods is required to reduce sudden transients.

Such a process also benefits by being able to apply differing levels of amplification to speech segments such that the overall speech amplitude envelope is preserved (for example, allowing whispering or other long-term variations in speech loudness).

Segmentation of speech into different classes for selective amplification also introduces the possibility of applying alternative processes to different parts of speech. An example of this is to selectively amplify fricative and sibilant sounds so that they stand out from a background of white noise, which has a similar frequency distribution.

2.3.4 Temporal methods

Speech processors often split speech recordings into frames of equal length. Thus, an evaluation of the speech power and frequency distribution over that period will be useful references, when compared to the background noise evaluated over the same period. Upon this basis, the non-temporal processing methods considered previously may be applied. This allows the close matching of processing to the time-varying background noise.

Certain speech processing methods exist that utilise the gaps in conversational speech to compress transmitted information. Use of these techniques for segmenting speech and performing simple cut-and-paste operations may allow a process whereby a short word or syllable, about to be obscured by a large incidence of noise, could be shifted slightly in time. The shift could place the word or syllable a few hundred milliseconds away, but still within the range found in normal speech. The likely effect on speech intelligibility, if the warping delay were limited, would be to increase intelligibility at the expense of, perhaps, a slightly unnatural rhythm to the speakers' voice.

Under normal conversational circumstances, a talker would use non-verbal clues to determine if a listener had understood or heard his/her speech, repeating or re-phrasing him or herself as necessary to ensure that the message was conveyed. For remote communication channels, non-verbal clues are not passed, and listeners must explicitly ask for information to be repeated. If all else fails, and speech enhancement is not sufficient for a listener to understand the transmitted speech, repetition will be the only solution. This requires a bidirectional communication channel.

It is possible to predict (although with limited accuracy) if a listener has not heard a message by consideration of the absolute limits to an average listeners ability, the specifics of the transmitted signal and the background noise or distortion. However the use of this information to either delay words until conditions are more favourable, or to repeat words would require a considerable degree of processing.

2.4 Acoustic background noise

In order to consider speech enhancement methods, and develop working algorithms, an appreciation of the acoustic background noise within the target environment is essential. The target environment of a police vehicle is investigated to determine the type and degree of noise likely to interfere with speech.

2.4.1 Generation of noise

Vehicle noise arises predominantly from the tyres' interaction with the road surface, the engine and fan, the exhaust system, the air intake, and aerodynamic or wind noise [40][86].

The noise produced by tyres is heavily dependent upon tyre type, vehicle weight and the road surface, however the graph in fig2.7 illustrates a typical tyre noise spectrum.

Note that the noise amplitude peaks between about 300Hz and 2kHz - having unfortunate consequences for speech communication, which relies on the same frequency band. The noise peaks appear to rise about 2.5dB for every 10mph increase in vehicle speed, indicating that tyre noise may become very loud at high speed.

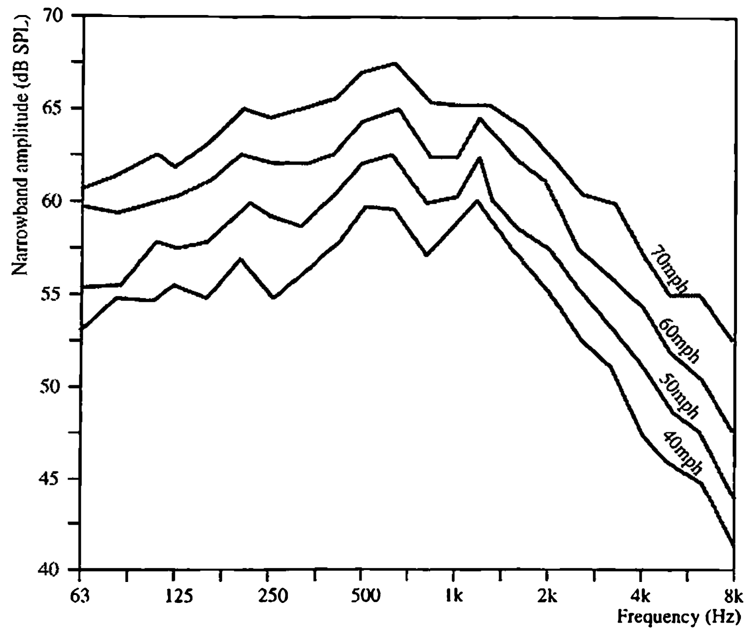


Figure 2.7: Spectral analysis of continuous-rib tyre noise, constructed from examination of a figure in [129].

An analysis of the difference between exterior vehicle noise when under maximum acceleration and that when travelling at constant speed [122] indicates that noise at frequencies between 100Hz and 1.5kHz is increased by about 20dB, with the increase being roughly constant at speeds of between 30 and 60mph. During acceleration, the throttle opens, causing air intake noise to increase, the rotational speed of the engine, and the firing rate increase, but tyre noise and aerodynamic noise remain relatively constant. Thus, during acceleration, a 20dB increase in sound intensity is mostly due to engine and intake noise.

At low speeds of up to 30mph, the exhaust system is likely to contribute most to the overall noise figure, along with engine fan noise, if present.

2.4.2 Noise dynamics

Most of the research in this area has been conducted on test tracks, with values averaged over relatively long periods of time, and collected under artificial test conditions. In reality, many factors will intrude to cause the generalisations offered by the research to become invalid.

Personal observations suggest that engine noises change when the vehicle is under acceleration or the engine is under strain (perhaps travelling up a steep incline). Road surface materials, as well as type material have a dramatic effect upon noise [21][126].

Under normal conditions, a vehicle may drive over manhole covers, rumble strips or white lines, all of which cause a temporary noise level increase. Wet roads are noisier, due to the impact of surface water sprayed on the inside of wheel arches and of rain hitting the surface of the vehicle the degree of loudness increase is highly dependent upon the type of vehicle. Occupants of vehicles passing through tunnels or along cuttings will experience an increase in noise due to reflected sound from their own and other engines. Reflection of sound also occurs whilst passing other vehicles, but this is generally less significant than the direct sound originating from the other vehicle.

Our expectation is that total noise in vehicles depends to some extent upon speed, however the relationship is surprisingly linear, as shown in fig2.8.

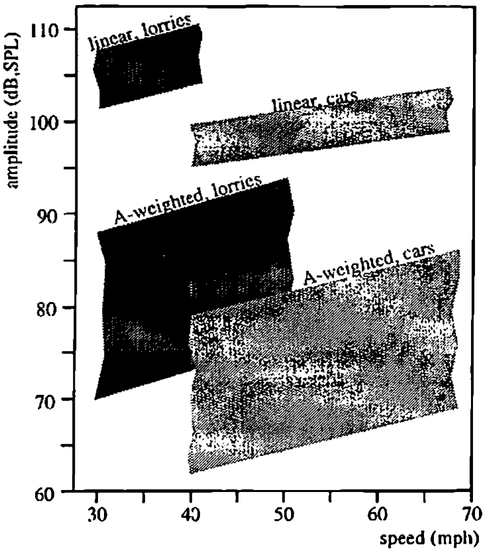


Figure 2.8: Variation of two measures of interior noise with speed. Constructed from examination of a similar figure presented in (114).

The shaded areas in fig2.8 represent the range of values found at the given speeds for the 47 lorries and 68 cars tested. The A-weighted measure is described in section A1.3. The apparently linear relationship between the interior noise envelopes and speed, suggest that interpolation of these results for higher speeds may be possible. For example, we may expect to find a maximum of around 108dB$_{SPL}$ in a car travelling at 90mph, with a minimum result about 5dB lower.

2.4.3 Frequency distribution

Total A-weighted noise within a typical vehicle, is around 65 to 85dBA at 60mph [110], however this figure, a weighted average, does not give a true picture of the amount of noise present in the vehicle, as fig2.9 illustrates.

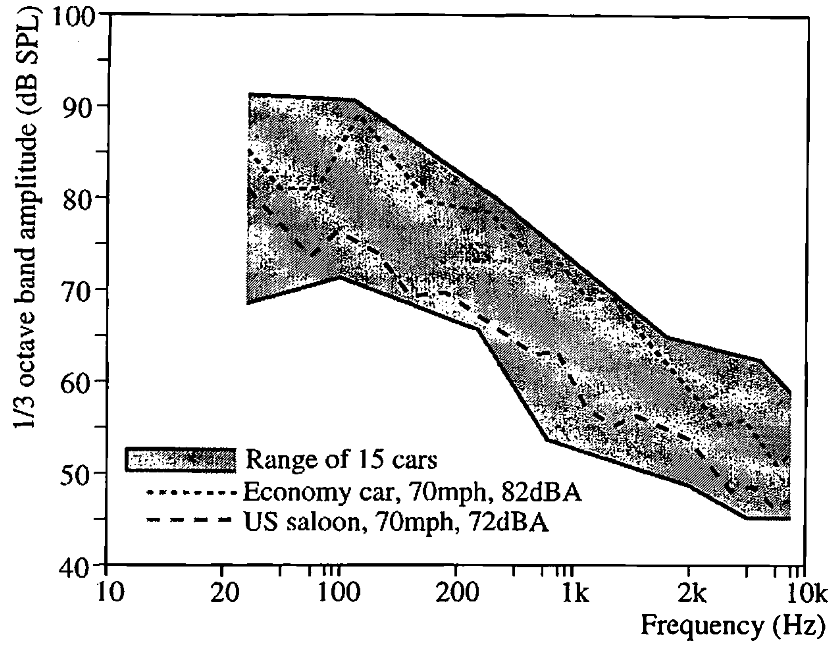


Figure 2.9: Noise spectrum of the interior of various cars. Constructed from data in (129).

In addition to showing the wide variation between different types of car (the shaded area), fig2.9 illustrates that even in a car with a noise rating at 70mph having a rather low figure of 72dBA, certain frequencies may contain noise power in excess of 80dB.

It is worth noting that the British Standards Institute have developed a way of assessing noise within vehicles [13]: BS6086 requires measurement of A-weighted noise, octave, and one-third octave band noise levels, in order to evaluate speech interference levels and determine the risk of hearing damage.

The predominant displacement of noise towards low frequencies follows the tyre-noise spectrum to some extent, and may be accentuated by resonances set up in the vehicle passenger space [20]. For a 2m wide and 3m long passenger cabin, resonances would occur at frequencies of 160Hz and 83Hz. Further research has been conducted on infrasonic noise within cars as illustrated in fig2.10.

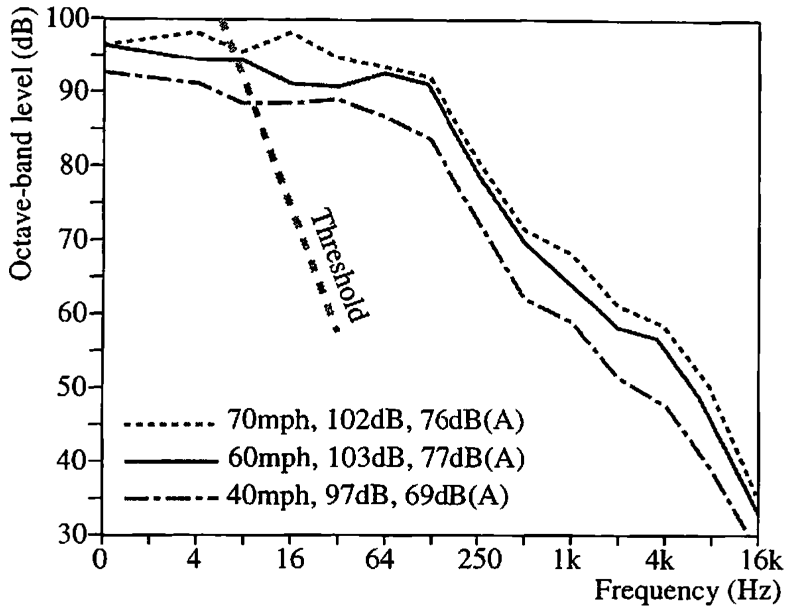


Figure 2.10: Low frequency spectrum of interior noise, constructed from data in (114).

In addition to the high levels of low frequency noise shown in fig2.10 at even quite moderate speeds, research has shown that levels of 110 to 120dB, due to random air turbulence, were present in the 2 to 32Hz frequency bands inside cars travelling with one window opened by about six inches [110]. Some studies suggest that this infrasonic noise, although supposedly imperceptible, does play a significant role in the subjective assessment of vehicle noise levels [130].

The data presented in fig2.10 has been used to construct a filter which, when operating on Gaussian white noise, produces a noise distribution similar to that found inside a vehicle. This forms the basis of a vehicle interior noise simulation, discussed in section 8.5.

Opening the window of a moving vehicle immediately creates a sound path between the exterior and interior of the vehicle. Such a path effectively presents a low impedance to sound compared with existing channels which are likely to have been treated with structural noise mitigation methods by the manufacturer. The effect of opening a side window by one inch in a lorry travelling at 50mph [130] is to increase interior noise, predominantly in low frequency regions as shown in fig2.11. Although effects will differ between lorries and cars, the shape of the frequency distribution and the degree of increase in noise are expected to be similar.

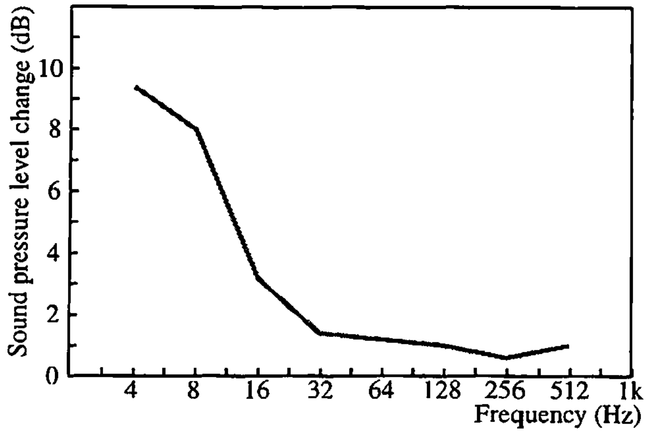


Figure 2.11: Changes in the spectral power of lorry interior noise at 50mph caused by opening a side window (130).

2.4.4 Target environment

In order to appreciate the levels, types, and subjective effects of noise, measurements were made in a typical West Midlands Police Force patrol vehicle.

Such testing is not intended to be statistically representative of such situations, but rather to provide an example of a situation in which noise levels regularly interfere with vocal radio communication, and in which the speech enhancement systems presented in this thesis should lead to improvements.

The 3.5 litre V8 Rover 800 automatic transmission vehicle, in which most tests were conducted was extremely well soundproofed. Personal observation suggests that engine noise was not intrusive even at 60mph in second gear or at 100mph in fourth gear. The 200W siren was located beneath the bonnet facing forward, and was not loud inside the vehicle, even when stationary, unless a window was opened.

The radio system speakers in the vehicle were standard manufacturers' 3 to 3½ inch cone speakers located in the driver and passenger door, and under the rear parcel shelf. Microphones were custom-fitted, handheld for the passenger and hands-free, located on a stalk ending just below the top rim of the steering wheel, for the driver.

West Midlands police currently use three siren sounds;

① *wailer*: a long drawn out up and down chirp with period around 4s and frequency bound of approximately 480 to 980Hz.

② *yelper*: similar to above but with reduced period of around 0.5s and identical frequency range.

③ *two-tone*: alternating tones of around 500 and 620Hz with a period of 1.2s and mark-space ratio 0.65:0.35 respectively.

the wailer is used on long stretches of road, changing to yelper as the vehicle approaches busy turnings or roundabouts. The two-tone siren may be used simultaneously with these, but is usually used when navigating through slow-moving traffic.

Measurements of noise within the police vehicle were made using an A-weighted noise meter located approximately in the centre of the vehicle cabin, and a DAT recorder attached to one of two arrangements of electret microphone selected via a switch. The first arrangement located a microphone at the entrance to each pinna (outer ear) of the driver, whilst the second arrangement located one microphone at the entrance to the drivers left pinna and one microphone on the surface of the radio system enclosure (itself located similarly to a factory-fitted car radio).

Table 2.1 contains a summary of the average A-weighted noise measurements obtained in the vehicle interior under the given conditions. Note that the weather was warm and dry at the time with little wind.

Condition	Noise ($\mathrm{dBA_{SPL}}$)
police vehicle stationary[1]	
ambient noise	50
siren on (loudest)	74
and window open (loudest)	78
police vehicle moving[2]	
30mph with radio call (long T_c)	70
60mph	72
70mph	77
90mph + wailer	80
90mph + yelper	80
90mph + two-tone	81
90mph window open (all sirens)	84
motorcycle radio (max volume) & siren[1]	112

Table 2.1: A summary of typical noise amplitudes noted from measurements with police vehicles. ([1] located in a walled courtyard at an inner-city police station, [2] typical average for various common A- and B-road surface conditions).

It must bc notcd that the values given in table 2.1 are typical values for the situations given, and are for an A-weighted sound measurement (section A1.3). This means that when the car window is opened at 90mph the noise amplitude is noted as 84dB, when in fact the low frequency noise increased substantially and was actually peaking above 100dB in the 0-100Hz range, a range of frequencies whose influence on the overall A-weighted figure is small.

Points to note in particular are the amplitude increases due to the siren noise (24dB when stationary) and that due to opening the window (3 to 4dB). In addition, a simple measure of noise generated from the siren and radio test (at normal working volume) for a stationary motorcycle indicated extremely loud sound levels. Comments made at the time indicated that police motorcyclists often experienced a degree of hearing loss, both temporary and permanent from everyday usage.

The relevance of these measurements is that noise levels in the situations shown routinely exceed levels at which conversation is normally held, and that noise levels exceed the amplitudes at which increases in speech amplitude improve intelligibility as discussed in section 2.3.1.

3 Speech compression using CELP

3.1 Introduction

This chapter begins by introducing the CELP speech compression algorithm and discusses aspects of CELP relevant to speech enhancement.

CELP is an analysis-by-synthesis speech compression algorithm that has evolved over a number of years. It is the logical conclusion of much research into speech coding as it collects together various techniques of speech modelling, into a coding system that provides good speech quality combined with low-bit rate transmission capabilities [98]. CELP utilises a source filter model of speech and so the algorithmic operations can be compared to the human vocal process.

3.2 A description of CELP

3.2.1 Derivation of algorithms

CELP is derived from basic linear predictive (LPC) coders (section A2.2.2), first applied to speech compression in 1971 [4][5][96]. These rely on the fact that speech is pseudo-stationary over intervals of around 30ms [62], and that linear prediction of such a speech frame can provide a good approximation to the original speech [6][61][62][87][88]. An example of an LPC coder is the US Department of Defense developed Federal Standard 1015 algorithm of 1975. This 2400 bits/second coder is used for military communications.

Analysis of the difference between the linear predicted signal and the real speech, called the residual, reveals a waveform having a spiky shape whenever the speech under consideration is voiced [131]. The spikes or impulses have a period related to the pitch of the speech. This is because the pitch signal is generated in humans by the glottis as a blast of air resembling an impulse train [81], and the linear predictor is incapable of adapting to such sharp transients. Therefore a substantial difference exists between the actual and predicted waveforms, and this difference is predominantly caused by pitch spikes.

The second generation of coders detects the presence or absence of pitch (by detecting if speech is voiced or unvoiced) and the pitch period [87]. For a frame of speech that is voiced, the decoder excites its linear predictor with a train of artificial pitch pulses having the detected period [12]. For unvoiced frames, the excitation is random noise. An example of such a speech coder is the full-rate GSM (Groupe Speciale Mobile) European Cellular

Standard of 1988.

CELP, as the next generation of speech coder, now generally uses pitch information to define a long term predictive (LTP) filter that adds pitch information to the excitation signal. This operates in series with an LPC filter. Most importantly, the CELP coder contains a codebook of artificially constructed signal frames which it uses to model the vector produced when pitch is removed from the residual [87][98]. Some examples of such a coder are the CCITT G.728 standard of 1992 (designed for use in mobile telephony) [14], the US Federal Standard 1016 coder for military and police communication, and the TETRA codec [120].

3.2.2 Operating principle

Although many variants of CELP coder now exist, the basic operating principle of most of them remains the same:

① A frame of speech is analysed to determine its pitch characteristics. These pitch signals are removed from the speech (using an inverse LTP filter), and a linear predictor generates coefficients modelled upon the remainder.

② A large codebook, containing many subframe-sized vectors, presents each in turn for filtering. Two filters operate on each vector: the pitch filter (with parameters constructed using the results from the pitch analysis) and an LPC synthesis filter (with parameters being the just-derived linear prediction coefficients).

③ The result of filtering each codebook vector, which is a synthetic speech vector, is weighted and compared to the actual subframe of original speech.

④ The index number of the codebook vector which results in the synthetic speech vector best matching the actual speech subframe, is transmitted to the decoder along with pitch, LPC and gain parameters, all of which describe the current subframe of speech.

The decoder uses an identical codebook indexed by the received codebook vector number, and the received LPC, LTP and gain parameters to recreate each synthetic speech subframe. These subframes are then joined together in a fashion which matches the subframe splitting process occurring at the coder, to produce high-quality synthetic speech.

In fact, subframes are usually overlapped by 50% or more, and due to a difference in the stationarity of the pitch and linear prediction signals in human speech, the linear prediction coefficients are usually calculated and updated less frequently than the pitch values.

3.2.3 Description of operating process

Fig3.1 shows a block diagram of a simple forward-adaptive CELP coder. The blocks outside the dotted box occur once per speech frame. The codebook search loop processes within the dotted box must occur for each codebook entry, every speech frame (ie. 1024 times as often for a typical 1024-entry codebook).

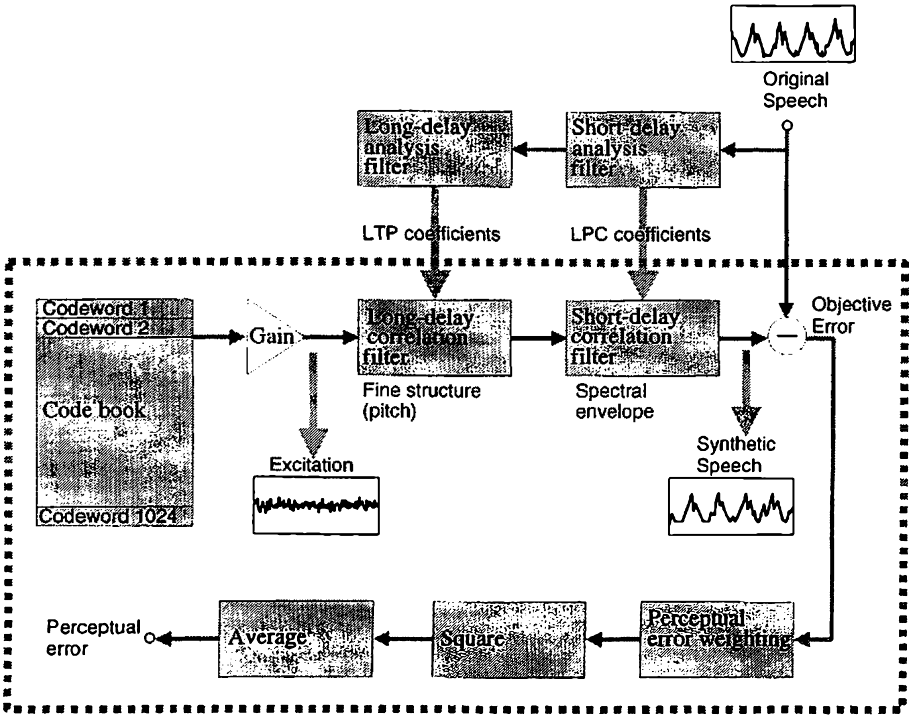


Figure 3.1: Block diagram of simple CELP algorithm. Codebook search loop operations are shown within the dotted rectangle.

Input speech to be encoded is compared with a reference signal of synthetic speech, created by the algorithm. The difference between these two signals is called the objective error. This error is passed through a perceptual weighting filter - which adds an interpretation of human aural perception to the signal. The result, the perceptual error, is an approximate measure of how closely the synthetic speech matches the original speech to a listeners ears.

The synthetic speech is obtained by exciting two filter structures that together mimic the human vocal tract and speech production physiology. The two filters are short (LPC) and long-term (LTP) predictors. The filter excitation signal is derived by amplifying a subframe-sized vector taken from a set stored in a codebook. During operation, each vector in the codebook is filtered in turn and the results compared with the actual speech. The two filters have adaptive coefficients, updated by an analysis of the current speech frame.

The algorithm scans through each of the codebook waveform vectors, filtering and then comparing each to the original speech (but also applying a perceptual bias to the comparison). Once each code vector has been tried, the one that corresponds best with the input speech is transmitted to the decoder along with LTP, LPC and gain values. Analysis then begins on the next input frame.

The comparison process subtracts the artificial from the original speech and then filters the result using a perceptual-error weighting filter [57], which amplifies frequencies located around the speech formant regions and attenuates frequencies away from the formants (the artificial speech having been calculated from the current codebook entry processed by LTP and LPC synthesis filters). The comparison process then calculates the mean-squared error for the current frame, interpreted as being a measure of how closely the two frames match each other to a listeners ears

3.3 A vocal description of CELP

The CELP coder generally mimics the human vocal tract in its arrangement, and is termed a source-filter model (with the source-filters being human-vocal system related). This indicates that parts of the CELP algorithm, similar to parts of the human vocal system, impart distinct characteristics to the resultant speech.

The CELP codebook, which represents the lungs, contains vectors that are usually Gaussian distributed random noise. Reports [98] suggest that a Gaussian distribution matches closely the distribution found on average in the excitation air emitted from human lungs. The air pressure from human lungs is modelled by the CELP system gain parameter, which operates separately from either of the filters (this is partially because each filter operates on normalized values within a typical implementation).

The LTP filter adds pitch to the excitation signal, as do the human vocal chords to air from the lungs. The LPC filter adds a spectral shape in a similar manner to the human throat and mouth. This is illustrated in fig3.2 which shows human vocal actuators grouped in terms of CELP processing blocks.

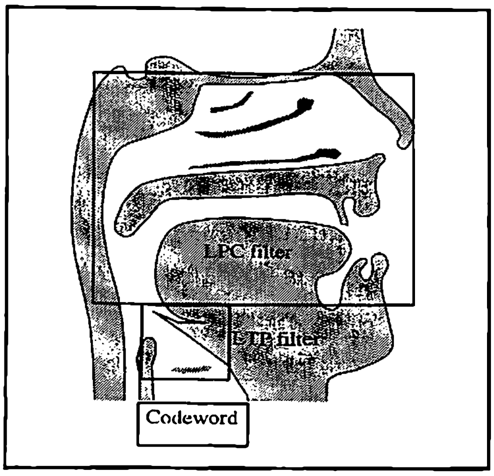


Figure 3.2: Human voice production apparatus with a CELP function interpretation overlaid.

CELP differs from a purely human model, however, in that it processes the speech on a frame-by-frame basis rather than continuously, and the linear prediction model has been found to model air passage through the nasal cavity rather poorly (this is because the LPC filter is an all-pole filter, and the nasal cavity introduces zeros into the equivalent circuit. Although many authors agree that a zero can be modelled with two poles [90] , there is nevertheless a degree of mismatch).

The dependence of CELP on a vocal model does not end with structural similarities: LPC and LTP filters are fixed over their frame periods (usually a frame of 20 to 30ms and a subframe of 5 to 6ms respectively). These periods of time are derived from the maximum rate of movement found in the muscles of the corresponding vocal tract regions. For example, the throat muscles move relatively slowly in speech, and can be assumed to be pseudo-stationary over 30ms, whereas the glottis moves quicker and pseudo-stationarity can only be assumed for around 6ms.

3.4 Speech enhancement with CELP

Speech enhancement is based upon altering speech to match the acoustic conditions in the listener's environment. Alterations to speech are best made in a domain that is relevant to vocal communications. The CELP coder, as described in section 3.3, divides the speech signal into lung, glottis and vocal tract descriptive parameters each of which may be altered to create a speech signal that has been changed in a way relevant to spoken communication.

Furthermore, the CELP coder undertakes a thorough analysis of the spoken signal resulting in the transmitted parameters. These parameters thus encode important information concerning the speech signal, which may be examined by any enhancement system in order to modify enhancements dynamically depending on the speech type.

Some of the aspects of the CELP coder that are relevant to speech enhancement are as follows:

① The long-term pitch filter in CELP requires pitch timing and strength information, thus any pitch-based post-processing scheme can benefit by using the same information, saving calculation. In addition scaling of these parameters may cause the CELP algorithm to pitch shift or scale speech.

② The short-term linear predictive filter uses coefficients that can be analysed to reveal information regarding formants, or can be used on their own to implement spectral sharpening [2][95]. These LPC coefficients encode an efficient spectral representation of the signal.

③ The LPC coefficients themselves are usually transformed to LSP parameters within the CELP encoder, which may be exploited for enhancement as described further in chapter 5.

④ Analysis of LTP and LPC parameters along with gain value can allow distinction between segments of speech such as voiced or unvoiced parts [39][56][79].

⑤ As CELP processes on a frame-by-frame basis, the coded LPC, LTP and gain values are available to any post-processor about 30ms (a typical frame length) before the output from the decoder. Although any such post-processor could introduce a latency of its own, such additional delays are generally unattractive in a working system. An external system using these parameters has them available before the speech is decoded. In implementational terms, this is far more advantageous than the common arrangement where analysis parameters are only available only after speech has been heard and analysed.

To summarise the opportunities for speech enhancement combined with the CELP algorithm, the coder transmits four basic parameters to the decoder for speech reconstruction. These are:

① gain

② LPC coefficients encoded as LSP parameters

③ pitch (LTP) coefficients

④ codebook index

The proposed speech enhancement system would firstly analyse these parameters to gain information on the characteristics of speech being transmitted for the current frame (if indeed speech is being transmitted rather than, perhaps, a pause between words). Then an efficient method of speech enhancement would be an adjustment of these parameters prior to decoding. These parameters are highly condensed representations of the speech waveform and so adjustment of these few values (typically around 5kbits/s) is far more efficient than a process adjusting the decoded speech waveform directly (approximately 128kbits/s).

3.5 Effects of CELP coding on speech

The entire CELP coder is tailored to process speech signals, specifically those produced by an "average" speaker. This causes a CELP coder to process any signal as if it were average speech, transforming non-average speech, and even incidental signals such as noise, into an average, speech-like form. Thus the output from a CELP coder is limited by the following processes:

① The synthesised output is stationary over an LPC analysis frame, irrespective of whether the input, assumed to be pseudo-stationary is actually so (A2.2.2).

② Pitch strength and period (represented by the LTP parameters) is constant over each pitch analysis subframe (A2.2.1).

③ Pitch strength is limited by quantization to upper and lower limits associated with speech, thus non-speech signals are quantized into a speech-like range.

④ Pitch period is also limited by the pitch extraction search process (A2.2.4) to upper and lower limits associated with average speech.

⑤ Gain is constant over each frame, and is limited by the quantization process to that occurring during average speech.

⑥ Filter excitation source is derived from a known codebook entry which was itself chosen to be applicable to average speech.

The processes listed above confer advantages to any post-processing schemes due to the limitations which they create: the CELP coder ensures that all parameters are speech-like to some degree and so any alterations will not be able to adjust speech to outside this range.

Unfortunately the CELP algorithm also acts to degrade speech signals in a number of ways. These include problems associated with the choice of analysis frame. If major changes in sound frequency distribution occur within the span of an analysis frame, then the resultant synthetic frame will be an amalgamation of the two frequency distributions. This averaging effect also occurs when significant changes in signal amplitude are evident, leading to *precursory noise* [46].

Precursory noise, illustrated in fig3.3, arises when an amplitude step occurs during an analysis frame. The synthetic frame, has had constant gain applied to it, causing amplification of the samples within the frame that occur immediately before the amplitude step. Thus the original noise floor before the step has become audible due to a gain contribution from the wanted signal.

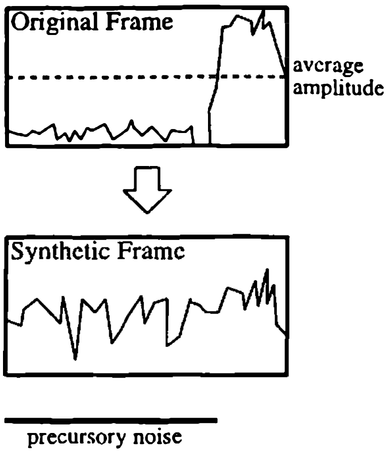


Figure 3.3: Precursory noise due to an inter-frame amplitude step.

Other signal degradations caused by the CELP algorithm include the breakup of rapid changes in pitch, such as those produced by a glissando or slur. Each synthetic frame has a pitch equivalent to an average of the actual pitch and thus if actual pitch changes rapidly then the resultant synthetic output may be a number of pitch steps. A similar effect occurs when formant frequencies change rapidly. This may explain why CELP coders are particularly poor at coding music (as discovered during informal CELP coder listening trials).

4 Enhancement methods chosen for implementation

4.1 Introduction

For a CELP-based communication system, operating in the target police vehicle environment, possible speech enhancement methods were investigated to choose the most promising in terms of their potential to improve performance and in terms of their degree of integration with existing CELP functions.

4.2 Selective amplification

CELP encodes speech on a frame-by-frame basis, with each frame being described by amplitude, pitch, vocal tract and excitation values (gain, LTP, LPC and codebook index respectively: refer to section 3.2 for further information).

An increase in gain value will amplify the contents of the current frame by the required amount. In addition, the transmitted parameters are an information-rich description of the sound contained within that frame, and can be analysed to determine the type of speech, if any, contained within that frame.

Thus, two of the requirements for selective amplification are already present within the CELP coder: an amplification arrangement for short periods of speech and a means of determining the class of speech (i.e., whether the type of speech is applicable for selective amplification).

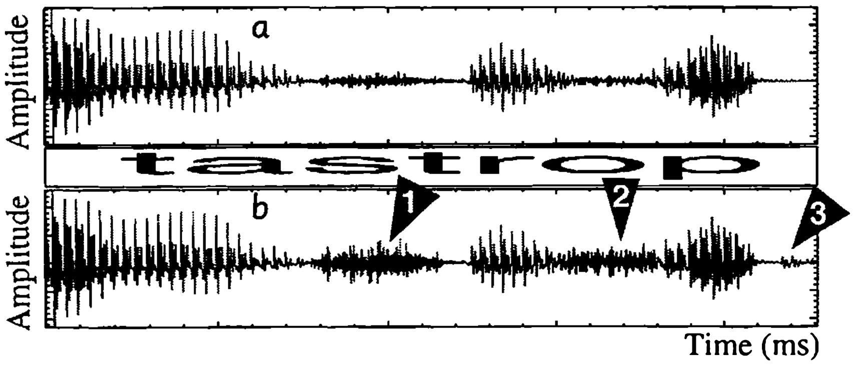


Figure 4.1: Selective amplification of the word 'catastrophe'. a) plots the unmodified speech waveform, and b) the waveform after selective amplification of the phonemes regions 1, 2 and 3.

Selective amplification is illustrated in fig4.1, where a segment of the recorded word ‘catastrophe’ has been classified to determine its phonemic content (the classification method is described later in chapter 6). Three of the regions within the segment were found suitable for selective amplification, and were amplified to give the new waveform as shown.

4.3 Formant sharpening

CELP encodes spectral information in a condensed format, with most of the spectrum represented as linear prediction coefficients or LSP transformations of these. Section 2.2.4 reported some methods of spectral alteration: the information-rich nature of the CELP parameters suggests that spectral alterations may be made by adjusting these few parameters as opposed to the obvious alternative of filtering the entire array of speech samples.

Even if the filtering alternative is considered, operating on decoded speech, the spectral modification filter response can be added to the existing CELP filter response to effect the alteration with a few additions rather than the many multiply-accumulate operations of an additional filter.

The evidence presented in section 2.2.4, and the results of testing (described later in section 9.3) indicate that spectral sharpening is a valid method of improving intelligibility. This process is demonstrated in fig4.2, where a speech spectrum is plotted with a psychoacoustically-weighted background noise spectrum. The latter is the masking level: the level at which a tone must rise above, when listened to in the noise shown, to be audible. The left hand graph in fig4.2 shows that the peak, or formant, in the speech spectrum does not rise above the noise masking level and is thus inaudible. Formant sharpening has narrowed the formant in the right hand graph, and increased its peak amplitude, which now rises above the masking level and is thus audible.

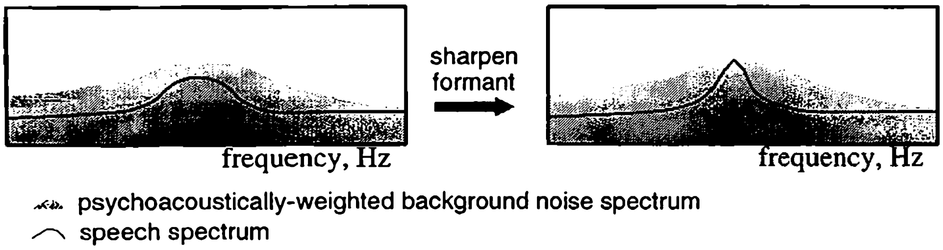


Figure 4.2: Illustration of a masked formant (left) being sharpened to rise above the masking level (right). Assumptions of masking effect and audibility may be made using these plots for average-hearing listeners.

One further effect should be mentioned here: the effect of spreading signals across as many critical bands as possible [69][110][68]. Evidence suggests that when the bandwidth of a wanted signal is present across more than one critical band (section 2.3.2.1), the hearing system is able to utilise correlation effects to improve audibility [54][73][69]. Spreading the bandwidth of a spectral peak is the opposite of sharpening, it is spectral broadening.

So it would seem that spectral sharpening, and the converse, spectral broadening can both increase audibility under certain circumstances. It is likely that spectral sharpening improves the intelligibility of speech in the presence of wideband noise, and that spectral broadening improves intelligibility in the presence of narrowband noise (where one critical band is saturated with the noise, the signal can be extended into an adjacent, non-saturated critical band). This is explored further in section 9.5.

4.4 Formant shifting

As noted in section 3.4, the CELP coder utilises linear prediction coefficients, an information dense representation of the spectrum of the encoded speech. For most phonemes, the speech spectrum contains formants - these formants can be adjusted through spectral modification as discussed.

The concept of masking (section 2.3.2.1) dictates that a tone, or formant, will be inaudible if it is coincident with (or close in frequency to) another tone of higher amplitude. If a frame of speech contains a formant frequency that is coincident with a louder interfering frequency in the acoustic background noise environment of the listener, then this formant will be inaudible, possibly rendering the speech unintelligible.

Spectral modification can be used to adjust the frequency of formants. Evidence suggests that a certain degree of mistuning can be tolerated by the hearing process with little subjective signal degradation [23]. If we consider a vowel, made up of a series of related tones, or formants, it has been found that mistuning a non-fundamental formant by about 8% has a similar perceptual effect to that experienced by removing it from the series [73]. Of course, the sound timbre changes slightly, but this result is equivalent to saying that up to 8% mistuning causes little effect. Furthermore, there is reported evidence which suggests that formant 'mistuning' can, under certain circumstances , improve intelligibility [73].

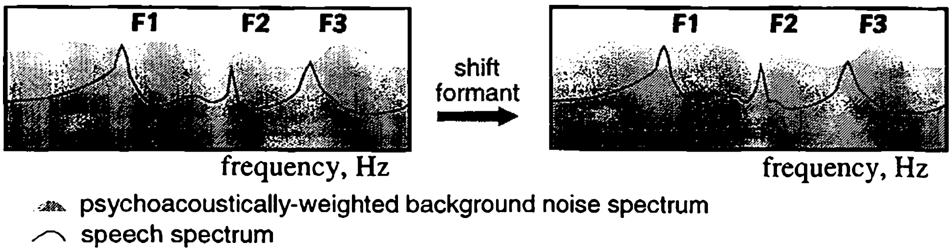


Figure 4.3: Illustration of the spectrum of a speech utterance containing three formants. In the left graph, formants F1 and F3 are audible but F2 is not. F2 has been shifted in the right graph so as to become audible.

The effects of formant mistuning can be beneficial. Consider the speech spectrum of fig4.3, showing three formants in the presence of masking noise. On the left hand graph, formants F1 and F3 are audible, whilst F2 is inaudible. Spectral processing, operating on F2 has altered its frequency in the right hand graph, moving it into a region of lower acoustic background noise where it is now audible.

Of course, any speech alteration, especially a high degree of formant shift, can be expected to degrade the quality of processed speech to some degree [23].

4.5 Rejection of other methods

Although the clipping of speech (section 2.3.3) demonstrated good intelligibility results, it severely reduces the quality of the processed speech. For a CELP coder system, developed in part with the intention of maintaining good speech quality at high compression ratios, and designed to analyse, transmit and reproduce natural speech, clipping the reproduced speech appears a retrograde step. Thus clipping will not be considered further. Selective amplification has been chosen to preserve the naturalness and quality of speech whilst operating on the same intelligibility enhancement premise as clipping (section 2.2.1).

CELP coding systems are generally designed for two-way communication, as is true within the target environment. Time domain processing techniques (section 2.3.4), which add latency are therefore undesirable.

Furthermore, the temporal processing methods of section 2.3.4 require an advance knowledge of the interfering acoustic background noise, so that phonemes may be advanced or delayed slightly around loud periods of noise. It is possible to predict the occurrence of periodic sounds in advance, and to adjust decoded speech playback correspondingly. However, the noise types found within the target environment, and which cause most interference (such as wind, rain, tyre rumble etc.) to listeners in that environment are predominantly non-periodic (section 2.4). Thus temporal methods of speech processing were not considered to be promising enhancement schemes for a CELP communications system operating in the target environment.

PART II: Analysis and Implementation

Part I of this thesis has reviewed existing speech enhancement methods, and proposed additional methods. These have been considered alongside the CELP algorithm for use in the target environment. The CELP structure and the characteristics of the target situation have allowed a subset of algorithms to be chosen for further investigation. These are selective amplification, formant shifting and formant sharpening/broadening.

Part II considers the chosen enhancement techniques and investigates methods of implementing these, and the analyses that are required for their use. Already, it is possible to examine the CELP structure and note the additions required to implement speech enhancement.

FigP2.1 shows a simple block diagram of the CELP speech coder that would be located in the police vehicle of the target environment. The blocks located in the dotted area are those additions to the existing system (shown outside the dotted area) that are required to perform speech enhancement.

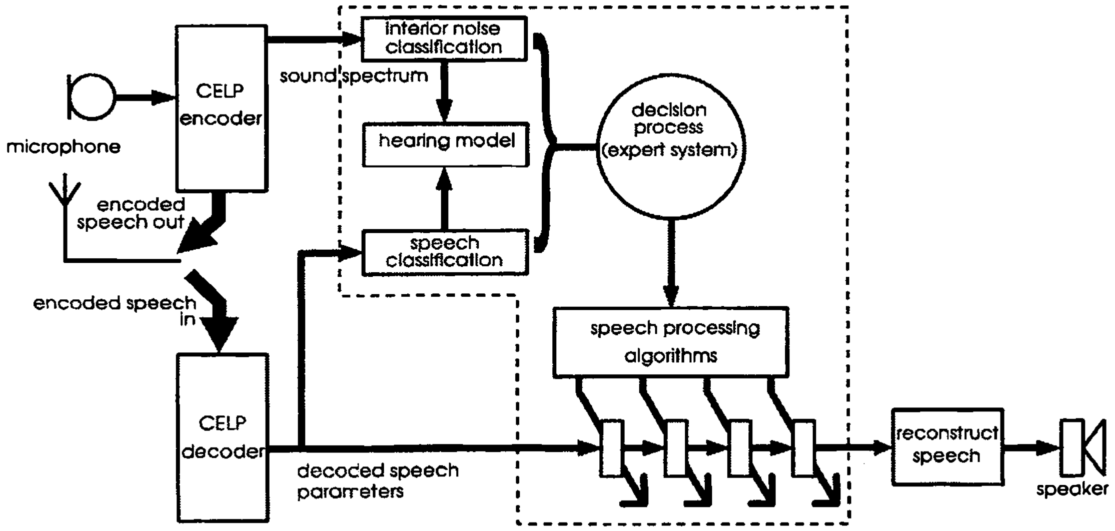


Figure P2.1: Enhancing CELP system block diagram (blocks within the dotted enclosure are enhancing additions to the existing CELP components outside the enclosure).

Part II provides more detail of these operations. It begins, in chapter 5, by introducing the line spectral pair representation, and its relevance to speech enhancement. Chapter 6 then discusses existing and proposed methods of speech analysis and classification, before presenting methods of noise analysis and the definition of a hearing model for speech intelligibility prediction.

Finally, chapter 7 compiles the analysis and speech enhancement methods into a speech enhancing CELP coder: based upon the then completed structure of figP2.1.

5 LSP-based analysis and processing

5.1 Introduction

Modern standard CELP coders, such as the TETRA codec (and also including G.728 [16] and FS1016 [15]) used in the target situation, employ LSP parameters as a means of efficiently encoding LPC coefficients for transmission from coder to decoder [15][120]. LSP parameters are derived via a mathematical transformation of LPC coefficients.

LSP parameters are used firstly because they appear more information-dense than other representations, in that LSPs can be quantized more severely than other representations whilst retaining equivalent levels of speech distortion [37][48]. This is partly a consequence of the LSP property that the parameters are of equal importance to the underlying spectrum, and thus quantization can be equal across all parameters (giving a quantization effect spread over the entire frequency spectrum. LPC parameters by contrast are of unequal importance) [37][48]. Secondly, LSP parameters can be interpolated between frames or scaled, and the LPC filter deriving from the altered parameters is always stable. Whereas, injudicious LPC parameter quantization often results in an unstable filter [90].

It is for their advantages in the transmission of spectral information between CELP coder and decoder that LSP are commonly utilised. In addition, two authors have attempted to perform speech recognition using a feature vector of raw, uninterpreted LSP values for hidden Markov model based speech recognition [30][80].

As the LSP values convey spectral information from CELP encoder to decoder, adjustments to these values before reaching the decoder enables known spectral adjustments to be made extremely efficiently. In this way, LSPs have been found to have specific usefulness to speech enhancement. This chapter discusses the line spectral pair representation before demonstrating some of its properties and relating these to speech enhancement algorithms.

5.2 Line spectral pair representation

Line spectral pairs are parametric representations of linear prediction coefficients, with the useful properties of being resident in the frequency domain and being relatively more resistant to the effects of quantization, shifting and interpolation [51][90].

Line spectral pairs represent the resonances of the all-pole filter model when the standard form of the PARCOR (or reflection coefficient) representation of linear predictor is advanced to an extreme conclusion. The PARCOR process compares the linear predictor to a system of joined pipes of constant length but differing width (the number of pipes being determined by the order of the system). The PARCOR coefficients encode the degree of back-reflected energy from each pipe join, where the entire system is assumed to be perfect, but the back-reflected energy is then assumed to be lost.

If the PARCOR representation is modified to form two distinct cases where the start of the system of tubes is either a perfect opening or a perfect closure, then a series of standing waves are set up; equivalent to the resonances or the poles of the linear predictor. The frequencies of these resonances are the line spectral frequencies. Pairs of line spectral frequencies, one from the fully-open case and one from the fully-closed case then act together to define the peaks and troughs of the underlying spectrum.

To relate this in some way to reality, imagine that the system of interconnected tubes is representative of the vocal tract. The vocal tract begins at the rapidly opening and closing glottis, and although the throat muscles are pseudo-stationary over the period of analysis (for example, a 30ms frame period), the glottis is not. The two analysed cases of open and closed tubes thus correspond to open and closed glottis - in fact the glottis is neither fully open nor fully closed and thus the real spectrum (the resonance of the system of tubes) occurs between the two cases. Fig5.1 illustrates the line spectral pair representation of a frame of sound. The LSPs derived from the open and closed tube conditions are shown as dotted and continuous vertical lines respectively overlaid on a spectral plot. It can be seen that the peaks in the spectral plot (spectral resonances) occur between closely spaced dotted and continuous lines.

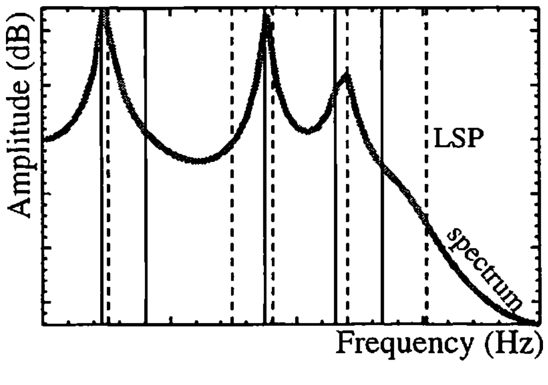


Figure 5.1: LPC - derived spectrum and corresponding line spectral frequencies. Data under analysis was derived as in figure 5.4, section 5.5.

5.3 Evaluation of LSPs

It is apparent through simulation and testing that line spectral frequencies track the underlying spectrum in a manner that suggests a predictable relationship, this is substantiated by Paliwal [80].

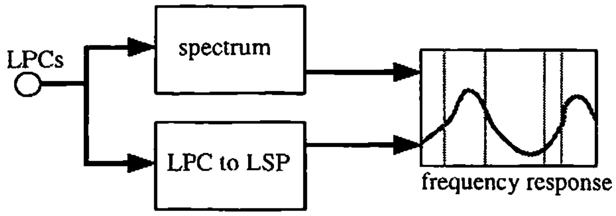


Figure 5.2: Method used to visualise LSP relationship to LPC spectrum.

Fig5.2 shows a method to plot line-spectral frequencies overlaid onto a linear prediction-derived spectrum. Fig5.1 is such a plot, showing the LPC frequency response (section A3.3) and corresponding line spectral pairs, symmetric being shown solid and anti-symmetric shown dotted, for a spoken vowel analysed using a 10th order linear predictor.

This method of visualising LSPs can be extended one step further as shown in fig5.3, where the effect of changing some or all of the LSPs for a given analysis frame of sound can be compared by noting differences between the original and altered spectra.

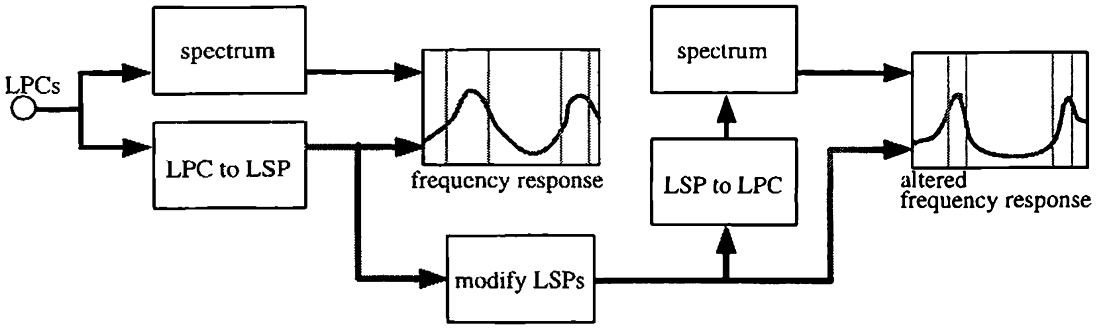


Figure 5.3: Method used to visualise the spectral effect of altering LSPs.

5.4 Important mathematical properties

The derivation of LSP parameters is given in appendix 3, however, certain properties of LSPs are important to their understanding. These are now examined mathematically.

Consider A_k and B_k the symmetric and antisymmetric LSP polynomials made up from the p LPC coefficients a_k (with initial condition $A_0 = 1$ and $B_0 = 1$ and assuming that the order p is even). The expressions relating LSP and LPC coefficients are described in section A3.1, and are reproduced here:

$$A_k = a_k - a_{(p+1-k)} + A_{k-1} \tag{5.1}$$

$$B_k = a_k + a_{(p+1-k)} - B_{k-1} \tag{5.2}$$

A theorem developed by Sugamura and Itakura [108] states that;

$$a_p = \frac{1}{2(A_k + B_k)} \tag{5.3}$$

where a_p is the linear prediction polynomial. If we now calculate the power spectrum from the linear prediction parameters;

$$\left|H(e^{j\omega})\right|^2 = \frac{1}{\left|a_p(e^{j\omega})\right|^2} \tag{5.4}$$

$$= \frac{4}{\left|A_k(e^{j\omega}) + B_k(e^{j\omega})\right|^2} \tag{5.5}$$

and using the following expressions for $A(z)$ and $B(z)$ [107];

$$A(z) = (1 - z^{-1}) \prod_{i=2,4,\ldots\ldots,p} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{5.6}$$

$$B(z) = (1 \mid z^{-1}) \prod_{i=1,3,..p-1} (1 \quad 2z^{-1}\cos\omega_i + z^{-2}) \tag{5.7}$$

with the usual assumption that the array of LSPs, ω_i are ordered least first and are constrained to angular frequencies of between 0 and π.

Now eqns5.6 and 5.7 may be substituted into equation 5.5 to give;

$$\left|H(e^{j\omega})\right|^2 = 2^{-p} \Bigg/ \left\{ \sin^2\frac{\omega}{2} \prod_{i=2,4..,p} (\cos\omega - \cos\omega_i)^2 + \cos^2\frac{\omega}{2} \prod_{i=1,3,..p-1} (\cos\omega - \cos\omega_i)^2 \right\} \tag{5.8}$$

If the above equation were evaluated for all ω in the range 0 to π, the result would be the linear prediction spectrum. Note however that this spectrum reaches a maximum when the terms of the denominator $(\cos\omega - \cos\omega_i)$ are minimum for adjacent values of i.

Imagine that the equation is evaluated at a certain frequency ω. If line spectral value ω_2 is located here, then the left hand side of the denominator (even LSPs) will be zero. It then requires one of the odd-valued line spectral parameters to be close to this frequency for the right hand side of the denominator to also approach zero, and the spectrum to peak. This demonstrates the mathematical basis for assuming that peaks in the linear prediction spectrum are located where two adjacent LSPs are closest.

In addition, the sin^2 and cos^2 terms in this equation indicate that each half of the denominator is zero, respectively at values of angular frequencies of 0 and π. This property indicates that single lines approaching values of 0 or π cause spectral peaks.

5.5 Properties of line spectral pairs

In order to appreciate the properties of LSPs further, the method outlined in section 5.3 has been used to compare the spectra derived from the linear prediction coefficients of a typical sound with that resulting from LSP-adjusted linear prediction coefficients.

The reference spectrum, with which the figures from each LSP modification are compared has been derived from tabular test vector data presented in [86], and shown in fig5.4. This consists of 10th order linear predictive analysis, and thus 10 LSP values are plotted, for clarity, as vertical lines overlaying the spectrum.

It should be noted that one of the properties discussed in section 5.4, that lines located close together straddle spectral peaks, can clearly be seen in fig5.4 and each of the subsequent plots.

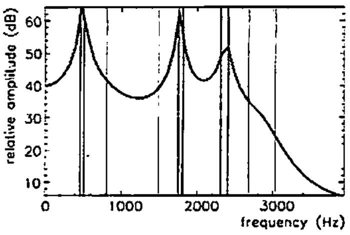


Figure 5.4: Reference spectral plot with overlaid LSP values.

The following subsection plots the spectrum with overlaid LSPs resulting from an alteration in the LSP values at the "modify LSPs" stage in fig5.3. The plots are those from the right hand graph in fig5.3, whilst the left hand graph is that of fig5.4.

5.5.1 Frequency shifting and scaling

The results of slightly separating the first pair of lines by increasing the frequency of line 2 by 20% are shown in fig5.5:

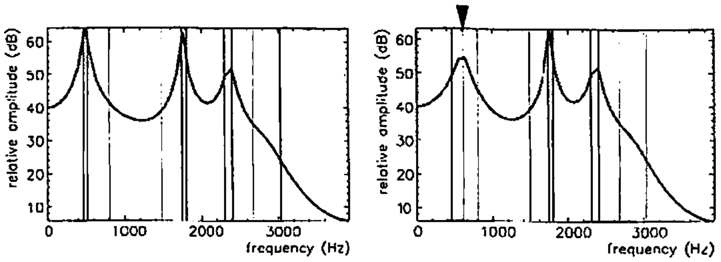


Figure 5.5: Spectrum obtained by widening first pair of lines (right) compared to original reference spectrum (left).

Note that the first peak of the spectrum has reduced in amplitude and spread out, indicating that separation of the LSPs affects both the amplitude and the bandwidth of frequency peaks. More specifically, widening the LSP separation reduces amplitude and appears to increase the bandwidth.

Next, both LSP 1 and 2 were increased in frequency by 20% of their original values, and the results plotted in fig5.6:

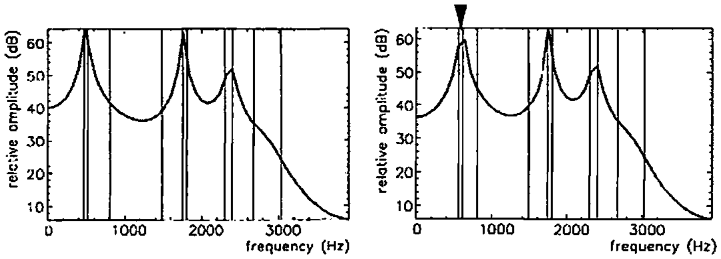


Figure 5.6: Spectrum obtained by applying upward shift in frequency of lines 1 and 2 (right) compared to original reference spectrum (left).

This change has resulted in the lowest frequency peak in the spectrum increasing in frequency by around 20%, but a slight reduction in amplitude is evident, perhaps due to some of the energy constrained between the line 1 and line 2 pair bleeding into line 3 which is now slightly closer.

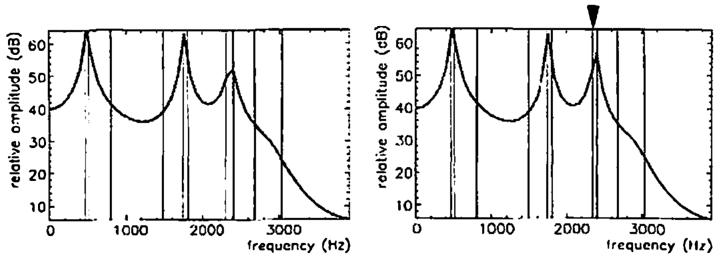


Figure 5.7: Spectrum obtained by increasing frequency of line 7 (right) comparod to original reference spectrum (left).

In fig5.7 the frequency of the seventh line has been increased. Note that the third frequency peak has become sharper, consistent with the effect of narrowing the separation between the pair of lines 7 and 8.

Finally, lines 7 and 8 were both increased in frequency from their original values, and plotted in fig5.8:

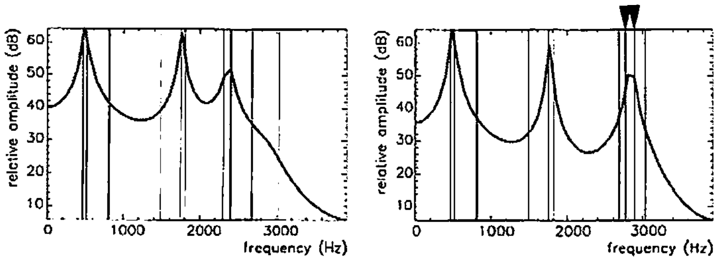


Figure 5.8: Spectrum obtained by shifting line spectral pair 7/8 upwards (right) compared to original reference spectrum (left).

This has caused the third spectral peak to increase in frequency, but has reduced in amplitude. The example is more complicated than is initially obvious, as the two lines have been shifted to *between* the pair of lines that were 9 and 10. The self-ordering property of LSPs has ensured that stability is maintained.

5.5.2 Spectral peak induction

Using the same test data, the frequencies of LSPs 3 and 4 were changed to narrow their separation. The effect of this is shown in fig5.9.

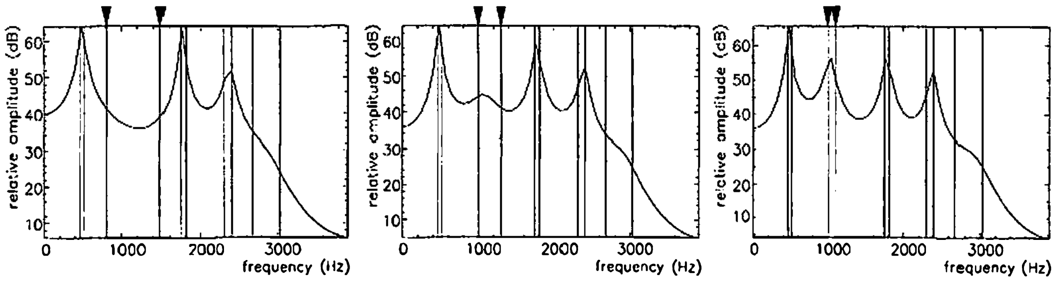


Figure 5.9: The effect of decreasing the separation of widely-spaced lines is shown with the separation of the second and third lines, indicated by arrows, decreasing from left to right.

This operation has resulted in the generation of an additional frequency peak as the lines close towards each other, with the penalty of causing a slight reduction in the amplitude of the existing peaks.

5.5.3 Spectral peak formation by addition

Apart from inducing the formation of non-existent spectral peaks by shifting lines, it is possible to add new line spectrum frequencies to an existing system. This, however, creates a potential problem in that the order of the LPC filter describing the system must be increased in accordance with the number of lines added.

Fig5.10 shows the spectrum obtained by the addition of two extra line spectrum frequencies.

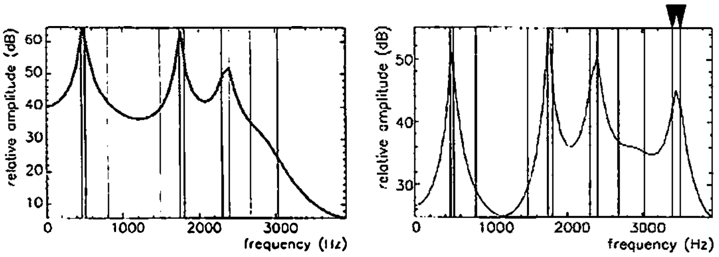


Figure 5.10: Spectrum obtained by adding two lines to the set of test LSPs (right) compared to original reference spectrum (left).

5.5.4 Further LSP properties

Another property of LSPs is that movement of a line towards zero, or addition of a line close to zero, will result in an increased DC component in the spectrum; it appears that a line equivalent exists at a frequency of 0Hz. This has been explained in section 5.4.

Mathematically, LPC coefficients derived from LSPs are guaranteed to be stable [30], however the audible consequences of altering LSPs are difficult to predict. It must be remembered that LSPs operate in the frequency domain, and are thus not directly related to audio quality in a perceptual manner.

The tests resulting in the plots shown in figs5.5 to 5.10, and other similar tests have revealed that inopportune alteration of LSP values can have effects that are difficult to predict accurately. Listening to the results of these tests has reinforced the belief that modification of line spectral pairs can either reduce or increase audio quality considerably.

5.6 The application of LSPs to enhancement

Line spectral pairs, being resident within many CELP codecs (including the TETRA variant present in the target situation), and containing important spectral information, can be utilised in a number of ways to perform enhancement.

5.6.1 LSP alteration in CELP coders

The figures presented in section 5.5 have shown the effects on test spectra of alterations to the line spectral pairs representing them. In fact, the reference test data of fig5.4 was derived from one frame of a speech utterance [94].

The test spectrum is thus known to represent a segment of voiced speech. In fact the three spectral peaks are actually the formant frequencies F1, F2 and F3 for that utterance. With this in mind, alterations to line spectral pairs have been shown to cause formant shifting, formant sharpening/broadening, the addition of extra formants, or the replacement of corrupted ones.

One of the inherent advantages of LSPs is that although the lines do not have any direct perceptual basis, lines usually cluster around formant locations. It is known that changes to lines predominantly affect the immediate frequency region of the line and thus LSP changes are mainly formant related: and are thus, after all, perceptually biased.

Fig5.11 shows a block diagram of one channel of a CELP communications system, with encoder on the left and decoder on the right, and parameters, including LSPs passing from encoder to decoder.

In addition to the standard CELP arrangement, fig5.11 shows an LSP adjustment process, working on the parameters transmitted from coder to decoder, before these are utilised by the decoder. The LSP adjustment process can cause changes in the speech spectral envelope, with these changes being dependant upon the current speech and interfering acoustic background noise (the dependence upon background noise being the reason for locating the enhancements at the decoder).

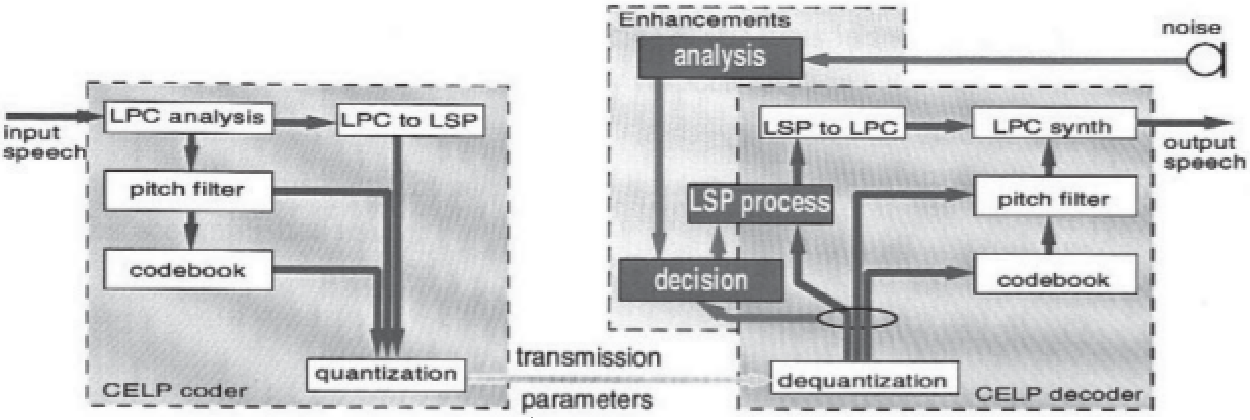


Figure 5.11: A diagram of one channel of a CELP communications system, showing speech enhancement additions performing LSP processing, requiring acoustic background noise and transmitted speech analysis information.

This section will now consider the content of the 'LSP process' block of fig5.11 in greater detail.

5.6.1.1 Formant amplitude and bandwidth adjustment

Considering two nearby LSPs, located either side of a spectral peak: further reducing the frequency of the lower frequency line by a small amount, and further increasing the frequency of the higher frequency line by a small amount will increase the separation between the lines. As has been demonstrated in section 5.5.1, the effect of this on the output spectrum is to reduce the amplitude of the spectral peak and to increase its bandwidth. Similarly, narrowing the lines will increase the peak amplitude and decrease the bandwidth of the spectral peak.

Where the peaks described by the spectrum correspond to speech features, they are usually formant frequencies. If such peaks are identified and the values of the nearby LSPs determined, formants can be adjusted.

If the line spectral pairs describing a period of speech are ω_p (where p is the number of LSPs, equivalent to the order of the linear prediction analysis filter from which the LPC and LSP coefficients were originally derived), and a formant is known to exist at an angular frequency of γ, the line spectral pairs relating most strongly to that formant are:

$$\omega_l = max(\omega_i) \text{ in the range } 0 \leqslant \omega_i < \gamma \qquad (5.9)$$

$$\omega_h = min(\omega_i) \text{ in the range } \gamma < \omega_i \leqslant \pi \qquad (5.10)$$

$$with\ i = 1 \ldots p$$

Where ω_l and ω_h are the angular frequencies of the lines immediately below and immediately above the formant frequency respectively.

Formant peak adjustment, by altering the separation between the two existing lines (ω), can be accomplished using:

$$\omega_l' = \omega_i - 0.5\lambda(\omega_h - \omega_l) \tag{5.11}$$

$$\omega_h' = \omega_h + 0.5\lambda(\omega_h - \omega_l) \tag{5.12}$$

where ω' are the altered values and λ is the fractional change in separation, positive values causing formant sharpening and negative values causing formant broadening.

Naturally there are limitations to this procedure, discussed further in section 5.8 Fig5.12 illustrates the effect on a speech spectrum of narrowing the lines describing one formant using eqns5.11 and 5.12 with a narrowing factor, λ, of 0.2, resulting in an increase in F1 amplitude around of approximately 5dB.

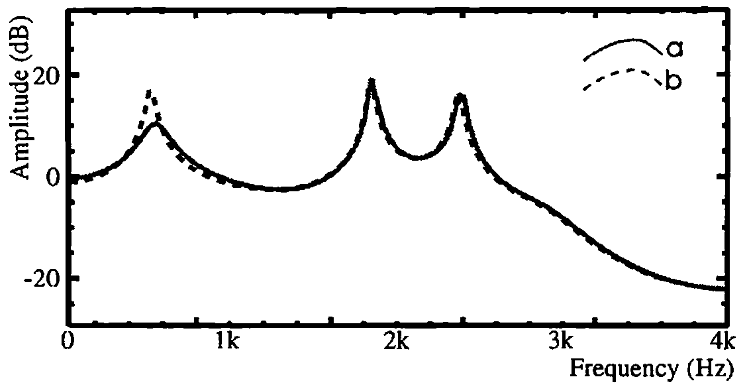


Figure 5.12: Amplitude-normalized plots of original (a) and LSP processed (b) speech spectra. Processing involved narrowing the separation between the pair of LSPs describing the lowest frequency spectral peak.

5.6.1.2 Formant frequency alteration

In a similar way to formant amplitude adjustment, spectral peak centre frequencies may be altered by moving the two LSP lines located immediately to either side of the peak (section 5.5.1).

Within limits further discussed in section 5.8, if a spectral peak or formant is located at an angular frequency of γ and the two LSPs describing it found from eqns5.9 and 5.10, then the spectral peak centre frequency may be altered by:

$$\omega_l{}' = \frac{\omega_h}{2}(\mu - 1) + \frac{\omega_l}{2}(\mu + 1) \tag{5.13}$$

$$\omega_h{}' = \frac{\omega_h}{2}(\mu + 1) + \frac{\omega_l}{2}(\mu - 1) \tag{5.14}$$

$$with \mu = 1 + \sigma\{\pi - 0.5(\omega_h + \omega_l)\}/\pi \tag{5.15}$$

where σ is the degree and direction of frequency scaling (positive is frequency increase, negative is decrease, with the value giving the increment factor) . The equations allow alteration in centre frequency, but preservation of the separation between lines, and thus giving a frequency shift but minimising any unwanted amplitude variations. In addition, eqn5.15 applies a reduction in degree of shift as the maximum frequency value, at an angular frequency of π, approaches. The identities may be applied to a single formant (via an adjustment to one pair of lines), or to the entire set of LSPs, thus scaling the entire speech spectrum.

Note that maximum and minimum frequency positions exhibit the characteristics of a line presence (see section 5.4 and 5.5 for mathematical proof or graphical demonstration), and thus adjusting formants too far upwards or two far downwards in frequency may locate these close to angular frequencies of π or 0 respectively, and cause spurious spectral peaks. Therefore limits must be established to excessive downward LSP frequency scaling as well as upward.

An alternative scheme is to introduce a perceptual scaling to LSP shifting. In this case, when all lines are adjusted, the arithmetic and geometric relationship between them is altered, but the perceptual difference is maintained. This relies upon shifting lines by a constant Bark [102] value, δ, as discussed in section A1.4:

$$\omega_l' = \frac{2 \times \pi}{fs} \times 600 sinh\{(b_k + \delta)/6\} - 0.5(\omega_h - \omega_l) \tag{5.16}$$

$$\omega_h' = \frac{2 \times \pi}{fs} \times 600 sinh\{(b_k + \delta)/6\} + 0.5(\omega_h - \omega_l) \tag{5.17}$$

where fs is the sampling frequency and b_k is the bark value of the centre frequency of the lines [42]:

$$b_k = 6log\left\{c + \sqrt{(c^2 + 1)}\right\} \tag{5.18}$$

$$and \quad c = \frac{(\omega_h - \omega_l)}{2.4\pi \times 10^7} \cdot fs \tag{5.19}$$

The sampling frequency is a required input to eqns5.16, 5.17 and 5.19 because the Bark scale is non-linear with respect to frequency - and is thus not convertible to and from angular frequencies, which convey no absolute frequency information.

The Bark-based and linear LSP frequency shifting schemes are illustrated in fig5.13, with the Bark-based scheme additionally having a hard cutoff at around 3000Hz (once the shifted value exceeds 3000Hz, no more upward adjustment is possible), found by subjective testing to give good results for speech formant scaling.

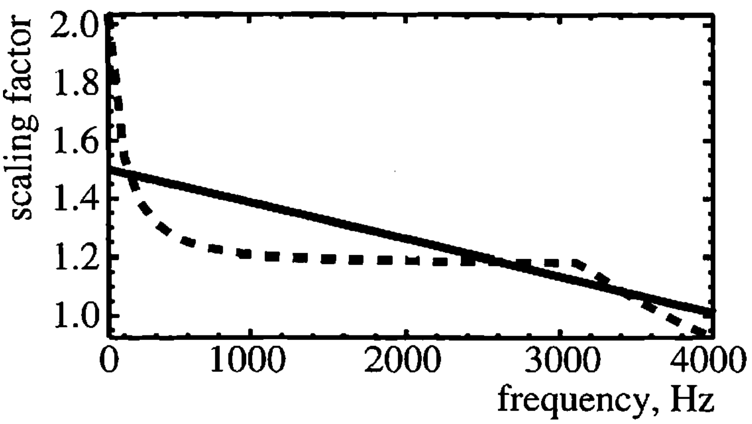


Figure 5.13: Effective LSP upward scaling factor for linear (solid line, with μ=1.5 from eqns 5.13 and 5.14) and Bark-based (dotted line, with δ=1 in eqns5.16 and 5.17) shifting of LSP centre frequency.

5.6.2 LSP-derived measurements

Given that the line-spectral frequency arrangement of a frame of speech relates closely to the speech spectrum, it is possible to use LSP information empirically to form general conclusions about the associated frequency spectrum.

Fig5.14 shows the speech waveform resulting from a recording of a North American female saying "oval face without an expression in the world", a manual analysis of the speech and a plot of the corresponding LSPs derived from a 160 sample frame, 10th order Hamming-windowed LPC analysis of the recording. The data, in the form of a speech recording, has been taken from the TIMIT database [118].

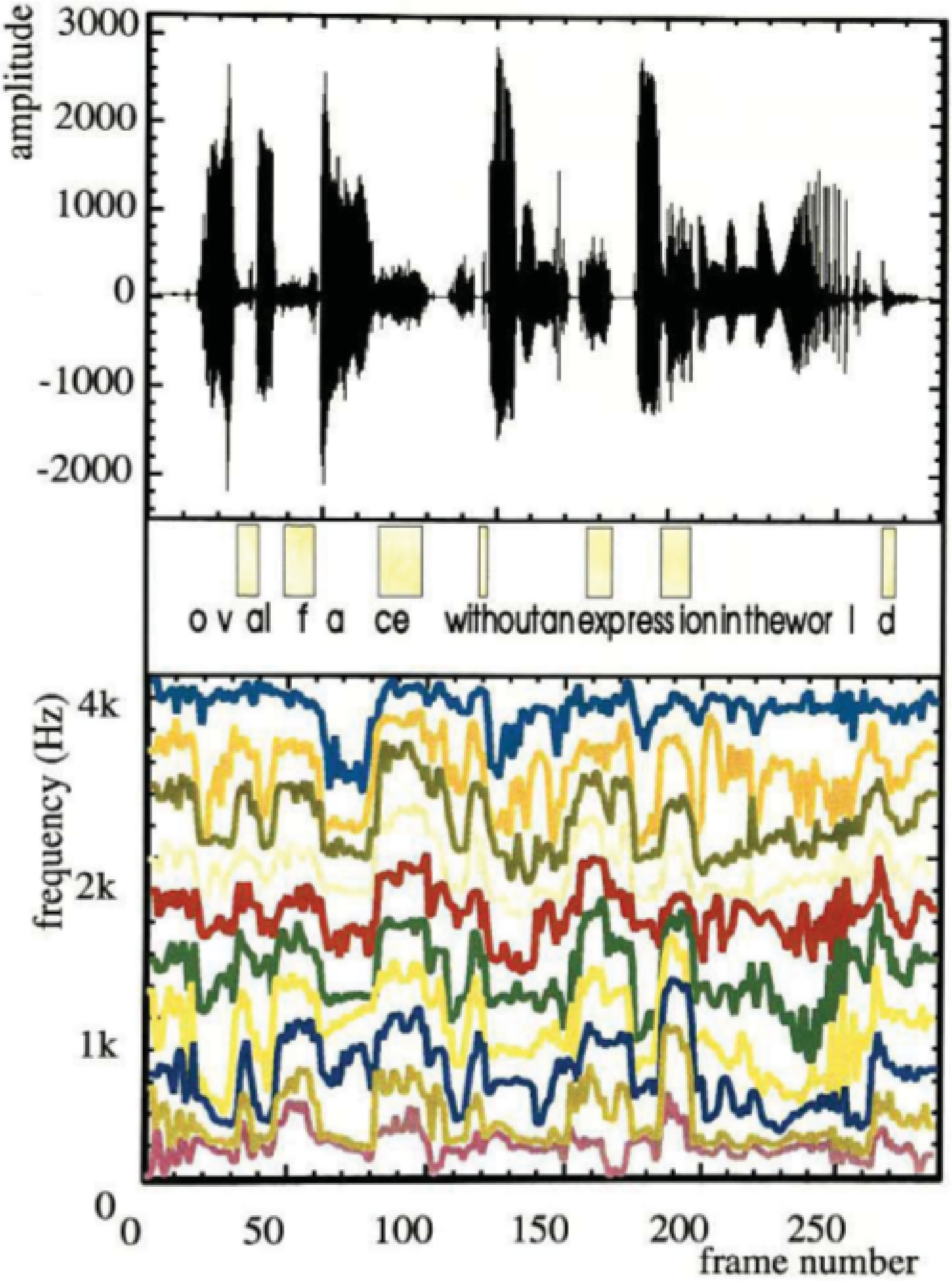


Figure 5.14: Speech waveform, manual analysis and corresponding LSP tracks.

Several trends are visible in fig5.14 involving the variation in LSP frequency over time. Most noticeably, once a period of voiced speech leads into a period of unvoiced speech (indicated by bars below the speech waveform plot and above the transcription), there is a general upwards shift in all of the line spectral frequencies. Tests (presented in section 8.3) confirm that the measurement of LSP values can be used to automatically distinguish between some of the constituent parts of speech.

5.6.2.1 Gross change in LSP value

Fig5.15 shows a measure of the sum change in LSP value between frames for the test recording. Note the correspondence of the measure to the phonemes, in that spikes occur coincidentally with changes in the waveform.

This measure, for frame *i*, in a *p*th order system with current LSPs ω_i is:

$$msr_i = \left\{\sum_{j=1}^{p} \omega_j^i\right\} - \left\{\sum_{j=1}^{p} \omega_j^{i+1}\right\} \tag{5.20}$$

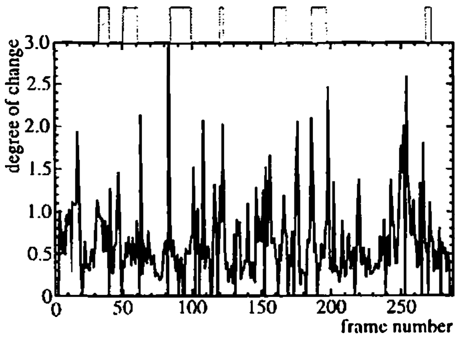


Figure 5.15: Gross change in LSP value over time, with manual analysis bars indicating the presence of unvoiced speech.

5.6.2.2 Deviation from median position

Any period when speech is not present tends to be described by LSPs that are approximately equally spaced and placed around their median positions (those that divide the frequency range equally, and which describe a totally flat spectrum). In a corresponding fashion, deviation from median positions can indicate the presence of unvoiced speech (causing a general increase in frequency) or of voiced speech (causing a general decrease in frequency). In fact, the comparison to median position, for a flat spectrum, may logically be replaced by any basic LSP location comparison. For example, in known interfering noise, the comparison positions could be made equal to the interfering noise spectrum in order to minimise its effect on the resultant measure. This measure is the sum of the square of the deviation.

If the *p* comparison LSPs are ϖ (in radians), then for median positions these would equal:

$$\varpi_j = j \times \pi / (p+1) ; \qquad j = 1 \ldots p \tag{5.21}$$

and the measure would thus be:

$$msr = \sum_{j-1}^{p} (\omega_j - \varpi_j)^2 \qquad (5.22)$$

Fig5.16 shows a plot of the total separation of the distance squared between LSPs in each frame and their median values.

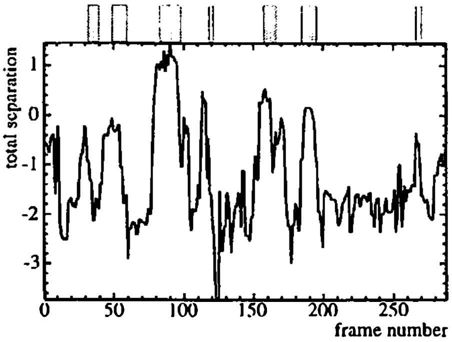


Figure 5.16: LSP deviation from median position over time, with manual analysis bars indicating the presence of unvoiced speech.

5.6.2.3 Average LSP distribution

A measure of the average LSP value indicates in a broad sense whether the spectrum is top-heavy (unvoiced) or bottom heavy (voiced) in that frame.

The measure is thus given by:

$$msr = \frac{1}{p} \times \sum_{j=1}^{p} \omega_j \qquad (5.23)$$

This measure works surprisingly well and is demonstrated in fig5.17, which plots data very similar to that in fig5.16:

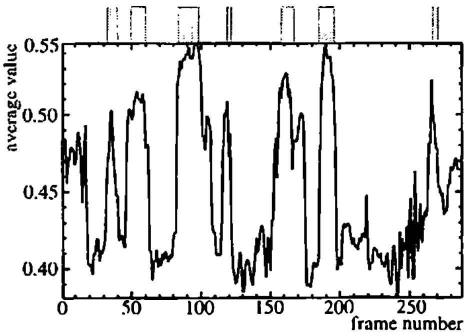


Figure 5.17: Average value of LSP within each frame, with manual analysis bars indicating the presence of unvoiced speech.

5.6.2.4 LSP vote measure: count of LSPs that increase

A subtly different measure is to count the number of line spectral frequencies which have increased from their median position (an increase in frequency often indicates a period of unvoiced speech). This vote measure again relies upon the LSP median positions to be found (eqn 5.21).

For each of the LSPs positioned above its median position, one vote is cast. The number of votes collected in each frame (up to 10 for order p=10) is thus the measure value.

Fig5.18 shows a count of the number of LSPs per frame which have moved above their median value.

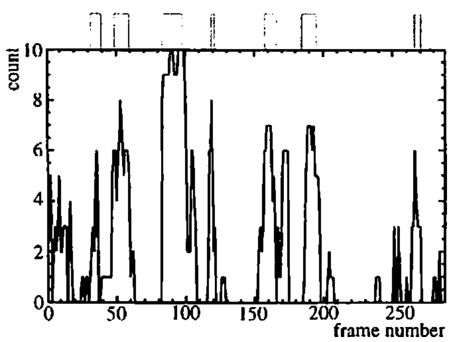


Figure 5.18: Count of LSPs per frame above median value, with manual analysis bars indicating the presence of unvoiced speech.

5.6.2.5 Summary of LSP measures

As noted in section 5.2, the separation of LSPs tends to indicate the presence or absence of frequency peaks or formants. Indications are that measures involving the summation of the few closest LSPs or a summation of the few most separate LSPs are, on average, able to distinguish between voiced, unvoiced speech or periods of silence. However, the thresholds and decision regions required to convert such raw measures into accurate indicators for speech class must be the subject of further analysis, as discussed in section 6.2.

5.7 Information required to enable processing

The LSP processing schemes described in section 5.6.1 are designed to adjust LSP values to influence the shape of the underlying spectrum. If these techniques are to be used to alter formant frequencies then the position of each formant must be known, to establish which LSPs lie to each side of the formant, and consequentially describe it.

Many methods have been developed and published to determine formant frequency. These include the obvious frequency domain techniques of power spectrum peak-picking, spectrum differentiation and certain time domain techniques. Chapter 6 will discuss formant detection more thoroughly.

Whilst the processing algorithms operate on line spectral pair values, and it is known that LSP locations are related to formant positions (see section 5.6.2), with a line pair positioned around each formant peak, it must then be possible to predict formant location from an analysis of LSP distribution.

In general, the centre frequencies of the narrowest pairs of lines correspond closely with the formant frequencies (as determined by alternative methods). More specifically, if formants are present, the three narrowest LSP pairs correspond to the three highest power formant positions. Chapter 6 investigates this novel and other alternative formant detection methods.

5.8 Limitations and advantages of LSP-based methods

5.8.1 LSP-based processing

Consider that line-spectral pair processing provides a method of inducing specific localised changes in the underlying frequency spectrum. A possible disadvantage, that changes are quite localised to LSP position, is irrelevant to this formant alteration application, because some LSPs will always be located close to formants: the regions requiring change.

For a typical implementation, only the six lines corresponding to the first three formant locations would be altered, although if an entire spectral shift was wanted, then all lines (usually no more than 10 or 12) would be altered. The most common alternative method of spectral alteration is by filtering - requiring multiply accumulate operations on each of the samples in the current analysis frame (usually around 240), a much less efficient method. In addition, the LSP method is by nature adaptive, because LSPs are already positioned around formants: creating an adaptive filter alternative is considerably more complex.

The processes of formant sharpening/broadening and formant shifting are limited in extent by their effects. Too great a formant broadening will cause that formant to cease to exist, whilst extreme formant sharpening could produce a sound resembling a tone rather than speech. Formant shiftings (and broadening) are limited by the observation that too great an adjustment to LSP lines may produce a situation where two previously distant lines become close. The spectrum will then exhibit a spurious peak (as demonstrated in section 5.5.2).

LSPs that are moved to approach angular frequencies of 0 or π will induce low or high frequency peaks respectively. Thus the functions describing LSP shifting were designed to reduce degree of shifting as these limits approach.

Rather narrower than the above limitations on LSP adjustment, is the extent to which values can be adjusted without the sound quality degrading excessively. Extreme sharpening, broadening, and shifting will produce sound that, although originally speech, is no longer recognisable as such. To ensure that the listener does not notice changes, formant shifting should be limited to around 8% (section 4.4), however subjective listening tests have established that a shift of up to 20% can maintain speech quality within acceptable limits, improve intelligibility under certain circumstances, and allow the speech to remain recognisable (discussed further in section 7.4). Such results are presented in chapter 9.

5.8.2 LSP-based analysis

In addition to adjusting the characteristics of speech, LSPs have been shown in this chapter to convey spectral information. Section 5.6.2 introduced several measurement techniques based on interpretation of LSPs. In fact, one of the few publications based around LSPs involved interpretation of LSP position. Erzin et. al. [30] used a feature vector composed of raw line spectral parameters to drive a Hidden Markov Model based speech recognition system, and although his system did not rely upon a predetermined LSP interpretation, the positive results indicate that the Markov model was able to construct a good internal recognition framework.

It is probable that, similar to alternative spectral measures, LSP based analyses would need to be combined with other signal measurements such as zero-crossing rate or band power calculations in order to construct an accurate speech classification system. LSP-based analysis is, however, extremely efficient, with around 10 to 20 operations required per frame for the measurements described in section 5.6.2 (on a typical 10th order analysis): even the efficient AMDF (average magnitude difference function) analysis requires at least 240 operation per frame (for a typical 240 sample analysis frame) [1].

5.9 LSP analysis and processing summary

This chapter has introduced the line spectral pair representation, discussed relevant properties and related these to signal processing and analysis. The outcome has been a number of LSP-based techniques relevant to speech enhancement. These are either directly relevant, by performing spectral alterations, or are indirectly relevant by aiding in the analysis of signals (speech or noise) to allow the speech enhancements to adapt to the type of speech being decoded and the type of background acoustic noise in the target situation. The methods are as follows:

① Formant frequency alteration through adjusting the LSPs describing a particular formant (section 5.6.1.2).

② Spectral frequency alteration, through adjusting the LSPs describing the entire spectrum (section 5.6.1.2).

③ Formant broadening or sharpening, through narrowing the gap between certain pairs of lines (section 5.6.1.1).

④ Formant detection, by measuring the gap between adjacent lines (section 5.4 and 5.5).

⑤ Speech detection, by measures in section 5.6.2 combined with existing indicators.

⑥ Speech classification (voiced/unvoiced/fricative etc..) by measures in section 5.6.2 combined with existing indicators.

Speech classification and detection will be explored further in chapter 6. Chapter 7 will integrate speech classification, the LSP-based processing and other speech modification methods into a speech enhancing CELP codec.

6 Sound Classification

6.1 Introduction

In the introduction to part II of this thesis, it was explained that an adaptive speech enhancement system was envisaged, integrated into a CELP codec, as shown in figP2.1. Adaptive enhancements are required firstly because it is known that certain of the speech intelligibility enhancement strategies to be employed can reduce speech quality, and are thus not wanted unless necessary (for example, speech modification is not required if there is no acoustic background noise in the environment of the listener), and secondly because each enhancement method may be specific to speech type and noise type (for example, shifting the frequencies of formants is not likely to improve intelligibility when listened to in white noise, or when the current type of speech contains no formants).

An adaptive enhancement strategy relies upon an analysis of both the acoustic background noise in the presence of the listener, and an analysis of the speech to be reconstructed by the CELP decoder. The acoustic background noise within the target environment is already sampled by a microphone, and analysed by the CELP encoder that transmits speech from the vehicle to the base station. Thus the output of this encoder is available within the CELP codec unit (where the enhancements are located), and can be analysed. The following analyses are performed:

① It must be determined whether the listener is speaking. If not then the microphone signal is assumed to derive from acoustic background noise and can be analysed. If the listener is speaking, a previous measure of the acoustic background noise must be used (and some measure of variance kept to determine how much the noise is changing between analysis frames).

② The CELP parameters from frames of acoustic background noise are analysed to determine the level and type of acoustic background noise.

③ The CELP parameters received by the vehicle are analysed to decide if speech is present in the received signal.

④ If speech is being transmitted to the vehicle, then this must be classified.

⑤ The acoustic background noise and the currently transmitted speech are compared by a *hearing model* to determine the degree of intelligibility of the current speech in the current noise to an average listener.

Subsequent to these analysis stages, a decision can be made as to whether speech enhancement is required, and if so, which type and degree of enhancement should be used.

6.2 Available measures

Before describing speech and noise analysis, it must be emphasised, with reference to figP2.1, that parameters for analysis derive from two locations: listeners environment and speakers environment. Both sets of parameters are limited to the standard CELP gain, pitch, linear prediction and codebook information.

Sound analysis is a well engineered topic, with many solutions, such as measurement of zero-crossing rate, power, average magnitude difference function (AMDF), cepstral techniques, autocorrelation [1][71], higher-order statistics [89][75] or others [8], often combining several techniques. Unfortunately the data available in the CELP transmission stream does not contain many of the parameters used in reported analysis schemes, and thus alternative measures must be developed.

The CELP gain parameter encodes the energy contained in each analysis frame, and is thus similar to AMDF or power measures. The pitch parameters comprise two values encoding the strength of the pitch component in the speech, and the pitch period (β and M respectively from section A2.2.4). The LSP values, as discussed in chapter 5, relate closely to speech type. Fig6.1 shows a typical example of a male speech utterance and the various CELP parameters relating to this. The LSP measure shown is that described in section 5.6.2.2, but now valued in Hertz.

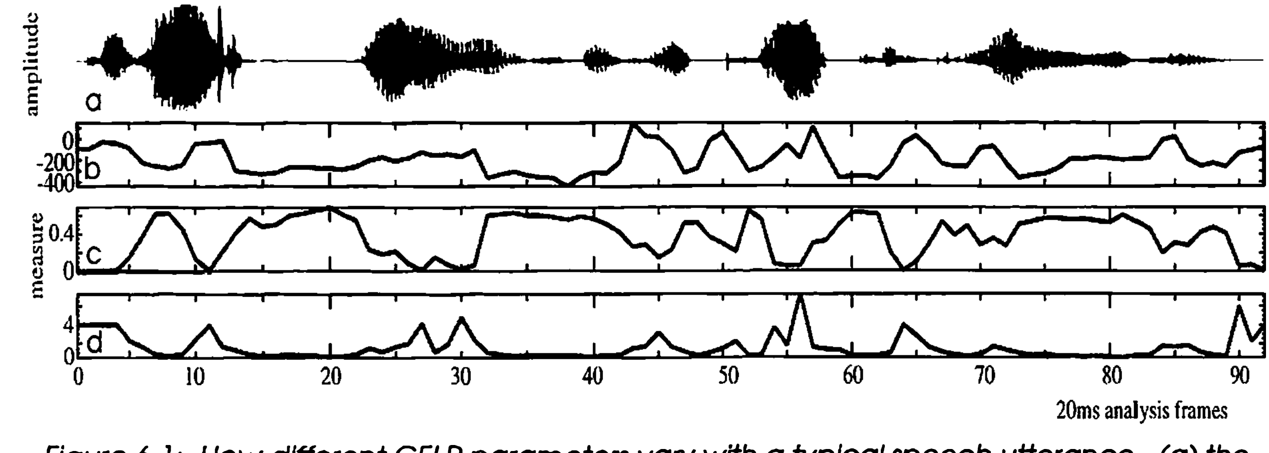


Figure 6.1: How different CELP parameters vary with a typical speech utterance. (a) the speech waveform is compared to (b) LSP measure, (c) pitch strength and (d) gain.

Examination of fig6.1 reveals that pitch is strong during certain speech periods. The gain generally follows the speech amplitude envelope (when pitch is not present) and the LSP vote measure is low during voiced speech, and high during unvoiced speech.

Each of the measures here can be confused to some extent when the speech signal contains noise with characteristics resembling speech. The solution to noise confusion is to not rely on a single measure, but to combine them, and if necessary, to average measures over a longer time period, or to perform statistical analyses on measure distribution over time. This means that only the relatively unlikely occurrence of noise having speech-like characteristics in every measured respect will be misclassified.

The development of classification and detection models is a time consuming process, involving much testing of different measures under different conditions, and using various speakers. Appendix 4 presents more detail concerning the implementation of the analysis systems - the remainder of this chapter summarises the outcome of the tests described in the appendix.

6.3 Speech detection and classification

Speech detection analysis is required for both the listeners' acoustic background noise signal and for the decoded speech, however there are different requirements for each. Considering that no analysis is perfect and that certain misclassification will result, it is necessary to define thresholds and boundaries appropriately. For acoustic background analysis, it is extremely important that no speech contaminates the noise signal. If this occurs, then the signal being analysed would not be the signal interfering with communications for the listener (as the listener would stop talking when listening, especially in reduced-intelligibility conditions). For transmitted speech detection, speech enhancement is not required during periods of non-speech, however it is more important to enhance all speech than it is to mis-enhance non-speech. Thus misclassifications should tend to exclude speech for the noise analysis path, and include non-speech for the speech analysis path.

For the decoded speech signal, classification is performed during the analysis process, with non-speech being assigned a separate class. For the noise signal, the presence of any speech must be detected. In fact the parameters available for noise analysis are identical to those available for speech analysis, and thus speech detection in both the speech path and the noise path utilise similar mechanisms.

6.4 Speech analysis

6.4.1 Speech type classification

The candidate enhancement schemes of formant adjustment and selective amplification operate on speech containing formants and unvoiced or fricative speech respectively. Thus these are two speech classes that must be detected. In addition, the presence or absence of speech must be determined.

The speech classifier operates on the LSP parameters, pitch strength and gain of the transmitted speech to classify each frame of speech as *voiced, fricative*, *other speech* or *non-speech*. The names of the classes do not exactly match their detected content, which is designed to match the enhancement strategy requirements: to classify the speech signal in terms of the type of enhancement to apply.

Thus speech frames detected in the *voiced* class are generally vowels, containing formant information, and are therefore most susceptible to enhancement through formant adjustment. Speech detected as being *fricative*, on the other hand, is better enhanced by selective amplification. Speech in the *other* speech class is indeterminate, and thus formant bandwidth adjustment or selective amplification may be appropriate, depending on the maximum amplitude of the current frame.

Speech classification is performed by comparing the available measures to fixed thresholds. The measures are normalized, and through extensive testing (section 8.3), suitable thresholds are found that correctly classify speech type for multiple speakers.

6.4.2 Formant detection

The detection of formants is required in order to firstly measure if each formant is audible (using the hearing model described later in section 6.6), and secondly to direct any formant adjustment, by the correct choice of which LSPs to move.

Although formant detection is an active research topic, with many reported techniques relying on analysis of such diverse parameters as linear prediction coefficients [2][11][64][134], cepstrums [97] and even a Newton-Raphson method [72], the parameter constraints imposed by CELP dictate the methods that can be used through the availability of only LSPs or LPC coefficients. In fact, the LSP parameters do describe spectral information, and are converted to LPC coefficients in the CELP decoder - it is relatively simple to construct a spectral representation from the linear prediction coefficients (section

A3.3).

Once a spectral representation has been obtained, it is possible to detect formants (spectral peaks) by a number of methods such as differentiating the spectrum to find maxima, peak picking, or determining maximum value within ranges corresponding to usual formant location (as seen in fig2.2 in section 2.2.1). In fact, once LSPs have been converted to LPCs, it is possible to solve the linear prediction equation numerically, with the roots corresponding to spectral peaks. However it is reported [64] that the former spectral techniques are preferable in terms of computation: both the root solving and the further processing required for the LPC polynomial solving technique are computationally more intensive.

Of course, spectral peaks may also be derived directly from the line spectral pair representation, usually being equivalent to the locations of the narrowest pairs of lines (section 5.4).

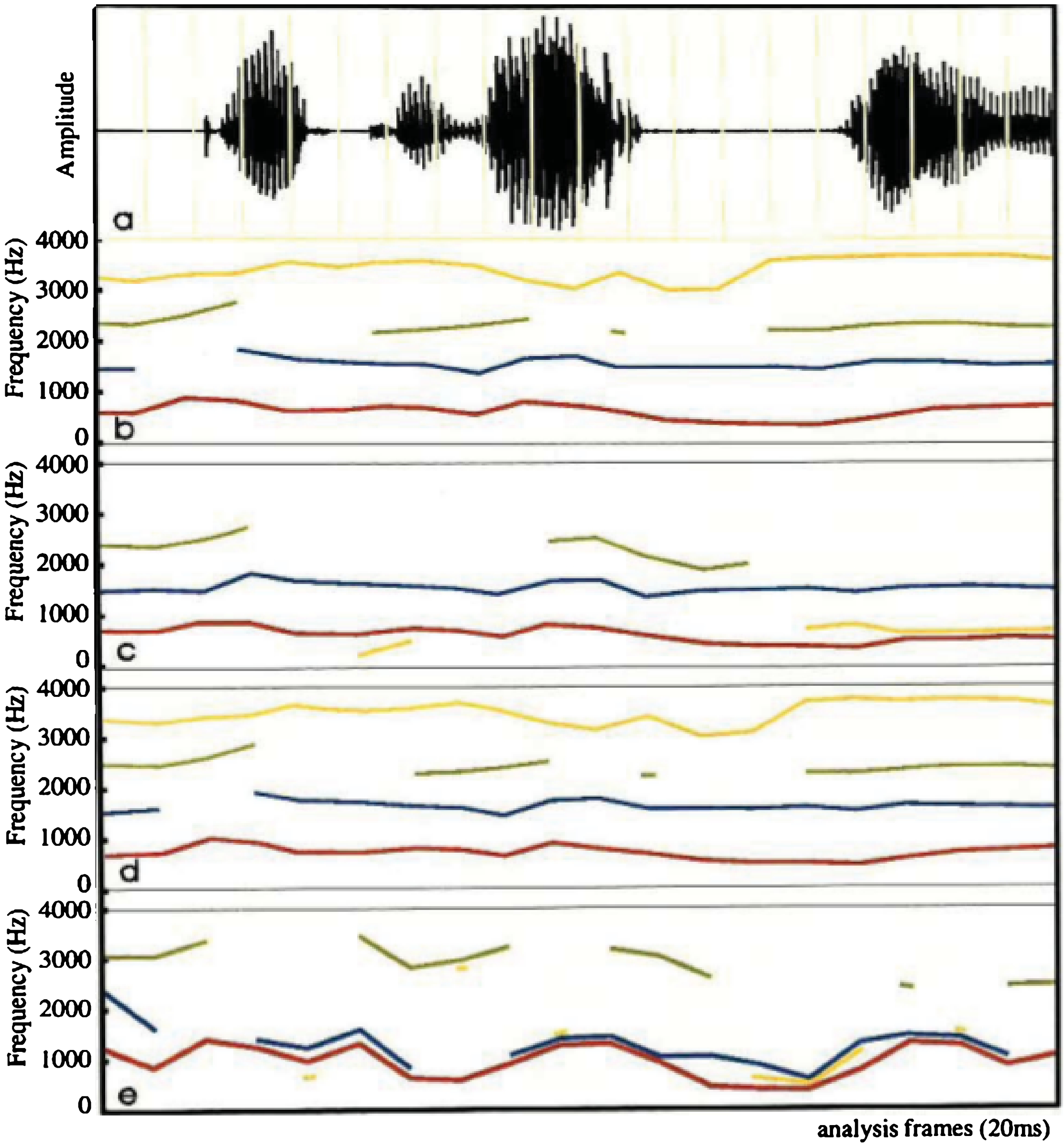


Figure 6.2: Speech waveform (a) alongside the output of different formant detection routines: (b) by differentiation of LPC spectrum, (c) by solving LPC polynomial, (d) by peak picking formants from the power spectrum and (e) by detecting most closely spaced LSPs.

Fig6.2 compares the formant locations derived from spectral differentiation, solving of the linear prediction equation, from picking peaks from the power spectrum, and detecting close pairs of LSPs. In each case the likely first three formant have been noted, and spurious peaks also plotted. No time-domain processing has been applied to any of the graphs, but a manual interpretation has been used, when 'joining the dots' between peaks detected in each frame to show the likely formant tracks for clarity. The strength and bandwidth of each detected formant has also not been shown.

It must be noted that there is no absolute measure of formant location, and that subjective analysis is often relied upon to interpret the raw spectral peak detection obtained through any particular algorithm. Details of the algorithms used to derive the graphs in fig6.2 are given in section A4.1.

It can be seen from fig6.2, and was observed through a number of tests on recordings of different speakers, that the formant position derived from a differentiation of the linear prediction spectrum, or from peak picking the spectrum is most accurate. With little to choose in terms of computation cost, the former method was chosen for use in the enhancer.

All methods of formant extraction suffer from incorrect results: formants may merge into single peaks (or appear to when the frequency resolution of the representation is too low), formants may appear and disappear at different locations, and the number of formants may vary between none and five or more (although the number of formants conveyed in the linear prediction spectrum will never exceed half of the analysis order [90]). Reported methods of formant extraction almost always apply a further time-domain constraint to the analysis process. This confines formant locations to smoothly varying positions (with the maximum frequency change gradient derived from the maximum rate at which the human throat muscles can move).

To summarise, a method of formant detection is used which derives the formant positions and amplitudes from the linear prediction power spectrum. The spectrum is differentiated to find turning points, and further differentiated to determine which of these points are maxima. The amplitudes of the maxima are found by averaging the spectral amplitudes of the three points located around the turning points (this is because a frequency resolution of 20Hz is used for the spectral representation, and the averaging reduces the effects of spectral quantization). The lowest frequency-first ordered array of spectral peaks is considered to be a relatively good estimation of the formant frequencies present in the current frame.

The formant shifting/sharpening/broadening process does not require prior information on the formant location: it simply applies to each spectral peak encoded by the line spectral pairs. However the formant locations are required for intelligibility analysis, as described in section 6.6, where the lowest frequency spectral peak (scoring in the top three in terms of amplitude) is considered F1 and the two spectral peaks of the top three, next highest in frequency, are considered to be F2 and F3. The relatively low resolution (20Hz step) of the spectral representation and the averaging process inherent in the formant location reduces the likelihood that spurious double peaks will be flagged as being formants (in contrast, note that the LSP-based formant detection algorithm with results shown in fig6.2e suffers greatly from double spectral peak problems - seen as a number of misdetected formants clustering together).

6.5 Noise analysis

As discussed in chapter 4, the different enhancement methods are effective in different types of noise. Formant frequency shifting is ineffective in white noise, but formant sharpening/ broadening is most effective in white noise.

It is impossible to predict every type of noise that will occur in the target situation, but at least generic car-interior noise, and the noise of the three police sirens will be present (section 2.4.4). The speech enhancement process must therefore be capable of reacting to any noise type or level. Despite this, there are good reasons for pre-formulating strategies to deal with common noise types. For example, if a siren is detected in the listeners acoustic background noise, and one frame of noise is contaminated by speech, then using the previous noise frame estimate for analysis will not work because the masking frequency will have moved. However, it is possible to predict the masking frequency due to a siren at any instant.

So noise analysis can be considered for two cases. Firstly, if the current noise type fits one of a predetermined codebook of known noise types, then enhancement accuracy is improved. Secondly, when the current noise can not be found in the codebook, then the enhancement strategy must be calculated for this general case.

The design of the noise siren codebook classifier is given in section A4.2.

6.6 The hearing model, and intelligibility

The hearing model is designed to compare the background noise and the speech in a perceptual fashion to determine if that speech would be intelligible to an average listener in that acoustic background noise. In addition, the hearing model may be called upon to predict if modified speech can improve intelligibility (and if so, what degree of modification is required).

In order to effect a perceptual comparison, the hearing model must account for psychoacoustic effects in its analysis. Psychoacoustics is the difference between a purely physical audio measure and the experience of a listener, and is well researched for a number of consumer products such as A-law (or μ-law) compression in telephone systems [47], the Philips DCC and Sony MiniDisc systems, both of which rely upon equal loudness and masking models to compress high quality audio [44].

6.6.1 Psychoacoustic effects

Although many processes may be defined as psychoacoustic, only the following relevant subset are discussed:

- Masking (section 2.3.2.1)
- Equal-loudness (section A1.3)
- Frequency discrimination (section 4.4)
- Speech perception

Masking as a psychoacoustic effect may be modelled in two ways. Firstly, each tone in a complex sound may be analysed and a masking effect determined. The masking effects for each tone are then summed to yield an overall masking level. Unfortunately there is evidence that masking effects are not entirely additive [73], however, some empirical methods have been developed to account for the non-linearity. A second method is to consider the ear as containing a number of critical-band filters (defined in section 2.3.2.1), and the masking level as a weighted sum of all noise powers that reside within each band. Implicit in such a model is a method of accounting for the effects of equal loudness.

The perception of speech may loosely be classed as psychoacoustic: whether or not a given period of speech is intelligible depends upon various factors. These include the audibility of the sounds that make up the speech (and this will vary with time), and the relative importance to the speech communication of those sounds. Other perceptual factors include the predictability of the speech, the redundancy and the familiarity of the listener to that speech.

6.6.2 Reported hearing models

It is usual for most authors to begin to model psychoacoustic effects by considering the frequency resolution of the ear. The most popular measure of frequency is the Bark scale (see section A1.4), which is a frequency scale with the property that unit Bark increases are of equal perceptual relevance. The audio signal must be represented in this Bark scale. This may be accomplished by using a filter bank, with each filter bandwidth being a constant Bark, and the absolute position of each filter related to the critical band frequencies [124].

More usually, the audio signal is represented in the frequency domain as a spectrum (usually derived via FFT analysis), and this signal is warped into the Bark scale [25][101][102].

The Bark-scaled spectrum is then convolved with a spreading function [124] (a precalculated masking effect), or applied to an equal-loudness pre-emphasis filter and converted to perceptual loudness units [42].

Most auditory models of masking work in similar ways, the following method is derived from Hermansky, 1990 [42], and is compared with the methods of other authors.

6.6.2.1 Spectral analysis

A frame of speech is weighted by a Hamming window and discrete Fourier transformed to yield a frequency domain representation linear in Hertz, and converted to a power spectral representation.

6.6.2.2 Critical band warping

The power spectrum $P(\omega)$ must be warped in frequency to fit a Bark scale. If the Bark frequency is Ω and the (Hertz-linear) angular frequency ω, then [42] for a sample rate of 4kHz:

$$\Omega(\omega) = 6 \cdot log\{\omega / 1200\pi + [(w / 1200\pi)^2 + 1]^{0.5}\} \tag{6.1}$$

now that the spectral index is represented in Barks, the effect of the critical band filter must be calculated using a relationship such as [42];

$$\Psi(\Omega) = \begin{cases} 0 & for\,\Omega < -1.3 \\ 10^{2.5(\Omega + 0.5)} & for\; - 1.3 \leqslant \Omega \leqslant -0.5 \\ 1 & for\; - 0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega - 0.5)} & for\,0.5 \leqslant \Omega \leqslant 2.5 \\ 0 & for\,\Omega \geqslant 2.5 \end{cases} \tag{6.2}$$

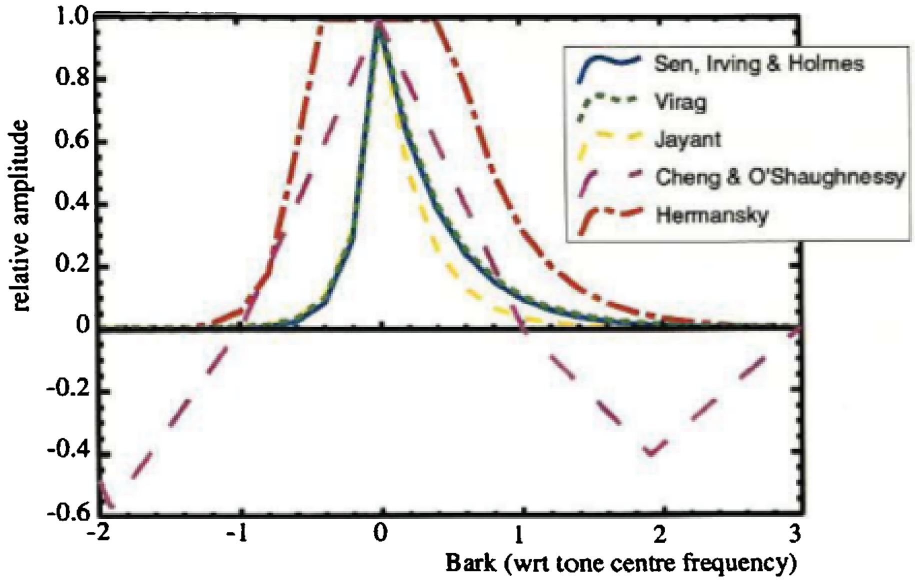


Figure 6.3: Comparison of critical-band spreading functions from various authors, Sen et. al. (101), Virag (124), Jayant et. al.(47), Cheng et. al. (18)(19) and Hermansky (42).

As the critical-band curve (also known as spreading function, lateral inhibition function and noise masking curve) is a psychophysical phenomena, it must be determined empirically. Most authors derive an approximation for this function such as that given in eqn6.2. This approximation is compared in fig6.3 with the functions used by various other authors. The curve of Cheng & O'Shaughnessy [18] additionally introduces some attempt to account for the lateral inhibition phenomena [69].

Jayant et. al.[47], Virag [124] and Sen et. al. [101] use curves that appear similar, differing here only in the upper-frequency side of the curve (the part of the curve that is most centre-frequency specific). In fact the latter of these authors corrects the upper-frequency side of their curve depending upon both absolute centre frequency and absolute power, to give a much better representation of a realistic function. Without such correction, accurate modelling of the critical-band function is not possible, and thus the curves of Jayant et. al.[47] and Virag [124] are clearly inferior models. Note that the curves shown in fig6.3, where modelled as absolute power and absolute frequency dependant functions, have been evaluated only for a fixed centre-frequency of 1kHz and a fixed power of 70dB$_{SPL}$.

The flat-top curve of Hermansky [42] not only approximates the human critical-band function well, but accounts for the dependence on absolute amplitude and centre frequency. In addition the flat top function minimises the very real problems associated with the use of such functions with discrete frequency arrays (where the curve peak will not always coincide with a sample frequency). This is in effect a reduction in sensitivity to frequency quantization errors.

6.6.2.3 Critical band function convolution

The chosen critical-band function (eqn6.2) must be convolved with the warped spectrum (eqn6.1) to generate a critical-band power spectrum:

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{6.3}$$

In general, a spectrum of around 200 samples will now have been convolved into a coarser Bark domain representation (of around 20 to 40 constant Bark-width sample bins).

6.6.2.4 Equal-loudness preemphasis

Many attempts have been made to quantify the equal-loudness function of the ear. As the function is both frequency and amplitude specific, and is usually derived for the case of a single tone, certain assumptions must be made for its general use. Most authors simply base their preemphasis around the 40dB curve (figA1.2 in appendix 1) such as the following used by Hermansky [42]:

$$E(\omega) = \frac{\omega^4(\omega^2 + 56.8 \times 10^6)}{(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)} \tag{6.4}$$

Note that this approximation to the equal-loudness curve is close up to around 5kHz, but above this should be extended with a further term:

$$E(\omega) = \frac{\omega^4(\omega^2 + 56.8 \times 10^6)}{(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)(\omega^6 + 9.58 \times 10^{26})} \tag{6.5}$$

The function in eqn6.5 is plotted in fig6.4.

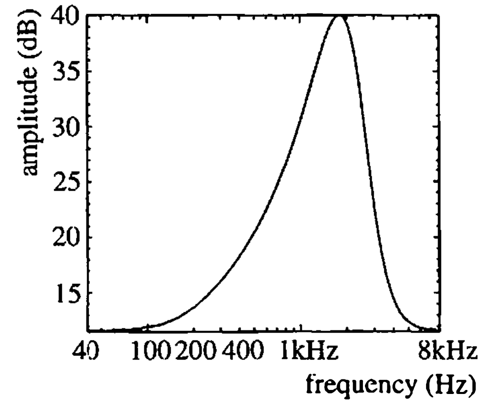


Figure 6.4: Equal-loudness preemphasis function of eqn6.5.

6.6.2.5 Intensity-loudness conversion

Finally, the equal-loudness function of eqn6.5 is combined with the critical-band power spectrum and a numerical conversion made to relate the power units to perceived loudness rather than power - this is known as the power law of hearing [42][73]. The perceptually relevant spectrum is now found to be:

$$\rho(\Omega) = \{E(\omega)\Theta[\Omega(\omega)]\}^{0.33} \qquad (6.6)$$

For a flat input spectrum, fig6.5 shows a graph of the convolved critical-band filters for each element of a frequency array with each frequency bin having constant bark width.

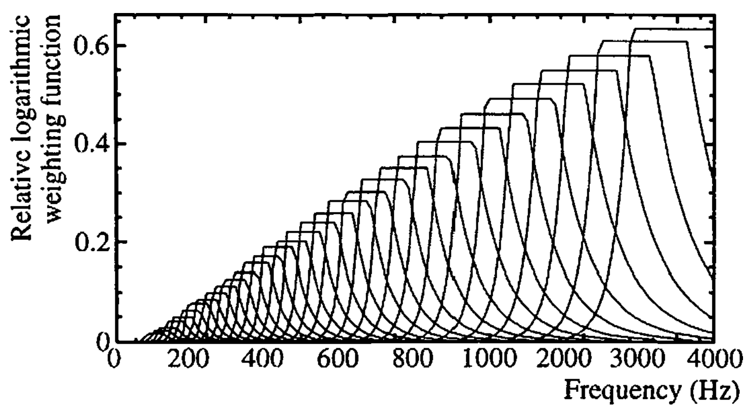


Figure 6.5: 40 equal Bark width convolved critical band filters of eqn6.6.

For comparing two perceptually weighted spectra, the power law of hearing is unnecessary if the absolute difference between the two spectra is not required. The simplification has been accepted in practice for the speech enhancement intelligibility estimation subsystem.

6.6.3 Hearing model outcome

The psychoacoustic processes modelled in section 6.6 can be used to analyse a given frame of sound in order to determine the overall masking effect of this sound. That is, a tone of given frequency would need to be louder than the calculated masking value at this frequency in order to be heard.

In fact, the output of most methods is a Bark domain array - with each bin holding the masking value for the critical filter centred at that frequency. It can be assumed that a harmonic whose components do not exceed the level of masker in any of the bins within which they lie would be inaudible, whilst a harmonic whose components all exceed the masking levels would be audible. If certain of the tones within the harmonic exceed the masking level, then there is a degree of audibility, but the sound heard may not be that of the full harmonic sound.

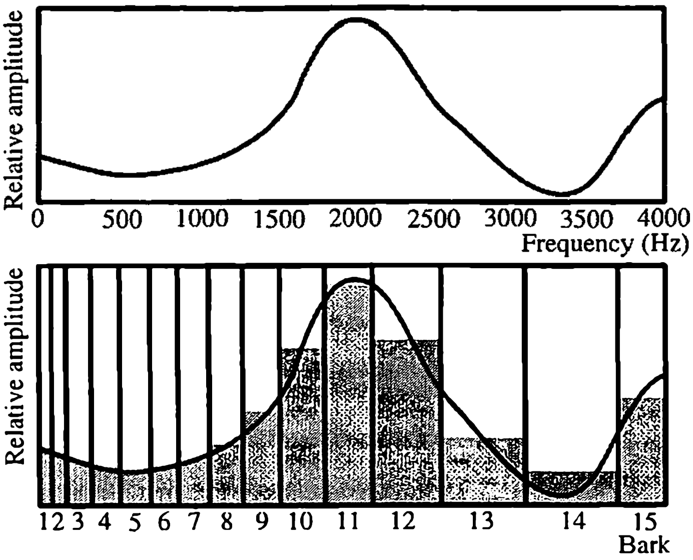


Figure 6.6: Masking level from spectrum for 16 constant Bark width subbands. Top shows frequency spectrum, with subband masking levels added below.

Fig6.6 illustrates the output from a 16-band psychoacoustic analysis of a sound, the spectrum of which is shown. Any tone must rise above the masking level in the subband within which it is located in order to be audible. For speech, formants rising above the masking level will be audible, formants not rising above the masking level will be inaudible.

Section 2.2.1 discussed the relative importance of speech formants, with F1 being found to be relatively less important to intelligibility than F2 and F3. In order to construct a measure of intelligibility, the audibility of each formant must be considered. For speech not containing formants, the audibility of the entire speech spectrum (approximately 250Hz to 2kHz) should be considered. Audibility is defined here as the amplitude difference between the (perceptually weighted) sound to be heard and the calculated noise masking level in the current critical band (Bark index). Negative values of audibility indicate that the sound will not be heard by an average listener in the given noise, but the value of audibility has no other meaning.

Subjective testing has determined that a good approximate measure of intelligibility for voiced speech is to construct a weighted average of the audibility of the first three formants, with the weighting acting to reduce the contribution of F1 by 50%.

In order to calculate the effects on intelligibility of shifting formants, it is necessary to first calculate the audibility of the formants at different frequency positions on the masking curve. Intelligibility is then found from the weighted average.

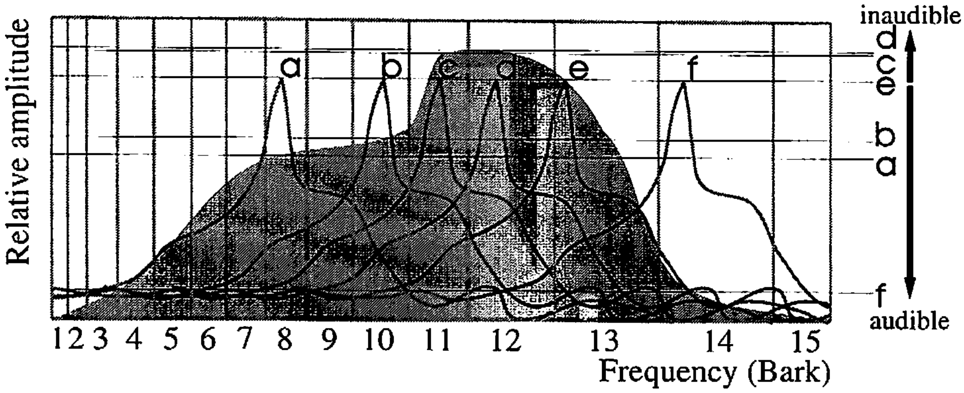


Figure 6.7: Audibility of frequency shifted spectral peaks (a to f) with respect to acoustic background masking noise.

Fig6.7 illustrates the method of determining the audibility of shifted spectral peaks (or formants). Peaks a to f have been shifted higher or lower in frequency, and in each position the peak signal to masking noise level has been found (on the right hand side). For this example, instances a,b and f are audible, and c to e are all inaudible. Thus a spectral peak as shown would need to be shifted to position f to be most audible. For formants, or multiple spectral peaks, the weighted average of the peak signal to masking noise levels would be considered instead.

7 Designing a speech-enhancing CELP coder

7.1 Introduction

This chapter describes the speech enhancing CELP coder. The enhancements are known: some of them are performed by adjusting LSPs, and others by adjusting the CELP gain parameter, whilst the type of enhancement is chosen based upon the current type of speech and noise.

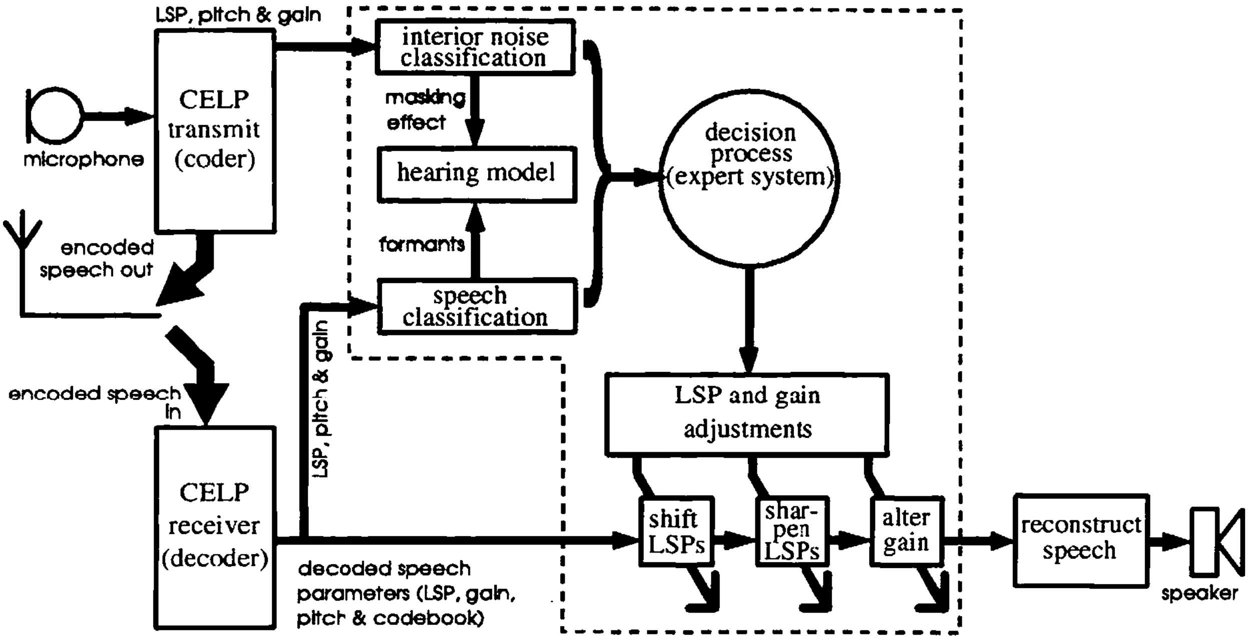


Figure 7.1: A block diagram of the enhancing CELP system. Original CELP functions are outside the dotted enclosure; enhancing additions are inside. Shown is a CELP decoder and part of a CELP encoder, sufficient for acoustic background noise analysis.

Fig7.1 contains a block diagram of the speech enhancing CELP coder, within which can be seen the three chosen types of enhancement algorithm, being selected and adjusted by an 'expert system' on the basis of an analysis of the speech currently being decoded and an analysis of the acoustic background noise in the environment of the listener. The parameters for the latter analysis being derived from a CELP uplink normally used to transmit speech from the listener to the speaker.

This chapter describes the operation of the speech enhancing CELP system in more detail - summarising the findings of the previous chapters on enhancement using LSPs, speech classification, noise analysis and others.

7.2 Data flow analysis

Speech is encoded by a CELP coder and transmitted to a CELP decoder located in the target environment. The received parameters (LSP, pitch, gain and codebook values) are then passed to the speech enhancement subsystem before being decoded by the CELP decoder into speech.

The speech enhancement subsystem analyses the CELP parameters to firstly determine the class of speech being received (non-speech, voiced, fricative or other) and then determines what formant frequencies, if any, are present in that speech. A perceptually-weighted speech spectrum is constructed from the LSP parameters (via spectral generation from LPC coefficients as described in section A3.3).

Concurrently with the speech analysis, the enhancement subsystem is also analysing the CELP parameters from a CELP encoder located within the target environment. Firstly the parameters are checked to ensure that no speech is present within the target environment, and secondly, a noise spectrum is derived from these parameters. This spectrum is then perceptually weighted and applied to a hearing model to determine its masking effect. If speech is contaminating the noise signal then the previously uncontaminated perceptual analysis is re-used.

The hearing model and expert system together select the type and degree of speech adjustment (if any) to be made to the CELP LSP and gain parameters. CELP parameters, having been adjusted or not, are passed from the enhancement subsystem to the speech reconstruction stage of the CELP decoder, which synthesises the speech to finally broadcast to the listener.

7.3 Hearing model and expert system

The perceptually weighted speech and noise spectra are compared in the regions of the formant frequencies to determine the audibility of each formant, and from this an intelligibility measure is constructed (section 6.6.3). If this intelligibility measure is found to be negative (ie. unintelligible) then the noise masking spectra is examined in a number of regions around each formant, extending as wide as the allowable formant shift, to determine if formant shifting will improve intelligibility. If, however, the speech was found to be intelligible then no further processing is required and the unaltered CELP parameters are passed on to the speech reconstruction stage of the CELP decoder.

If the speech is fricative, the intelligibility is still measured by comparing the perceptually weighted noise masking and speech spectra. If the speech is unintelligible then selective amplification is used to improve intelligibility. A check is made of absolute amplitude levels of speech and noise to ensure that selective amplification does not actually reduce intelligibility (discussed in section 2.3.1). If this looks likely then selective amplification is not used, and a fixed degree of formant narrowing is used. If speech in the *other* class is detected, or if *voiced* speech intelligibility can not be improved through formant shifting, then formant broadening is used instead.

7.4 Enhancement

Under the selective amplification scheme, the CELP gain parameter is scaled by an increasingly larger gain multiplier for each consecutive frame of unintelligible fricative speech, until a maximum gain multiplier value is reached. For non-fricative speech frames, the gain multiplier value is reduced towards unity. For consecutive non-fricative speech frames, the reduction is twice as steep as the increase was. The gain multiplier sloped attack and decay curves were chosen empirically through informal listening tests, which also demonstrated an obvious enhancement of fricative speech periods, and is shown in fig7.2.

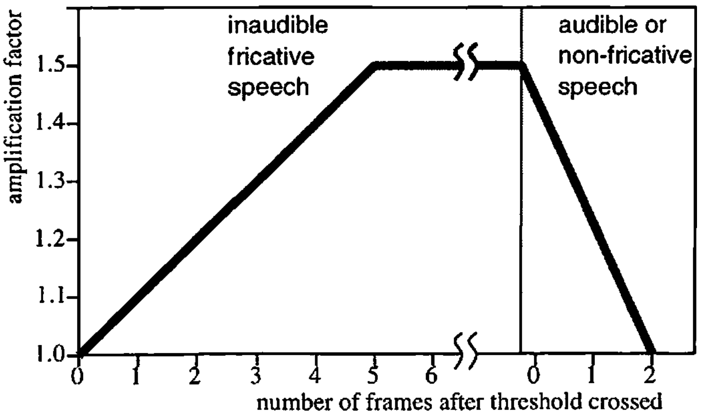


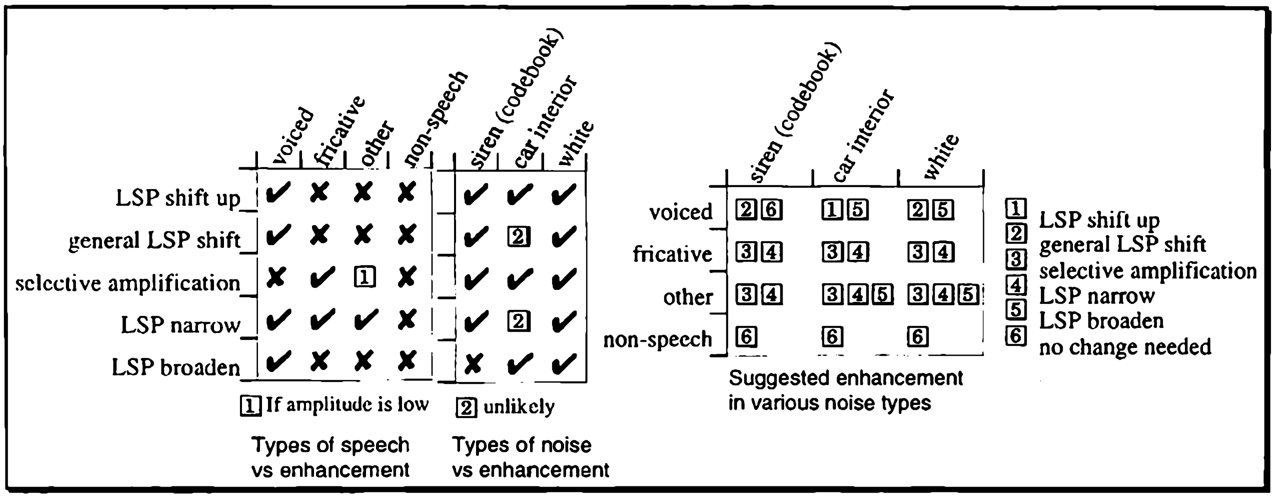
Figure 7.2: Selective amplification speech enhancement scheme amplification factor.

LSP adjustments have been explained in chapter 5. Maximum LSP shifting is ±20%. This figure has been established through informal listening tests and has been demonstrated as a value that does in fact result in enhancement (in section 9.3). Formant broadening to counteract wideband noise is fixed at 180% (established through testing in section 9.5.2, and also shown to provide enhancement in section 9.3). and formant narrowing of fricative speech that cannot be amplified is fixed at 75% (established through informal testing, impractical to test extensively but tested implicitly in final system tests of section 9.6).

7.5 Expert system

The expert system within the enhancement system is designed to select the most appropriate enhancement, and degree of enhancement for each decoded frame of speech. To do this, a number of rules have been established governing the type of enhancement that can be applied in any particular type of noise to any particular class of speech.

Fig7.3 shows the decision matrices implemented within the expert system. Firstly, the type of speech is determined, ruling out certain enhancement types. For example, only the *voiced* speech class has distinct formants, and thus formant frequency shifting can only be applied to voiced speech. However minor spectral peaks are also sometimes observed in fricative speech, and often in the *other* speech class. These may be sharpened to subjectively improve intelligibility.



	voiced	fricative	other	non-speech	siren (codebook)	car interior	white
LSP shift up	✓	✗	✗	✗	✓	✓	✓
general LSP shift	✓	✗	✗	✗	✓	2	✓
selective amplification	✗	✓	1	✗	✓	✓	✓
LSP narrow	✓	✓	✓	✗	✓	2	✓
LSP broaden	✓	✗	✗	✗	✗	✓	✓

1 If amplitude is low

2 unlikely

Types of speech vs enhancement

Types of noise vs enhancement

	siren (codebook)	car interior	white
voiced	2 6	1 5	2 5
fricative	3 4	3 4	3 4
other	3 4	3 4 5	3 4 5
non-speech	6	6	6

Suggested enhancement in various noise types

1	LSP shift up
2	general LSP shift
3	selective amplification
4	LSP narrow
5	LSP broaden
6	no change needed

Figure 7.3: Rules for enhancement type selection based on speech and noise types.

Selective amplification is designed to normalize speech amplitudes between louder and quieter phonemes, and thus does not operate on the *voiced* speech class containing mostly vowel sounds. However some speech in the *other* class may be amplified if it is of low amplitude. *Non-speech* is not processed.

Secondly, the expert system uses the type of acoustic background noise to restrict enhancements: a general case, or noise detected within a noise codebook (siren noise). Siren noise is very much a tonal noise (section A4.2) and thus only those speech frequencies close to the varying siren tone will be masked, making this an ideal situation for formant shifting. Fig7.4 illustrates how example formants will be shifted in the presence of siren noise. Note that the formant frequencies are shifted away from the changing siren frequency - this means both upward and downward shifts.

Fig7.3 also indicates the response to car interior noise, a specially shaped noise type used in listening tests. LSP shifting in such noise is generally upward due to the predominance of low frequencies in the spectrum (shown in fig2.10).

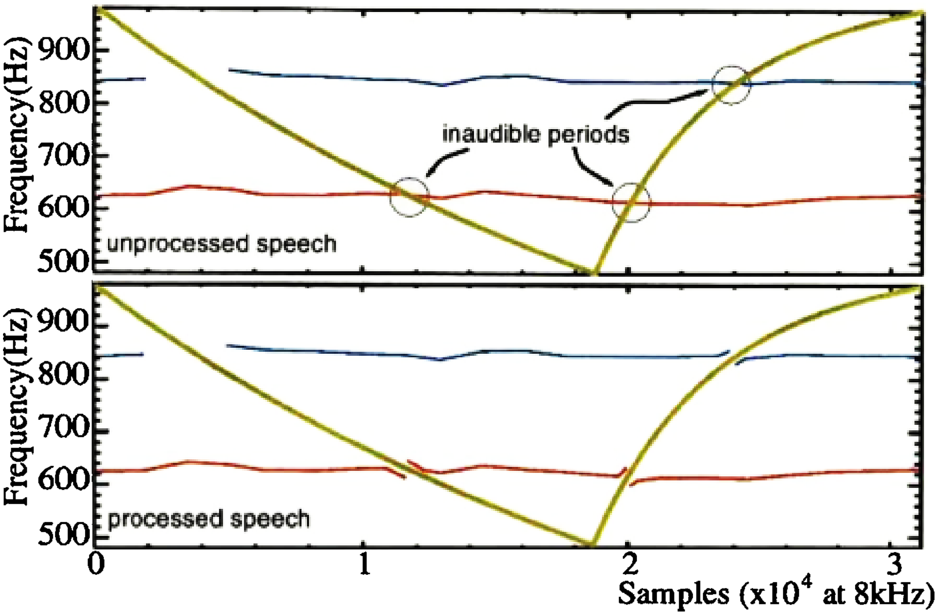


Figure 7.4: Formant shifting being applied to two formant tracks (roughly horizontal lines) to overcome periods of inaudibility due to interfering siren noise (one period shown, sweeping through its frequency profile).

Finally, the enhancements are performed. If multiple enhancements are possible on the current frame, the allowed enhancements are selected in the following order:

LSP shift → selective amplification → LSP narrow/broaden

Thus for *voiced* speech, if LSP shifting can not increase intelligibility, LSP narrowing/ broadening will be selected instead. For *fricative* speech, selective amplification is preferable, however if the amplitude is too loud this will be disallowed, and LSP narrowing/broadening will be selected.

PART III: Testing and Evaluation

Parts I and II of this thesis have presented speech enhancement methods, related these to hearing, speech, the CELP coder and the types of noise likely to be found in the target system, before proposing a subset of the methods for further integration with the CELP coder.

Methods were reported of integrating enhancements into the CELP coder, including by the use of a novel LSP-based speech modification technique. The requirements of the enhancement methods, and the systems needed to control them automatically were investigated and means of conducting the relevant analyses and deriving the required parameters were found.

Finally, a speech enhancing CELP structure was proposed, and its operation discussed.

Part III begins in chapter 8 by describing some of the experimental work conducted in order to develop the speech enhancements and integrate these with the CELP structure. Appendix 6 contains implementational details of and more results from the tests in chapters 8 and 9.

Chapter 9 will discuss the methods used to test the speech enhancements, and to test the final speech enhancing CELP system - in contrast to chapter 8, here aspects up to the level of the entire system are investigated using objective and listener trial methods. Test results are presented and analysed before chapter 10 summarises the entire speech enhancing CELP system, its capabilities and limitations.

Chapter 11 then concludes by restating the aims of the speech enhancements, and how the system developed and described in this thesis fulfils those aims. Consideration is also given to novel methods developed here to perform speech enhancement and the possibilities of extending those methods, and the enhancement system in general to other applications.

8 Test objectives and systems

8.1 Introduction

Much experimental testing was required in the investigation of the speech enhancing CELP codec. This chapter presents the CELP simulation upon which the methods were tested, and some of the experiments that have been conducted with it.

Evidence is also presented of the effectiveness of the speech classification, speech detection and noise classification tests and methods.

In order to test the speech enhancement functions, samples of realistic speech and noise were required. The speech type and noise simulations are described.

8.2 CELP simulator

The CELP codec, described in chapter 3, compresses speech into LSP, gain, pitch and excitation codebook values. Of these, the LSP, gain and pitch parameters are required for speech detection and analysis, and the LSP and gain parameters are adjusted by the enhancement methods.

It is well known that the most computationally intensive part of the CELP encoder is the codebook search loop [98][78], however for the purpose of speech enhancement testing, this is not necessary: the codebook index is not used. In fact, CELP can be simplified considerably to investigate such speech enhancements.

Fig8.1 shows the simplified CELP structure used for enhancement simulation purposes. The LPC analysis process determines LPC coefficients and removes the LPC contribution from the speech signal, which is then subject to LTP analysis: determining the pitch parameters and removing these from the speech signal. The speech signal gain is calculated from the speech signal (although usually performed as part of the LPC analysis process), although the residual after LPC and LTP analysis is not normalized and encoded as a codebook index as in standard CELP.

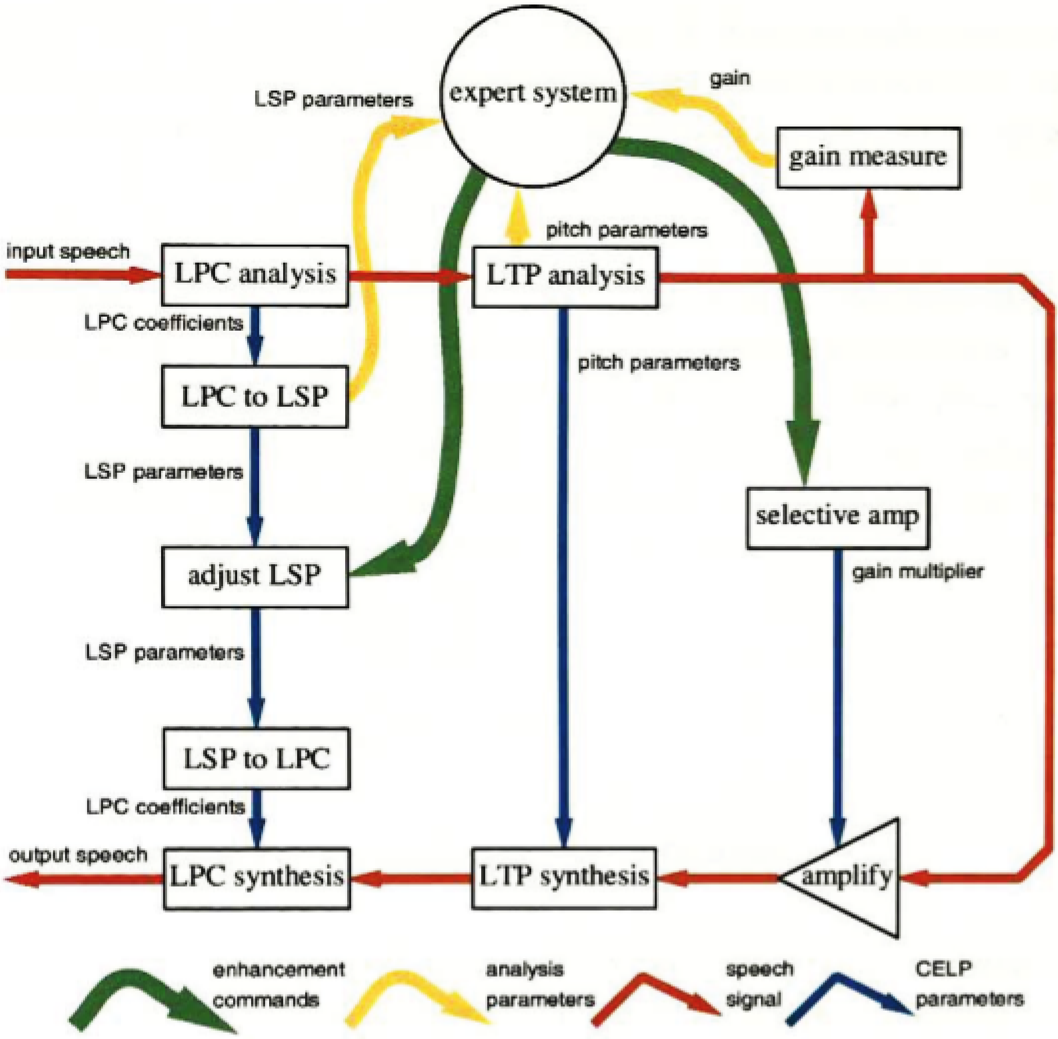


Figure 8.1: Simplified CELP structure used for enhancement simulations.

In the structure used, the residual signal is amplified by the speech enhancement gain multiplier, used for selective amplification, and then the pitch parameters added in. Finally, the signal is fully reconstructed with the LPC contribution added in. LSP adjustments act to alter this LPC contribution.

Standard CELP is a ‘lossy’ coder: all parameters are quantized, especially the residual being quantized as one of a number of codebook vectors. This CELP simulation however is an almost lossless coder. If no adjustments are made to LSP values, and the gain multiplier is unity then the output speech will be identical to the input speech, excepting floating point or fixed point rounding effects.

Simplification of the CELP simulation not only saves a significant amount of otherwise wasted processing time (for example simplification increases the speed of a MATLAB simulation by a factor of around 100), but removes some of the effects of CELP coding such as those due to parameter quantization, to allow the enhancements to be individually characterized: and to allow enhancement effect to be related directly to cause without consideration of any interfering processes.

This CELP simulator, developed under MATLAB, has been used for many tests, including most of the subjective listening tests, and the development of the enhancement strategies. The CELP simulation itself is investigated more thoroughly in [68] However, the final system testing, as described in chapter 9, uses a commercial CELP coder written in 'C' and modified with enhancement additions. This coder includes all of the effects of CELP coding, such as parameter quantization, and the use of the codebook (vector quantization).

8.3 Speech classification

8.3.1 LSP measure in relation to speech features

Section 6.2 explained that speech classification and detection must be accomplished using the available CELP parameters, while section 6.4 proposed a method of speech classification using the CELP pitch strength, LSP vote and power measures.

Tests have been conducted to further substantiate the assumption (section 5.6.2) that noticeable changes in LSP value when representing different classes of speech can be quantified into a speech measure. Tests used the LSP vote measure of section 5.6.2.4 compiled within the CELP simulation of section 8.2.

A number of sentences, totalling over 20 minutes of speech, were randomly selected from the TIMIT multi-speaker speech database [118], which contains not only a large selection of speech sentences spoken by various North American male and female speakers, but also a time-indexed phonetic transcription of the speech. The phonetic transcription, made manually by a panel of speech experts, notes the start and end positions of each speech feature within the recordings.

A simulation was constructed that calculated the LSP vote measure on fixed-sized speech frames in order to correlate these against the phoneme within which they are found. The

LSP measure value obtained from each frame was added to the total measure obtained for the current phoneme, and the phoneme occurrence counter incremented. When a phoneme boundary occurred within a frame, the measure was assigned proportionately to the phoneme on each side of the boundary, and the count of each phoneme occurrence incremented fractionally.

Fig8.2 gives the raw results in terms of average LSP vote measure with respect to type of phoneme for a selection of 5896 individual phonemes of 57 types. The phoneme names are those of the TIMIT database, with certain additional symbols. *h#* (padding frames before and after speech onset) and *pau* (pause between phonemes or words) indicating non-speech.

The results are difficult to interpret, but it can be seen that certain of the fricative phonemes (*jh, ch, f, z, sh, s*) show high average measure values, and that strongly voiced phonemes (*m, n, ao, aa, oy ,w*) show low average measure values.

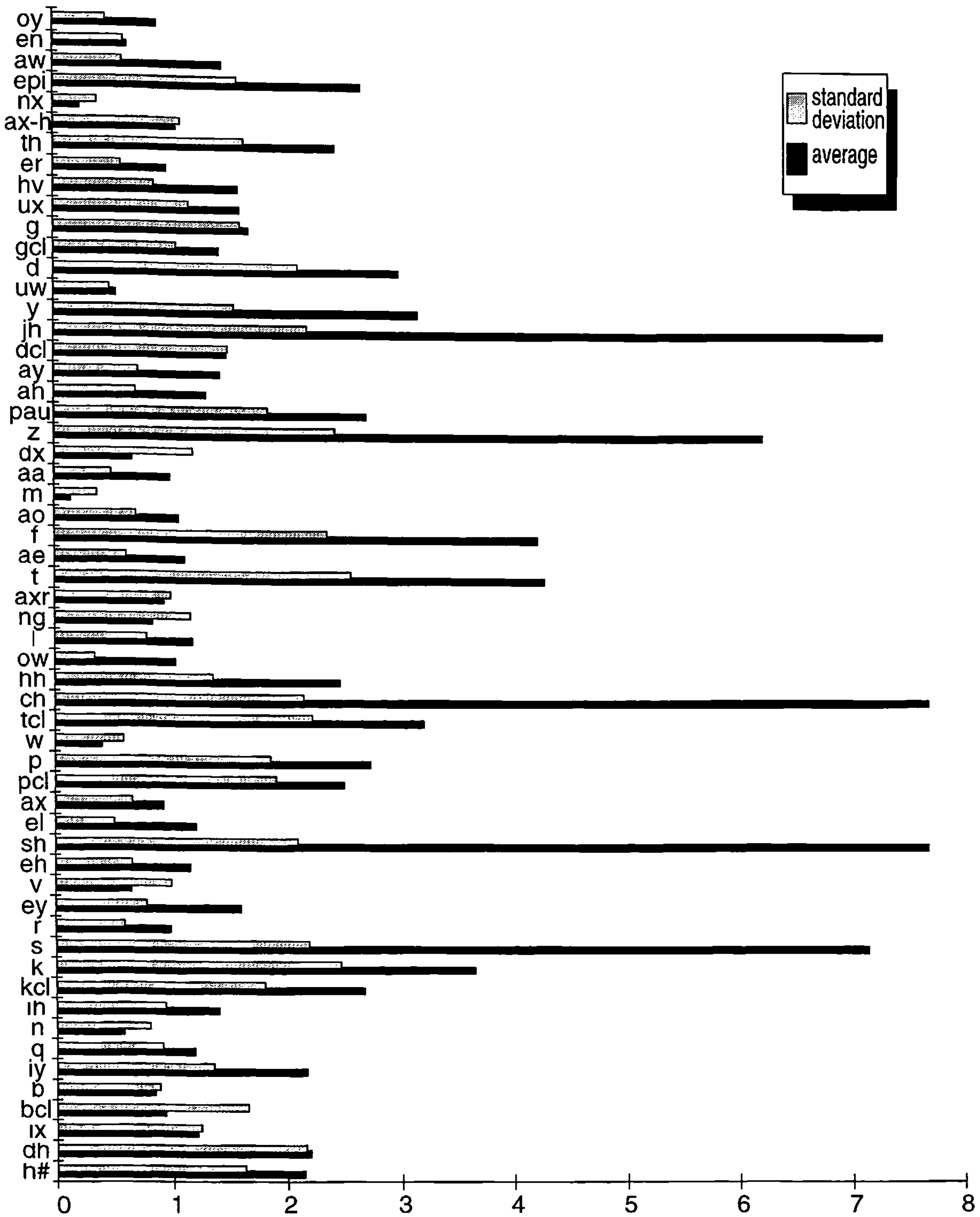


Figure 8.2: LSP vote measure average and standard deviation shown against phoneme type.

To aid the interpretation of results, in fig8.3, the phonemes have been grouped into five classes by type, and the average phoneme measure per group plotted with error bars indicating the maximum and minimum average phoneme measure found for any phoneme within that group.

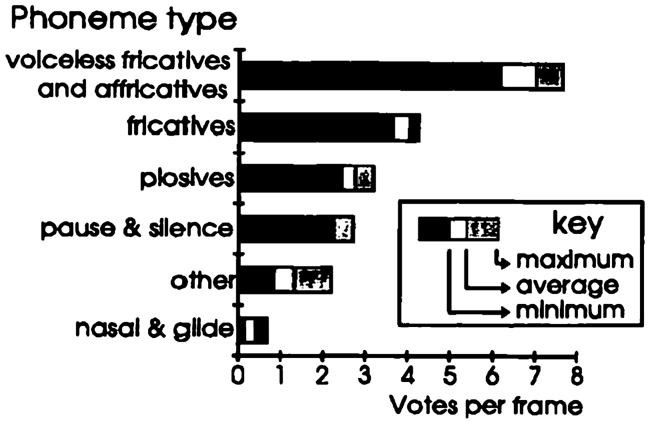


Figure 8.3: Average, maximum and minimum LSP vote measures per phoneme class.

Fig8.3 indicates that those phonemes that would most benefit from speech enhancement by selective amplification can be distinguished by the LSP vote measure. Fricatives and affricatives, in particular, can be detected for possible further selective amplification.

Of course, no single speech measure alone would be used for classification. Improvement would be noted when other measures (pitch strength and gain) were considered. In particular other measures must be relied upon to detect non-speech and voiced speech (contained here within the 'other' category) which the LSP vote measure distinguishes poorly.

8.3.2 LSP measure compared to other measures

Section 8.3.1 indicated that an LSP measure alone could be used to distinguish between certain classes of speech. In this section, an LSP measure is similarly applied to a number of sentences taken from the TIMIT database [118], however it is here compared to three other standard speech classification measures. In addition, the test has been conducted for the speech sentences mixed with each of five levels of interfering white noise.

Hereafter, the four tested speech measures are referred to by their abbreviations:

- LSP: the measure of section 5.6.2.3, equal to summing the differences between each LSP value in Hertz, but also subtracting the sum of the nominal LSP values in Hertz for a flat spectrum. (Equivalent to the measure of section 5.6.2.2 with no square prior to summation, and normalized by dividing the result by the system order).

- ZCR: zero-crossing rate is the number of times the sample value within a speech frame crosses the zero-axis, divided by the number of samples within the speech frame.
- POW: the frame power measure, equivalent to summing the square of each sample value within the frame, divided by the number of samples within the frame.
- AMDF: absolute magnitude difference function, of the sum of the absolute of each sample value divided by the number of samples within the frame (often used instead of POW when implemented on systems without efficient multiply instructions).

With the exception of the novel LSP-based method, these are all standard metrics by which speech can be classified [1][132][131]. Note that the tested speech samples were in 16-bit format, and speech was normalized to a maximum amplitude of 40% of full-scale. Variable amounts of noise were added to this as explained in section A6.5.

For the test, each speech sentence was analysed in a number of 240 sample speech frames and the four measures calculated for each frame. Use of the TIMIT phonetic transcription files enabled measures to be assigned to current phoneme type as in the experiment of section 8.3.1. Changes made between the previous and current experiments were an increase in the number of analysed sentences, and a rejection of measure values to be assigned to phonemes with less than 25% overlap with the current frame. There were 4169 logged measure occurrences for 58 phoneme types.

Each of six levels of noise were added to the sentences prior to analysis. As the amplitude of each speech recording was normalized, the maximum speech amplitude to maximum noise amplitude ratios tested can be given as 9dB, 6dB, 3dB, 0dB and -3dB.

For each of the noise conditions, Spearman rank correlation was used to determine the similarity between the different measures, and between the same measures when the speech was noise-contaminated and when it was not. The Spearman correlation relies upon the placings of each phoneme in terms of its rank, rather than the measure value itself, thus working in a similar way to speech classification which deals with the separability of classes of phonemes in the n-dimensional space defined by the n speech classification measures.

The results indicate that under noise-free conditions, the LSP and ZCR measures correlate well (the full interpreted experimental results are presented in table A6.7, section A6.5). In a CELP decoder environment, there is no possibility of a speech enhancement system having access to the original speech signal in order to calculate ZCR, and thus the significant correlation shows that the LSP measure can be used as an alternative for speech classification. Reassuringly, the AMDF correlates very well with the frame power measure. There is little strong correlation elsewhere, however both the ZCR and LSP measures are weakly inversely proportional to the power and AMDF measures, with the LSP measure slightly less weakly related than the ZCR measure.

As noise level increases, all correlations except the power-AMDF relationship break down with saturation. Evidence suggests that the LSP measure deteriorates less than the ZCR measure as noise increases (table A6.8 and figA6.1 in section A6.5). The frame power measure also appears more robust in noise than the AMDF measure, as has been reported [1].

Speech classification, involving the determination of phoneme type through parameter analysis of speech, can be described as defining a region in *n*-dimensional space within which occurrences of the desired phoneme reside, and where *n* is the number of feature measures used. Speech classification is explored further in section A4.3.

Fig8.4 shows the two dimensional zero-crossing rate versus frame power measure space, for noise-free speech, with the average locations of each of 58 phonemes noted (with the type being as described in TIMIT documentation [118]). With ZCR and power parameters being common speech classification methods, as discussed previously, it can be seen that a suggested classification region can be drawn to encompass phonemes judged most likely to benefit from selective amplification.

The meaning of the TIMIT phonemes can often be guessed from their symbol, perhaps with the exception of *q* which indicates the presence of 't' in "bat" and *dh* as the 'th' in "then". *pau* denotes a pause between words, *epi* is an epenthetic silence (such as the gap between the 'm' and 'b' in "thimble", and often found between a fricative and a semivowel or nasal). *h#* is the TIMIT marker prior to the beginning, and following the end of a word, denoting non-speech or silence.

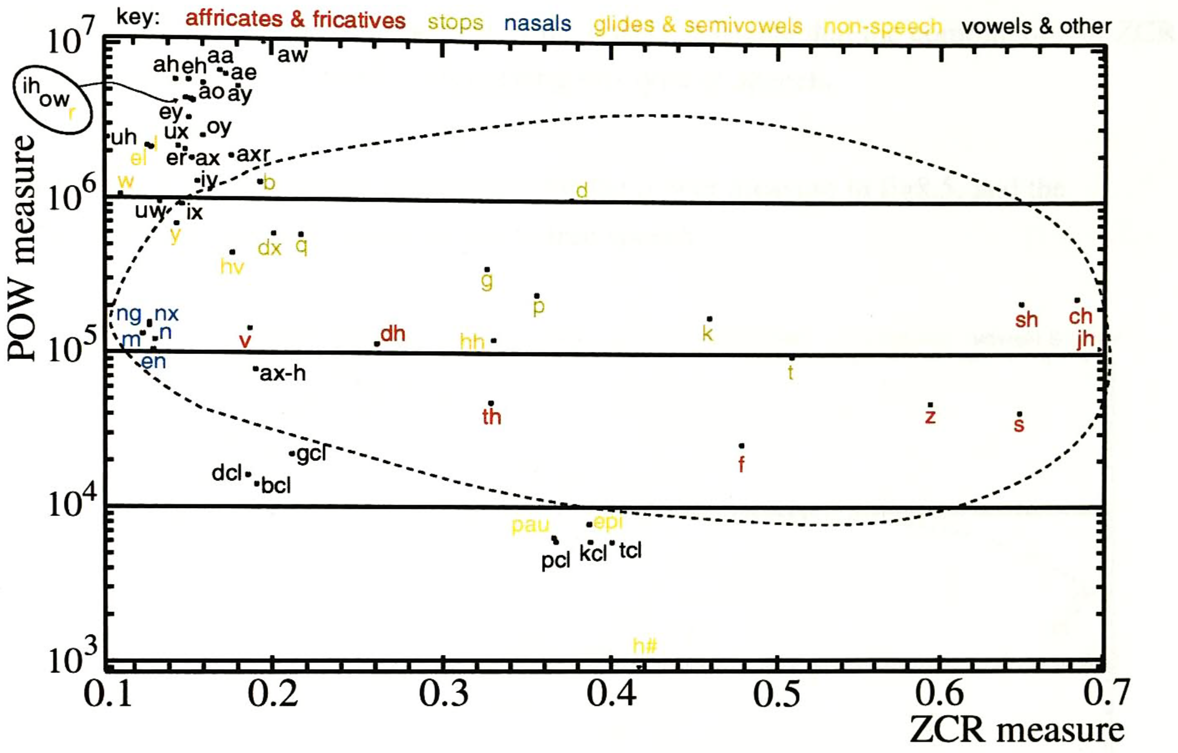


Figure 8.4: ZCR measure against frame power measure, showing phoneme classifications and approximate detection region bounded by a dotted line.

The suggested classification region encompasses all fricatives, stops and nasals, some glides and semivowels and two vowels. *ax-h* is called a devoiced schwa, a very short unvoiced vowel typically bounded by voiceless consonants such as the ‘u’ in “suspect”. *ix* is the ‘i’ sound in “debit”, an unvoiced breathy sound. The glides *hh*, *hv* and *y* are the ‘h’ in “hay”, the ’h’ in “ahead” and the ‘y’ in “yacht”. Each of these sounds is unvoiced and therefore likely to benefit from enhancement through selective amplification.

Further sounds, classed as ‘vowels & other’ are the consonant closures. The TIMIT documentation describes the closure intervals of stops *b, d, g, p, t, k* as being distinguished from the stop release wherever possible. The closures are thus *bcl, dcl, gcl, pcl, tcl, kcl* with the closures of *jh* and *ch* also being *dcl* and *tcl.* Closures occur when the path of air, and sound out of the mouth is momentarily stopped by throat, tongue or lips blocking the vocal tract. These may thus be classified as non-speech, although they, and some of the marked non-speech sounds are an integral part of speech communication. For the purposes of enhancement, however, each of these sounds can be categorized as features that would not benefit from the available enhancement methods.

Note from fig8.4 that the tight bunching of the nasals indicates that the combination of ZCR and POW measure is very good at classifying this type of speech.

In contrast, the LSP measure is plotted against the power measure in fig8.5, and the phoneme positions marked, again for noise-free speech.

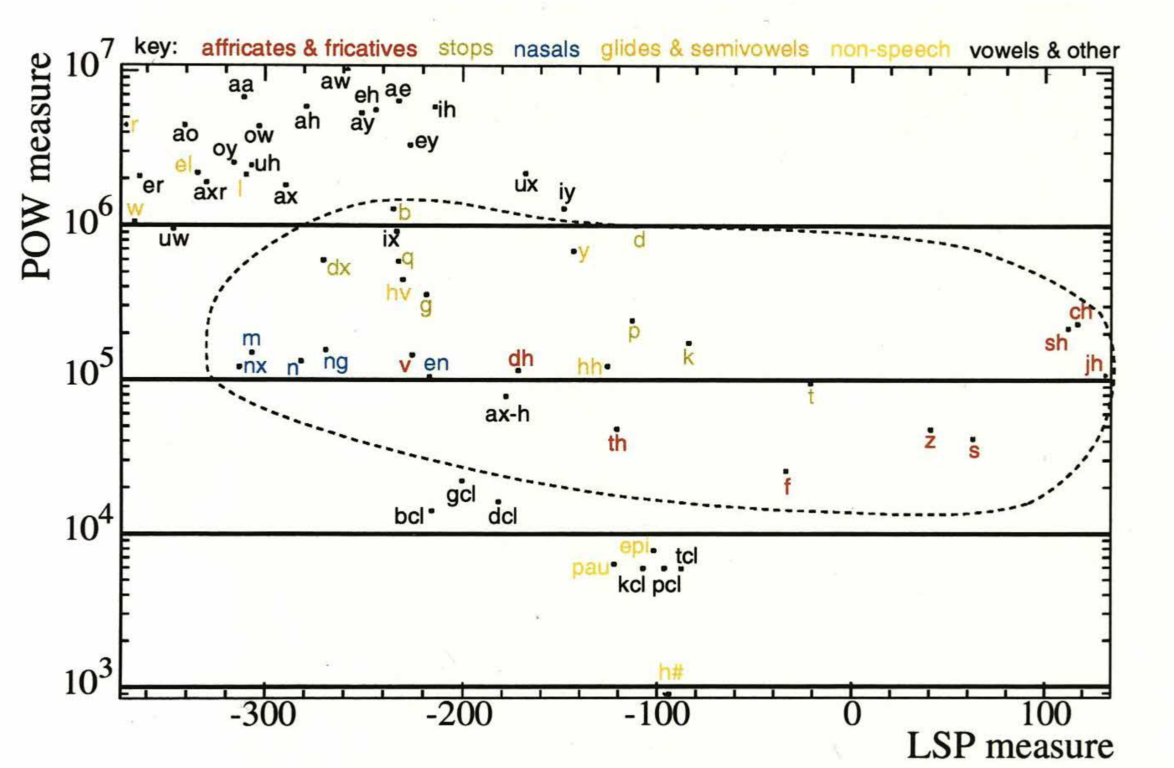


Figure 8.5: LSP measure against frame power measure, showing phoneme classifications and approximate detection region within the dotted boundary.

The classification boundary of fig8.5 encloses the same phoneme types as those in fig8.4, however it can be seen that the separation of the classification region and the bunch of vowels at the top left of the plot has improved over the ZCR case. This is evidence that the LSP measure is better able to distinguish phonemes in the suggested region than the ZCR measure, when used in conjunction with the POW measure.

When speech is corrupted by noise, the effect is to reduce the distinction between classification regions through saturation of measured values. Fig8.6 shows phoneme positions plotted against ZCR and frame power measures for speech contaminated with noise such that the speech to noise amplitude ratio is 9dB. Due to the saturation of analysis frames by additive noise, there is much less variation in both the frame power and ZCR measures (previously ranging from 10^3 to 10^7 and from 0.1 to 0.7 respectively).

Despite the evident bunching together of phonemes, it is still possible to apply a decision region and separate the wanted phonemes in this level of noise. In this case, the closure locations and non-speech periods also fall within the classification region. The latter must therefore be distinguished in another way.

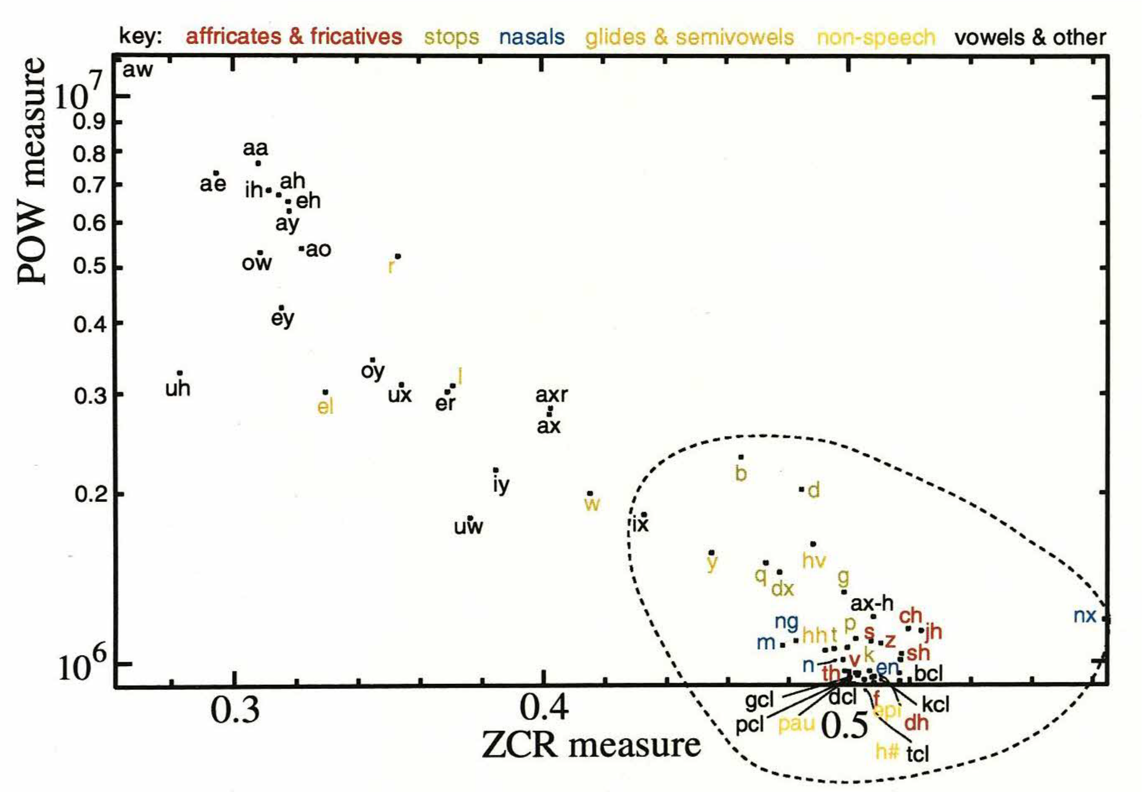


Figure 8.6: ZCR measure against frame power measure for 9dB SNR, showing phoneme classifications and approximate detection region within the dotted boundary.

Similarly, when the LSP measure is plotted against frame power for a 9dB speech-to-noise ratio condition, as in fig8.7, the measure scales are significantly reduced (LSP measures previously ranged from -400 to 140) and phoneme bunching occurs. However, as in the case of the ZCR measure, a classification region can still be drawn: even in such noise conditions, both LSP and ZCR measures, when combined with a frame power measure, are capable of applying the suggested speech classification region.

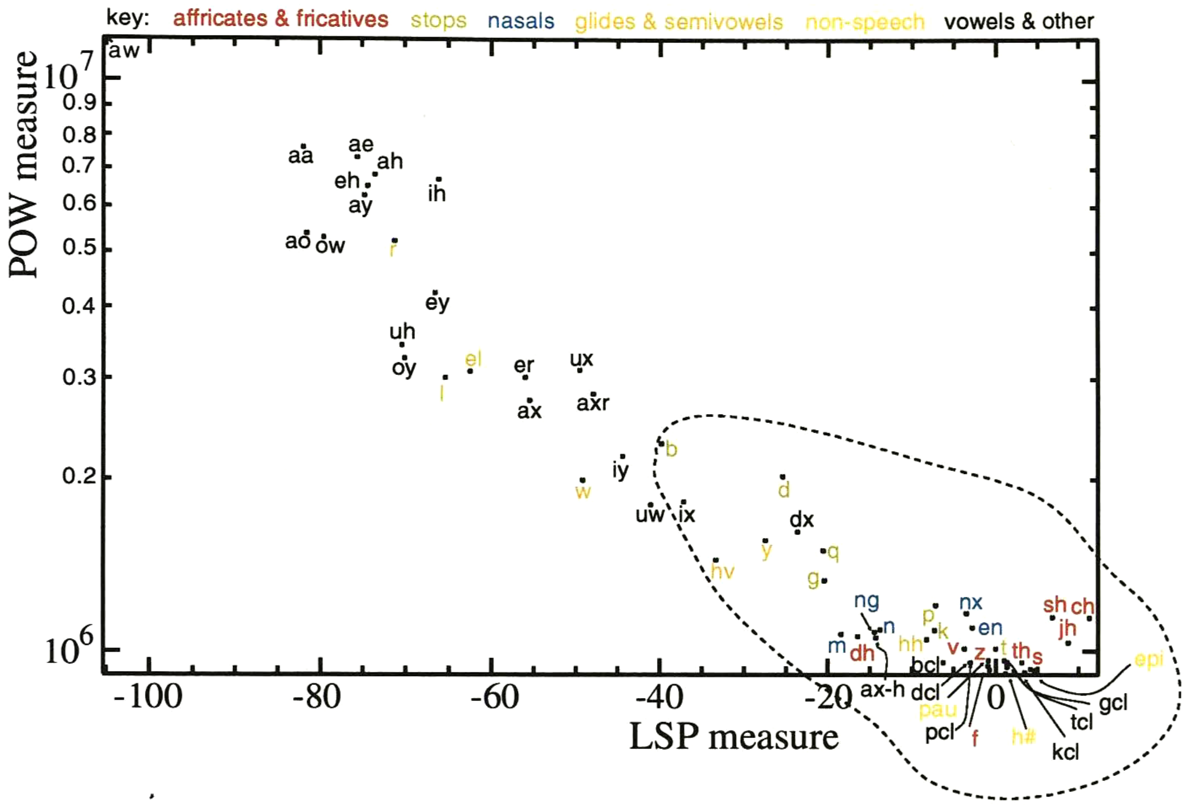


Figure 8.7: LSP measure against frame power measure for 9dB SNR, showing phoneme classifications and approximate detection region within the dotted boundary.

When the amplitude of noise added to the speech prior to analysis becomes extreme, all of the tested classification methods become unusable. This is shown in fig8.8, which plots the LSP measure against the frame power measure for a speech-to-noise ratio of -3dB.

The impossibility of separating phoneme types is illustrated by the presence of the vowel sounds *uh, uw* and *iy* within the main body of consonant phonemes. Note that due to the saturation of the frame power measure, this is no longer plotted logarithmically. The saturated LSP measure now extends from approximately -12 to 15, a range that extended from -400 to 140 under noise-free conditions.

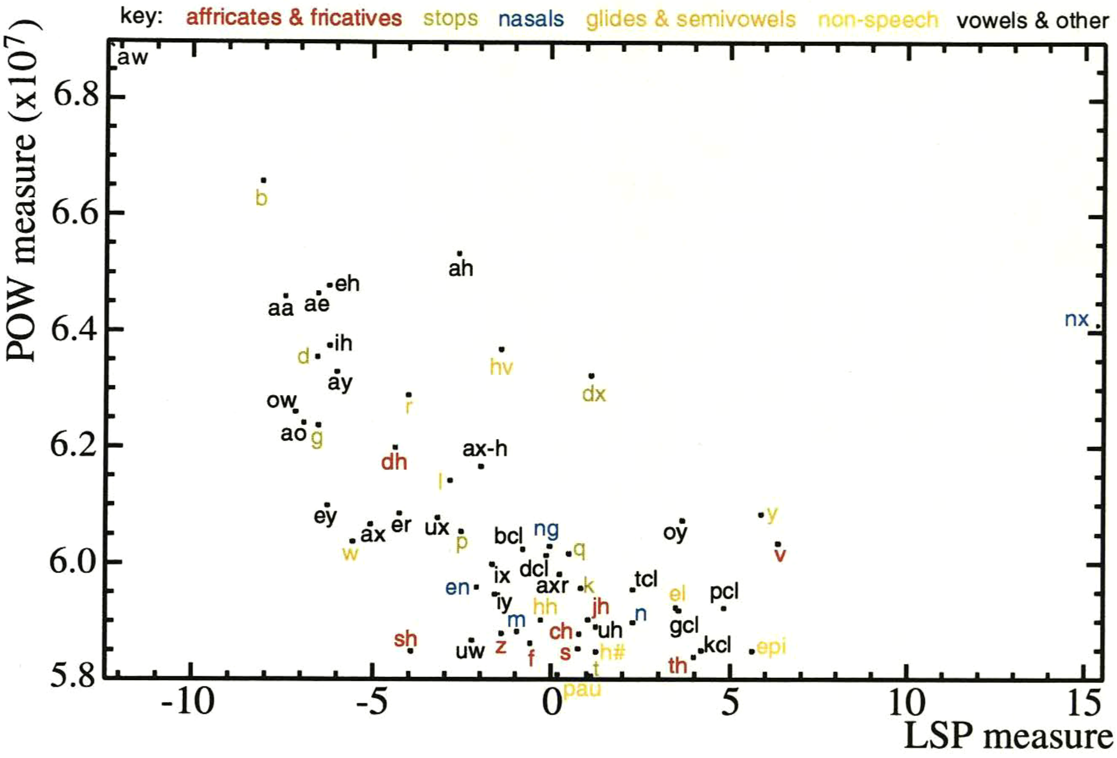


Figure 8.8: LSP measure against frame power measure, showing phoneme positions coloured according to classification, for speech mixed with equal amplitude white noise.

Despite the reduction in efficiency of the measures with added white noise, the experiment demonstrated that the LSP measure is more robust to noise than the well-known zero-crossing rate measure (this is also demonstrated numerically in table A6.8 in section A6.5). The strong correlation between LSP and ZCR measure in noise-free and 9dB SNR conditions indicates that both measures describe approximately equivalent speech features.

It should perhaps be noted here that zero crossing rate measurement, although a relatively computationally efficient measure, still requires *m* comparisons and up to *½m* additions (= *1½m* operations) to be performed within a speech frame of length *m*. By contrast, the LSP measure only requires *p+1* additions for a *p*-th order analysis system. Typical values (and those used in the TETRA codec [120]) are *m*=160 and *p*=10. Thus the LSP measure is over 20 times more efficient.

8.3.3 Interpretation of phoneme test results

Section 8.3.1 demonstrated that an LSP-based measure could be used to classify speech phonemes into different classes (particularly those classes of phonemes naturally suited to enhancement by selective amplification). Section 8.3.2 indicated the similarity between the LSP-based and ZCR measures, implying that the LSP measure may then be suitable to replace the ZCR measure in certain situations, such as speech classification. In addition, the LSP measure is more efficient to perform, more robust to noise, and better suited to classification for speech enhancement due to its ability to classify the wanted phonemes into a region that is more distinct than the region drawn for the ZCR measure.

The actual implementation of the speech classifier relies upon frame power, LSP measure and pitch strength features, and is described further in section A4.3, however fig8.9 demonstrates the classification output for a test speech waveform.

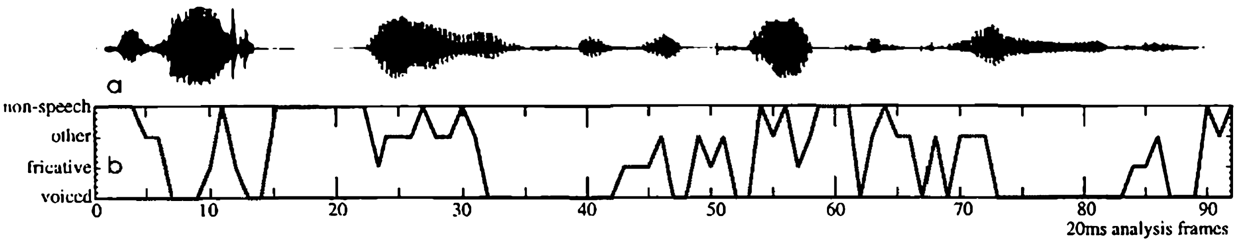


Figure 8.9: Speech classification result (b) for a test waveform (a).

A number of test plots similar to that of fig8.9 were made for different speakers to test the classification scheme, and to fine-tune the thresholds and decision regions.

Despite the signal-to-noise levels tested in section 8.3.2 ranging from 9dB to -3dB, it must be emphasised that in the target enhancement system, the speech analysis is assumed to operate on speech from a quiet police base station. Even SNR levels of 9dB are considerably more noisy than is likely: these tests were designed to characterise the LSP classification performance by comparison to the ZCR measure, and not to indicate probable noise levels.

8.4 Speech intelligibility

The speech intelligibility measure, based upon the hearing model described in section 6.6, identifies the audibility of each formant (or higher amplitude spectral regions when formants are not present), weights these, and calculates an intelligibility measure.

The operation of the hearing model can be tested by comparing a given sound to a given acoustic background noise to calculate the audibility, then mixing these and presenting them to a listener.

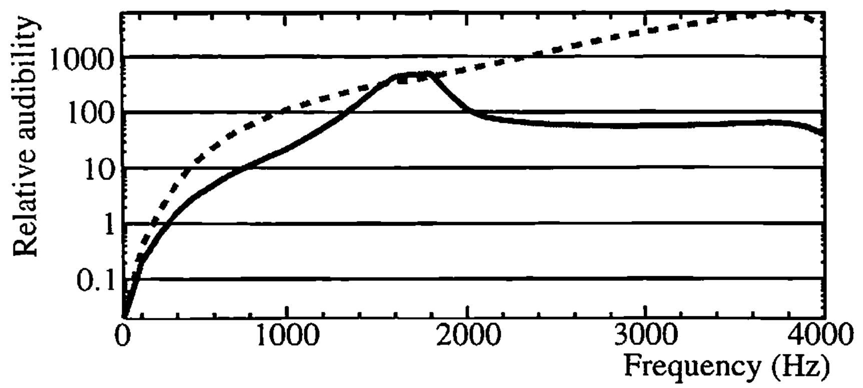


Figure 8.10: Hearing model output comparing white noise (dotted curve) and a single 1.6kHz tone (solid curve).

Fig8.10 shows a typical output from the hearing model, comparing the masking effect of white noise and the audibility of a 1.6kHz sinusoidal tone. The level of tone and noise chosen are those that are found to be just audible by a normal-hearing listener. Such informal tests have demonstrated that the hearing model can in general predict audibility and non-audibility for tonal sounds. When the tone is just-audible to some speakers, the model can be incorrect, and thus a safety margin is used when determining whether speech enhancement is required. A formant that is only audible by a small amount (up to 50 units above the noise on the currently used audibility scale) will be considered to be inaudible for enhancement purposes.

8.5 Noise simulation

A simulation of the noise likely to be found within the vehicle is an important aspect of the testing of the enhancements, and to this end, the vehicle noise has been characterised and simulated. Simulations have been made of a reasonably realistic vehicle interior noise, a simplified noise with similar shape (designed to be easily reproducible by other experimenters to allow direct comparisons of enhancement methods), the three sirens currently used by UK police forces, and a simplified tonal sound (again designed to be easily reproducible).

8.5.1 Realistic noise simulation

A simulation program has been written that generates a model of car interior noise based around previously published average vehicle spectrum analyses. In particular, the data presented in section 2.4 was used as a prototype for a filter that, when operating on Gaussian random noise, produces a similar spectrum to that found in the interior of a car. Listening tests confirm that noise from this process resembles that heard in a car interior under steady-state conditions. Fig8.11 shows a spectral plot of the car interior simulator, which can be compared to fig2.10 - the actual average of car interior spectra:

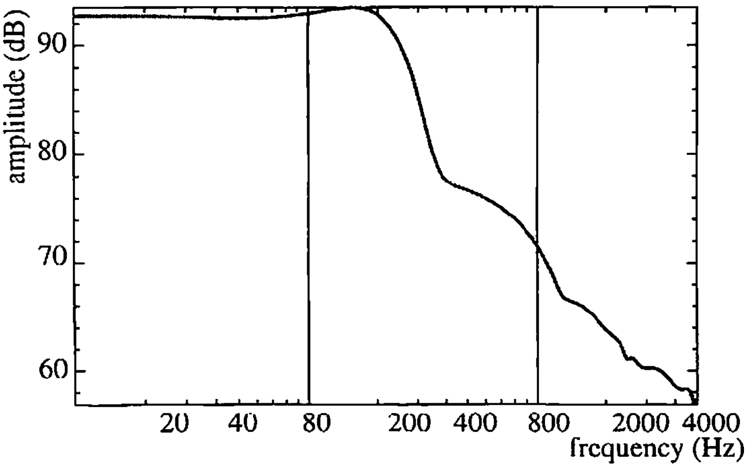


Figure 8.11: Frequency spectrum obtained from vehicle interior noise simulator.

A digital tape recorder was used to record various samples of noise generated from a 4-cylinder petrol-engined vehicle. Recordings were processed to remove the spectral weighting introduced by the microphone frequency response. A filter was derived for this purpose, based upon the inverse frequency response of the microphone (as given in the microphone specification document). This filter response is shown in fig8.12:

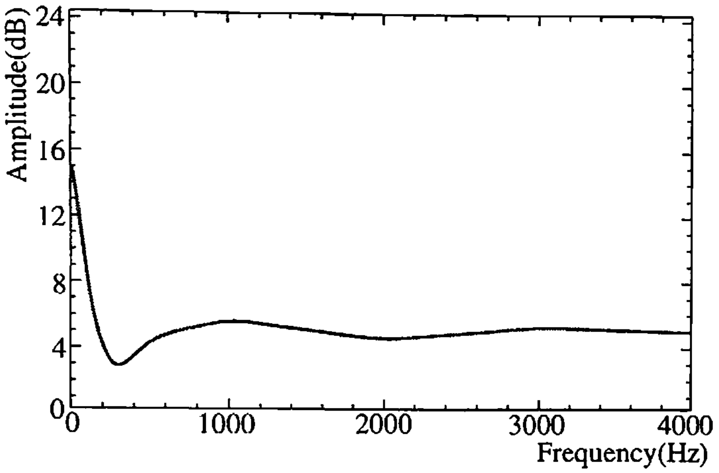


Figure 8.12: Spectral correction for microphone frequency response.

The recording of engine noise, shown in fig8.13, has been analysed in order to determine the shape of the noise waveform produced by the engine.

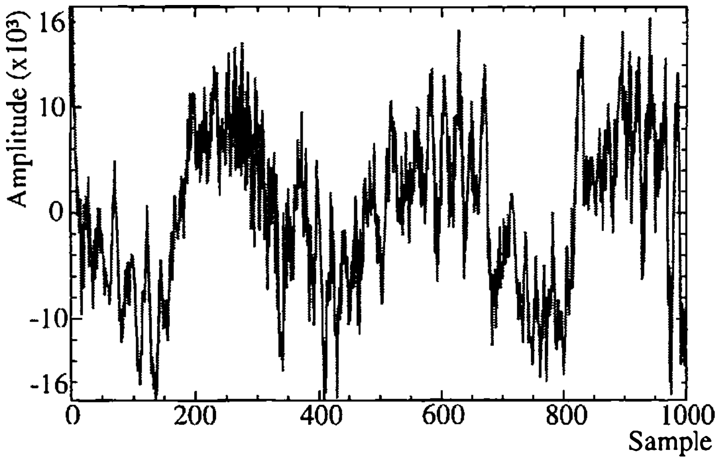


Figure 8.13: Waveform from engine compartment test recording.

The analysis began by determining the fundamental period of the noise, from examination of the auto-correlation function of the recording, as shown in fig8.14:

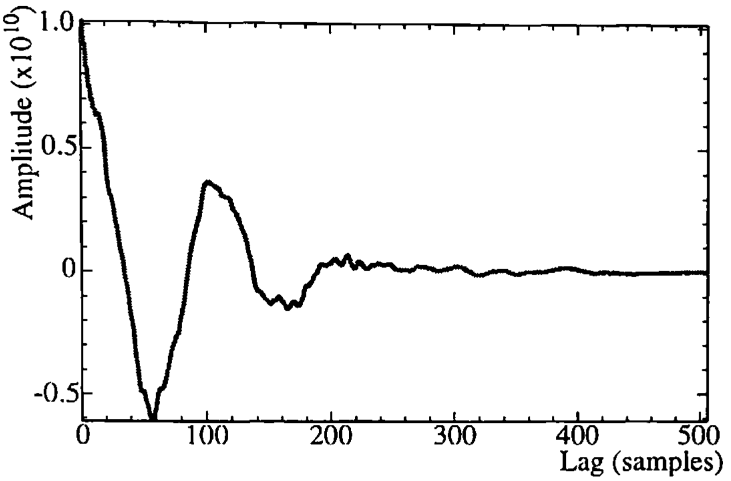


Figure 8.14: Correlogram of engine compartment recording.

The auto-correlation results indicate that periodicity predominantly exists within the engine recording at a lag of 57 samples, corresponding to a frequency of 70Hz respectively. The repetitive waveform occurring at this frequency was analysed further by adding together all samples within the recording at multiples of this lag to accentuate the periodic features and reduce the effects of uncorrelated features. The resulting enhanced waveform is shown in fig8.15:

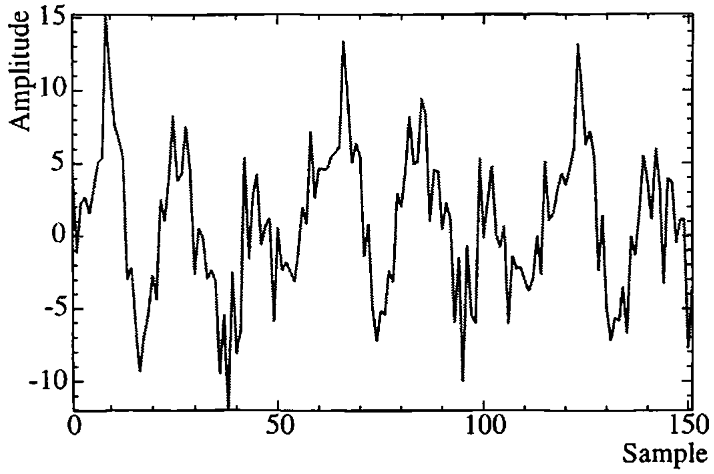


Figure 8.15: Enhanced periodic waveform.

In order to produce a simulation of vehicle engine noise, a model was constructed of the waveform periods (three of which are shown in fig8.15), as plotted in fig8.16:

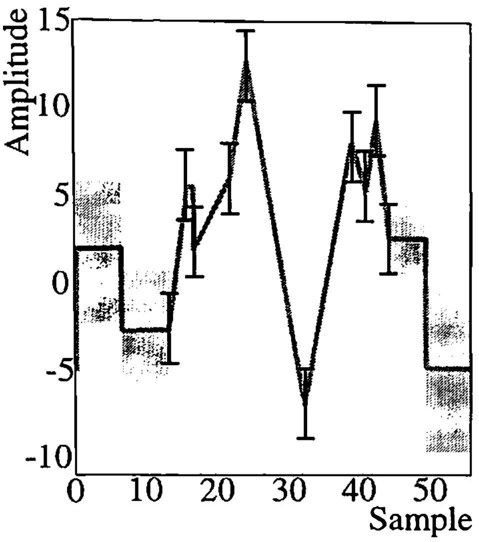


Figure 8.16: Simulation model of engine noise waveform.

The simulation model consists of a basic path having node points (represented in the figure by error bars) shifted randomly within the range found to occur in the analysed test recording. Shaded areas in fig8.16 represent further ranges of random variation from the nominal path. Thus the simulation generates a simplified engine waveform shape, having a similar sound to the original. Fig8.17a shows the original analysed engine waveform which can be compared to the simulated waveform shown in fig8.17b. Fig8.18 compares the auto-correlation of the real engine noise and that of the simulated noise, showing a close degree of similarity. Time domain scaling allows the simulator to produce different period waveforms by specifying the rotational speed to be simulated, in revolutions per minute.

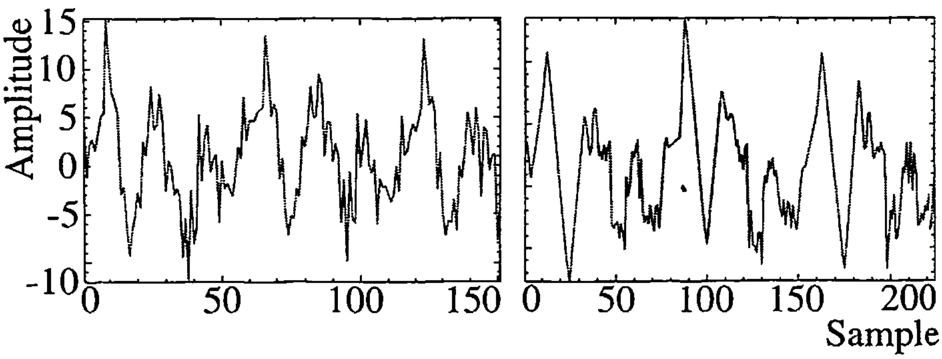


Figure 8.17: Waveforms of a) actual and b) simulated vehicle engine noise.

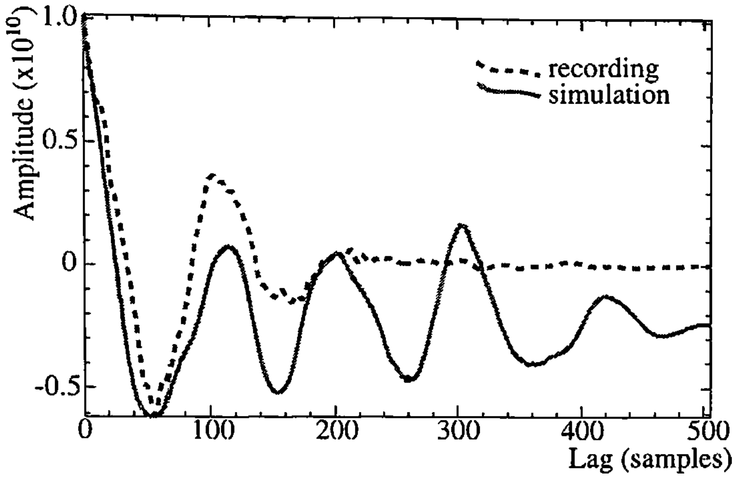


Figure 8.18: Comparison of simulated and real engine noise correlograms.

The vehicle interior noise simulator mixes a speech waveform with variable amounts of interior and engine noise, resulting in a simulation of the corrupted speech apparent to a listener within a noisy car interior. If enhanced speech, rather than raw speech, is mixed into the simulation, then the effectiveness of the applied enhancement may be determined.

8.5.2 Siren noise simulations

Siren noise analysis (see section 2.4.4) has revealed how siren frequency alters with respect to time for each of the three common sirens (two-tone, wailer and yelper). These are plotted in section A4.2, which also gives the siren sound generating equations.

The analysis was performed through inspection of spectrograms and short-term correlograms, with equations found empirically to fit the shape of the frequency curves.

Siren noise simulations may be added to vehicle interior and engine noise to create realistic acoustic background noise conditions, and have also been used in the speech enhancement analysis procedures as entries in the noise analysis codebook. The noise analysis process measures the degree of fit between acoustic background noise frames and each of the sirens in the noise codebook to determine if such a siren sound is present.

The simulated sirens sound similar to recorded sirens, but allow precise control over amplitude level and time alignment.

8.5.3 Simplified interior and siren noise simulations

Whilst the interior and engine noise simulations of section 8.5.1 produce realistic vehicle noise, based upon actual vehicle noise recordings (section 2.4.4), there exists a need for a more simplified sound. A simplified interior noise simulation is required to enable a comparison of speech enhancement techniques to be developed by other authors with the techniques described and tested within this thesis.

The simplified vehicle interior noise simulation is obtained by filtering white noise. The filter is a 20th order Blackman FIR low-pass filter with a design cutoff frequency of 0.001Hz. The frequency response of the filter is shown in fig8.19:

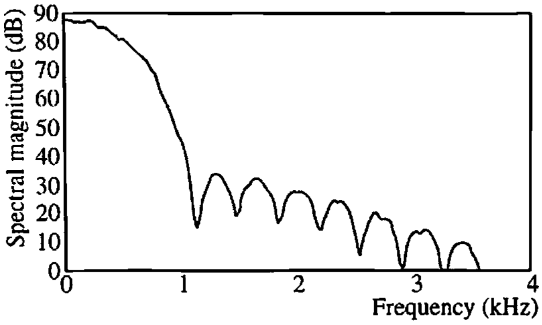


Figure 8.19: Frequency response of simplified vehicle interior noise.

Simplified siren noise is also necessary for other reasons. Due to the frequency movement of siren noise, and the frequency changes in formants, it can not be guaranteed that any instance of speech plus siren noise actually includes periods when formants are masked by the siren frequency. Thus a fixed frequency tone has been developed which coincides with the most likely F1 location of around 280Hz (see fig2.2 in section 2.2.1).

To slightly widen the frequency span of the tone masking F1, a set of 50 sinusoidal frequencies were combined in a range between 280 and 280.4Hz. The set of tones contributed to a beat frequency small enough that no beating effect was evident during its use to mask single words (section 9.6.2), lasting little over one second in duration. Individual listener responses revealed that the broader tone was found much less annoying to those participating in test procedures than a single tone, however the masking effect was judged subjectively to be greater (i.e. a pure tone would have to be louder, and thus more annoying to achieve the same degree of masking).

9 Testing

9.1 Introduction

Chapter 8 has discussed the development and testing conducted to define and refine the systems designed for use within a speech enhancing CELP codec. In this chapter, tests results are given that firstly demonstrate that LSP-based processing can enhance speech, secondly to investigate expectations of the enhancement possible through formant sharpening or broadening, and finally to characterize the performance of the speech enhancing system when integrated with a commercial CELP codec.

9.2 Test methods

Many test procedures using human listeners have been developed for speech system characterization by authors and standards bodies. Some of which are listed in section 9.2.3. The proliferation of such tests reflects the number of assumptions, independent variables, and considerations inherent in the testing processes. Here we only consider intelligibility tests, as opposed to quality tests. Some of the more important factors are discussed:

9.2.1 Test material

Speech intelligibility may be measured through phoneme, syllable, word, phrase, sentence, meaning, and any other arbitrary grouped, measured recognition rates. In general the smaller the unit tested, the more information is provided on the effect on individual parts of speech that the enhancement process has. However no reliable method has been developed of extrapolating from, for example, the results of a phoneme test, to determine effectiveness on sentence recognition (appendix 5 contains more information on aspects of intelligibility).

Test material may be familiar or unfamiliar to the listener. The latter provides a more true test of recognition rate, whilst the former provides a better indication of the effectiveness of a system to its users. Accent, phraseology and enunciation are also important.

Context plays a large part in the test results (see section A5.1.2), and must thus be accounted for when planning such a test.

9.2.2 Test conditions

Testing should ideally occur in an environment free from distractions and extraneous noise, and can be interactive, where a response is required to each question before the next is presented, or non-interactive where the test continues regardless of the listeners responses.

The listener can be presented with a choice of words in advance, or simply asked to identify an unknown word.

What is clear is that such tests do not replicate realistic conditions, or allow predictions of system performance in realistic conditions.

9.2.3 Standard tests

A survey of appropriate literature reveals the following set of standard procedures:

1. **DRT, diagnostic rhyme test** (ANSI S2.3-1989) - asking listeners to distinguish between two words rhyming by initial, such as {freak, leak}[11][104][3][125]
2. **MRT, modified rhyme test** (ANSI S2.3-1989) - asking listeners to select one of six words, half differing by initial and half by final, such as {cap, tap, rap, cat, tan, rat}[104][3]
3. **Phonetically balanced word lists** (ANSI S2.3-1989) - presenting listeners with 50 sentences of 20 words each, and asking them to write down the words they hear. [104][3]
4. **Diagnostic medial consonant test** [104]
5. **Diagnostic alliteration test** [104]
6. **ICAO spelling alphabet test** [104]
7. **2 alternative forced choice** - a general test category that includes the DRT procedure [25]
8. **6 alternative rhyme test** - a general test category that includes the MRT procedure [45]
9. **4 alternative auditory feature test** - asking listeners to select one of 4 words,

chosen to highlight the intelligibility of the given auditory feature [11]

10 **CVC, consonant-vowel-consonant test** - test of vowel syllable sandwiched between two identical consonants, with the recognition of the vowel being the listeners task. Example {T-A-T}, {B-O-B}. [2][123][112]

11 **general sentence test** - similar to the phonetically balanced word list test, but using self-selected sentences that may be more realistic in content. [112]

12 **general word test** - asking listeners to write down each of a set (usually of 100) spoken words, possibly containing realistic words. [115]

These test methods are briefly compared below:

9.2.3.1 Standard procedures

Every test shown in section 9.2.3 is recognised as a standard test procedure by the academic and industrial communities. Use of reproducible and standard speech libraries such as TIMIT and defined test settings is possible in every case. Such care is required for reproduction and comparison of the results by other authors.

9.2.3.2 Realism

Sentence testing will naturally be more realistic than phonetic testing. Thus a realism ranking would place test 11 followed by tests 12, 9, 7, 8 and then 3, 4 and 5. Tests 1, 2, 6 and 10 are most unrealistic.

9.2.3.3 Information content

The outcome of certain tests may be analysed to provide further insight into the ability of the speech enhancement system to enhance different categories of speech under various conditions. In general, tests 1, 2, 7, 8, 9 and 10 provide more opportunity to study, and perhaps further refine an enhancement system.

9.3 LSP enhancement

9.3.1 Testing requirements

When this research first proposed that line spectral pair alteration may be used to alter speech features usefully, and that this could lead to speech enhancement, there was only circumstantial evidence to support the proposal: authors had noted the relationship of LSPs to spectral features (section 5.3), spectral modification had been shown to enhance speech (section 2.2.4). However, it had yet to be demonstrated that LSP alteration could produce specific spectral changes, as shown in section 5.5, or that such alteration could enhance speech intelligibility.

A test was devised in order to demonstrate that speech enhancement could be accomplished through LSP adjustment.

9.3.2 Type of test

The LSP modification processes adjust either the position, or the amplitude and bandwidth of spectral peaks. Thus vowels, the speech features with the most distinctive spectral peaks or formants, were chosen to demonstrate enhancement. The C-V-C class of tests (number 10 in section 9.2.3) was chosen.

The test involved 15 listeners drawn from the general University population, each of whom reported no hearing abnormalities, and who were between 20 and 35 years old. Listeners had one of eight mother tongues (6 English, 3 Mandarin, 1 German, 1 Dutch, 1 Korean, 1 Hakka, 1 Cantonese and 1 Spanish).

A list of framed vowel syllables, mixed with background noise, were presented in turn to each of the listeners who were asked to identify them. The vowel syllable list included non-enhanced, formant shifted and formant broadened instances randomly distributed. Listeners were presented with a list of framing syllables and asked to fill in the vowel sound they heard between each framing set. Following the test, the responses were compared to the type of enhancement to obtain recognition rates.

9.3.3 Design of test

9.3.3.1 Speech

Recordings were made of several vowels and consonants and the most clear sounding sections sampled to computer. LSP adjustments were made using the CELP simulator of section 8.2 to copies of two of the vowels, (/a/ and /o/). The degree and type of adjustment was selected manually, and chosen by ear to provide the best degree of enhancement in the given noise:

Formant widening was set at a factor of -0.7 (a value of around 300%, shown in section 9.5.2 to provide a good degree of enhancement), and formant shifting was set to a frequency increase of 1.5 (using a non-Bark based shift as described in section 5.6.1.2).

The enhanced and non-enhanced vowels were all normalized to be of equal amplitude, as were the framing syllables.

A pair of identical consonants were used as framing syllables, selected from a set of three {/d/, /m/, /n/}, each of which was not enhanced. The vowels were positioned between the framing consonants to form a 'phrase' such as "m-a-m" or "d-o-d".

There were thus 18 possible phrases having one of:

- 3 types of framing consonant {/d/, /m/, /n/}
- 3 types of enhancement {non-enhanced, formant widened, formant shifted}
- 2 types of vowel {/a/ or /o/}

The framing consonants were used to help listeners to identify syllable onset and ending in high levels of background noise, and to prepare listeners for the vowel sound.

Section 2.2.4 commented on the complex normalization functions chosen by certain authors which can obscure results. In this case amplitude normalization was used. Informal tests performed subsequently on individuals, and using the same method, revealed that not normalizing the vowels resulted in better enhancement performance, and that normalization by sample power slightly reduced the enhancement performance.

In a digital system, volume tends to be controlled by an adjustment in signal amplitude, whereas a listener will experience psychoacoustic effects (section A1.3), related to sound pressure level which is dependent upon loudspeaker power. Thus justification exists for psychoacoustic, power or amplitude based normalization, but as CELP coders adjust gain by changes in amplitude, this method of normalization was advocated for all tests.

9.3.3.2 Noise

Simulated vehicle interior noise was used in the tests (section 8.5.1) with zero engine or siren noise components, and was mixed with the speech at one of four different relative amplitude levels before presentation to the listener.

A preliminary investigation, using three listeners, before commencing the main test found that speech mixed with the simulated noise having relative amplitude between 1.1 and 1.4 times that of the speech resulted in the listener hearing just-intelligible speech.

The just-intelligible speech condition was chosen to allow intelligibility improvements to be measurable. If speech is 100% intelligible after enhancement, this saturation prevents the degree of improvement caused by enhancement to be measured. Conversely, if speech is 0% intelligible prior to enhancement but intelligible after enhancement, the degree of improvement may be greater than the value of the resulting intelligibility - the true degree is thus not measurable. These arguments provided upper and lower intelligibility requirements to the tests, and which relative amplitude settings of {1.1, 1.2, 1.3 or 1.4} provided for each of the listeners.

The noise onset was timed to begin 1 second prior to the initial framing consonant, and end 1 second after the final consonant. This 1 second duration was long enough to negate any effects of pre- and post-stimulatory masking [70][33][68].

The simulated vehicle interior noise spectrum has a large low-frequency bias (section 8.5.1) and tails off towards high frequencies, hence the LSP shift parameter was chosen to shift formants upward in frequency to improve the formant-to-noise ratio on average (see section 4.4) as the chosen degree of limited shifting does not significantly alter formant amplitude, but noise powers are less at higher frequencies.

9.3.3.3 Test organisation

The 18 speech phrases (section 9.3.3.1), combined with the 4 relative noise amplitudes (section 9.3.3.2) yielded 72 unique test combinations for each listener. To account for possible learning effects, each listener was given a set of ten example questions before beginning the test, and a calibration set of a ten further questions. These ten calibration questions were a repeat of the final ten questions presented in the test.

The mixed audio phrases and the corresponding listeners response list were formulated randomly by a MATLAB program: the response sheet printed out, and the audio list recorded to DAT.

Each listener was seated in an anechoic chamber equipped with DAT player with active loudspeakers and given the printed response sheet. After being given examples, listeners were left alone to fill in the response sheet as they listened to the audio recording. Repeated questions were not allowed, and listeners were asked to write either 'a', 'o' or 'x' for don't know. The non-interactive nature of the test forced listeners to write down their initial response, and did not allow thinking time, or repeated questions.

9.3.4 Test results

The raw test results are given in section A6.1, along with some of the statistics obtained from these. To summarise, one listener exhibited an intelligibility reduction for both enhancements, and one listener exhibited an intelligibility reduction for formant shifting. The other 27 conditions showed improvements in intelligibility.

On average, 52% of vowels were correctly identified. With no enhancement, intelligibility was 44.4%, and improved to 54% and 59% for shifting and widening formants.
From this it can be seen that these enhancements have improved the vowel recognition on average by factors of 1.26 and 1.37 respectively.

A more detailed statistical analysis, performed over the 1148 sample questions to determine the mean improvement and its standard deviation found that, to a 95% confidence level (within two standard errors), formant shifting improved intelligibility by a factor of more than 1.15 and formant widening improved intelligibility by a factor of more than 1.21.

9.4 Selective amplification

Selective amplification was initially investigated using the CELP simulator of section 8.2 which provided parameters describing the current speech frame, and allowed adjustment of the amplitude of each reconstructed frame in order to perform the enhancement.

Unlike the speech classification system finally adopted in section 8.3, the speech classification here was performed solely using the CELP gain and LSP parameters (and not using the pitch strength value). In this investigation a full classifier was not required because LSP adjustment was not occurring. Thus only the parts of speech for which selective amplification is necessary required detection, and this detection was performed using LSP analysis, and an examination of the gain value to rule out non-speech.

The enhancement mechanism was that of the sloped attack and decay scheme described in section 7.4, and which was applied to a variety of speech recordings from the TIMIT database [118]. The enhancements resulted in an obvious increase in the audibility of fricative speech regions such as the ‘ch’ in *ch*urch or the ’sh’ in *sh*ip. This demonstrable intelligibility improvement was in contrast to the subtle enhancement through LSP adjustment, which required testing before a decision on whether it was effective or not, could be made.

Listening tests involving selective amplification in isolation were considered unnecessary due to:

- the obvious nature of the speech changes, visible as waveform alterations
- other authors have demonstrated clear enhancement due to phoneme normalization (although achieved using different methods, see section 2.2.1)
- the effectiveness of selective amplification depends upon the speech classification scheme effectiveness. This has been tested objectively in section 8.3
- the testing of a combined final system (section 9.6) was designed to indicate selective amplification effectiveness

Fig4.1 in section 4.2 demonstrated the effects of selective amplification on unvoiced or fricative parts of the speech recording.

9.5 Formant sharpening and broadening

Whilst the extreme limits of formant shifting can be established by considering the degree of formant shift that can be accommodated in speech, before that speech ceases to resemble speech (section 7.4), there is no such limit that can be applied to formant broadening or sharpening.

The broadening and sharpening of formants adjusts both the formant amplitude and bandwidth and thus both loudness and frequency effects are involved, which are difficult to quantify individually.

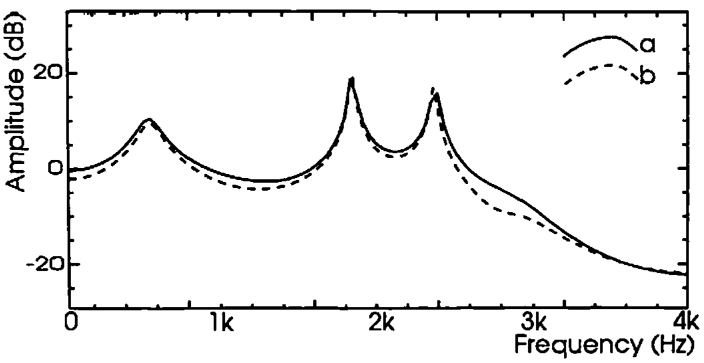


Figure 9.1: a) original power spectrum, and b) power spectrum resulting from the processing of the original speech by a standard CELP adaptive postfilter.

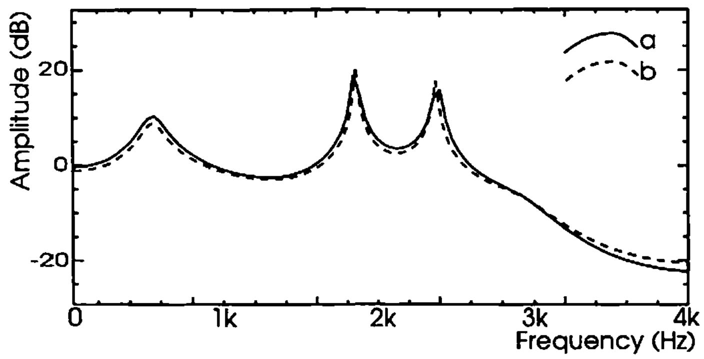


Figure 9.2: a) original power spectrum, and b) power spectrum resulting from the further narrowing of the three most closely spaced pairs of line spectral frequencies.

An example speech spectrum, and the spectrum resulting from the filtering induced by a standard adaptive CELP postfilter are shown in fig9.1. The effects of this can be compared with those shown in fig9.2 obtained through LSP processing. In both cases the three main spectral peaks exhibit little change in absolute amplitude (although the lowest of the three, possibly corresponding to F1 exhibits a small amplitude reduction in both cases), but the amplitude of the spectrum in the 'valleys' between formants has dropped. The effect being to increase the proportion of the energy of these speech frames devoted to the formant frequency regions.

A logical degree of LSP sharpening would be that which produced the most similar effect to the CELP postfilter, found by inspection to be λ=0.2 (section 5.6.1.1), and corresponding to a narrowing in LSP separation of 75% (section 7.4) for the three most prominent formants.

Despite the precedent for the degree of LSP sharpening given by the CELP postfilter, the degree of LSP broadening, and even the question of whether formants should be broadened or sharpened in noise has not been answered. For this, listening tests were required.

9.5.1 Formant shaping tests

In order to test the effectiveness of formant sharpening and broadening, a test was constructed, based upon the C-V-C method of section 9.3.

In this case, however, the tests consisted of fourteen listening sessions using six listeners and were conducted directly by computer, with the listeners in a quiet environment, wearing headphones.

Copies of the /a/ and /o/ vowels of the previous C-V-C test were subject to 11 degrees of formant widening and narrowing as shown in table 9.1:

	1	2	3	4	5	6	7	8	9	10	11
λ	2	1.7	1.4	1.1	0.8	0.5	0.2	-0.1	-0.4	-0.7	-1
%	400	270	240	210	180	150	120	90	60	30	0

Table 9.1: LSP scaling factors, and alteration in LSP separation resulting from these.

The value of λ relates to the LSP gap scaling factor of eqns5.11 and 5.12 in section 5.6.1.1, whilst the % figure indicates the percentage change in LSP separation caused by the processing, running from four times as wide (chosen as a value when distortion becomes annoying) to no line separation.

These eleven instances of scaling were applied to two vowels, each of which was framed between a pair of /m/ consonants before being mixed with a single level of simulated vehicle interior noise and presented to each listener.

In order to ensure that a single level of interior noise could be used for each listener with no possibility of all-incorrect and all-correct responses, the test was conducted interactively, and a series of calibration questions ensured that listeners correctly identified between 50% and 70% of vowels. If their responses lay outside this range, or if listeners heard almost every vowel as being identical (a common outcome, detected as being over 80% of identical responses to the randomly ordered question set) then the level of added noise was adjusted and the calibration procedure restarted.

Once calibration was complete, entailing between 20 and 50 questions per listener, the main test of eleven enhancement levels applied to two vowels (22 question phrases) began. The 22 phrases were each repeated three times and presented randomly to each listener, who was expected to respond by pressing the 'a', 'o' or 'x' keys to each question. The MATLAB program waited for a response to each question before continuing. Lasting around 15 minutes per listener, the test completed with the program calculating recognition rates for each degree of LSP enhancement.

9.5.2 Test results

Test results, tabulated in section A6.2, are plotted in fig9.3, where the average intelligibility for listeners (where the percentage of correct responses has been corrected for guesswork as shown in eqn9.1, section 9.6.1) is plotted against the degree of LSP narrowing.

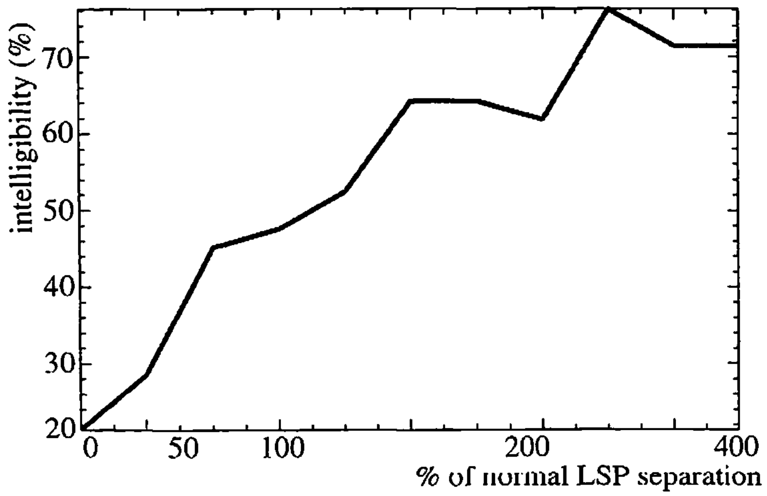


Figure 9.3: Percentage of correctly identified vowels for the given percentage change in the separation of the three closest line spectral pairs.

Inspection of the results show an average intelligibility of 52% for 120% widening. Intelligibility tails off to 21% with complete narrowing, and rises to 71% for widening line separation by a factor of four times. There appears to be a slight dip in the results for a widening of 210%, and the improvement in intelligibility due to widening tails off for the 270% and 400% results for the maximum recognition of 76% at 240% widening.

It appears that a good compromise widening of 150% to 180% produces an improvement in recognition from the unaltered rate of around 62%. The region where distortion due to processing becomes noticeable is from about 150% onwards: at 180% widening, distortion is not great, and certainly would not be noticed in appropriate levels of acoustic background noise (ie. those levels of noise occurring when enhancement must be triggered), whereas distortion at 240% widening may be.

9.6 Speech enhancement system testing

The speech enhancement algorithms, designed as presented in this thesis and operated with parameters derived from the tests described in chapters 8 and 9 were coded into ‘C’. The ‘C’ source code of the TETRA coder was modified by including the speech enhancement code. This modified commercial CELP system is described more fully in chapter 10.

In order to quantify the operation of the entire speech enhancement system, rather than the tests on component parts described do far in chapters 8 and 9, a speech intelligibility test was designed based upon this system.

The test chosen was the diagnostic rhyme test (DRT) defined in ANSI standard S2.3-1989 [3]. Such a test is a well recognised standard test, and is known to be repeatable. The results from such a test should involve finer characterization of the system and yield results more relevant to reality than the C-V-C tests conducted previously. Each enhancement has been shown to work in isolation through other testing, but this DRT procedure will detail the integrated system performance.

9.6.1 The DRT test

ANSI S2.3-1989 is intended “for use in measuring the intelligibility of English speech..”. It describes a number of trained speakers reading a list of words, the DRT list, which is recorded and later replayed to a number of audiometrically normal [3] listeners.

Listeners must be trained prior to testing: in other words, they should be familiar with the test material, and should have English as their native language. Each word is replayed to the listener who selects which word he has heard from two alternatives, rhyming by initial consonant. There must be no non-auditory cues to the sound, but the alternative words must be presented prior to hearing the sound.

Listeners must be trained to a plateau with respect to learning effects before the test begins, and ANSI S2.3-1989 sets out a considerable number of other constraints to the test procedure.

Once results are collected, the effects of guesswork and chance must be accounted for by a correction to the results:

$$R_A = R - W/(n-1) \tag{9.1}$$

where R_A is the number of correct items adjusted for chance/guessing, R is the total number of correct answers, W is the number of incorrect answers and n is the number of alternatives per question. Once such a correction has been made, R_A is considered to be a measure of the intelligibility under that condition.

The DRT questions comprise 96 alternatives (and thus 192 separate words) which can be divided into six categories [125] as shown in tables A6.3 and A6.4 in appendix 6. On completion, the results for each listener can be expressed as a percentage correct in each category for control (unenhanced) speech and enhanced speech. For the DRT test, with just two alternatives per question, the intelligibility can thus be calculated in each category by doubling the percentage correct and subtracting 100.

9.6.2 Test arrangement

The DRT word lists were supplied spoken, on DAT tape by Simoco International Ltd where they had been recorded in accordance with ANSI recommendations. These word lists were transferred to computer. Each of the 196 words was copied and enhanced by the modified TETRA codec whilst the remaining words were processed by a non-enhancing TETRA codec. Thus both sets of words had passed through the commercial CELP coder and decoder, with one set having been subject to possible LSP and gain value alterations.

A 'C' program was written to conduct the DRT test, and was run in an anechoic chamber on a portable computer. Listeners were seated in front of the computer screen which presented introductory notes, explanations, and a set of example questions before beginning anti-learning effect training and calibration. Listeners were given a hand-held button box upon which were two distinctive buttons to enable them to choose *left* and *right* words. Each test question began with two large words appearing on the computer screen (one to the left and one to the right) prior to the listener hearing the question.

The portable computer contained files of each DRT word in both enhanced and non-enhanced form. The entire set of DRT words was presented through four times: once each for enhanced and non-enhanced speech and repeated for simplified vehicle interior noise and simplified siren noise. The order in which each word, or enhancement condition was presented was randomly chosen, but with the vehicle interior noise test presented before the siren noise test.

For each word and condition in the computers internal random list, the appropriate word file was retrieved and mixed with the appropriate noise type. The recordings of each type of noise were 1.5 seconds long and the DRT words varied in duration with the maximum being below 0.75 seconds. The speech was positioned with 75% of the excess noise duration being before the speech start (in other words, there was always at least 0.5 seconds of noise prior to the speech onset - sufficient to negate temporal psychoacoustic effects [33][70]).

There was no 'don't know' condition for the listeners to report: they were asked to provide their best estimate of what they heard. A response was required before the next question was presented, and the interactive nature of the test enabled feedback: every 20 questions, listeners were told the percentage that they correctly identified.

A calibration procedure before the beginning of the test was used to adjust the relative levels of added noise with respect to the speech to ensure that their correct responses fell within the relatively wide range of between 50% and 80% answers correct. Whilst the test progressed, if less than 50% or more than 80% of the past 20 questions were correctly answered then the gain was adjusted. For both this and the calibration procedure, 20% gain adjustment steps were used.

With two enhancement conditions, two types of noise and 192 words, each listener was subject to 768 questions plus around 60 anti-learning effect and calibration questions. With each word requiring between 2.5 and 3 seconds to listen to and answer, the test progressed for between 32 and 48 minutes for each listener.

9.6.3 Enhancement

The nature of the DRT test placed certain constraints upon the enhancement system, although the intention was to test the performance of the system realistically: in as full and as unconstrained a manner as possible.

The speech enhancing TETRA codec was written in 'C', whereas the commercial use of the TETRA codec was based around DSP code, running on DSP systems. Although TETRA is a relatively efficient implementation, the compiled code can not be run at anything even approaching real time on computers, such as the chosen portable machine, that are capable of being transported to an anechoic chamber. This necessitated the TETRA processing being conducted off-line, with copies of the DRT words being processed by both the standard TETRA codec and the speech enhancing TETRA codec, and stored to hard disc.

The speech enhancement software chooses the enhancement degree and type by comparing speech and noise, and so the time-aligned noise file was input to the enhancing TETRA codec with each DRT word to be enhanced. There were thus three stored copies of the DRT words: the TETRA coded but unenhanced words, the enhanced words for use with simulated vehicle noise, and the enhanced words for use with simulated siren noise.

A final requirement is that each listener must find the words to be partially unintelligible. If words are fully intelligible then the improvement due to enhancement can not be calculated. If words are unintelligible then the degree of improvement can also not be quantified. The only appropriate means of adjusting intelligibility for different listener is alteration of the relative amplitudes of speech and noise, and in the test program, this is done by a calibration procedure, and performed if needed during the test.

The speech enhancement system included a check that amplitude levels had not exceeded the region where further increases in amplitude cause a reduction in intelligibility, as described in section 7.3. As the absolute sound pressure level experienced by the listener was not available to the speech enhancing TETRA program performing its processing off-line, this feature was disabled. The feature would have prevented selective amplification and formant sharpening (which causes an amplitude increase) when the appropriate amplitude levels were exceeded.

9.6.4 Test results

Twenty members of the general University population, aged between 19 and 55, were selected for testing, with each listener reporting no hearing abnormalities. Listeners were paid for their participation in the test and were unaware of the exact test objectives.

Sixteen of the listeners were native English speakers, three others had a high standard of English, and one listener was native Thai. The results from the latter do not differ significantly from those of the native English speakers.

DRT words from only one speaker were used in the test, selected prior to the experiment as the clearest sounding male speaker of the five alternatives supplied on DAT tape from Simoco (through a subjective comparison). The DRT test can be used to provide evidence of a communication systems absolute intelligibility. This feature is not required here: where the required outcome is only the evidence of an improvement in intelligibility between two conditions, and therefore gain alteration throughout the test and the use of only a single speaker are both acceptable (these would be disallowed if an absolute intelligibility value was required).

An initial viewing of the DRT test results indicate that enhancement processing has improved the ability of almost all the listeners to correctly recognise words in the given noise - as the percentage of correct results for enhanced words is higher than for unenhanced words.

In addition, the function of the automatic calibration procedures in the test program to maintain results in the 50% to 80% region have been verified in that only two average results fall outside this range.

The results, tabulated in section A6.3, also indicate that the simulated siren noise was generally less obstructive to intelligibility than the simulated vehicle interior noise, and that speech enhancement improved average recognition in the former noise by a more significant fraction than in the latter noise.

The rhyming word pairs used in the DRT test can be subdivided into six categories which describe the features that differ between the alternative words (section A6.4). Thus compiling average intelligibility for enhanced and unenhanced words in each category and dividing the former by the latter gives an intelligibility improvement measure. This measure relates the effectiveness of the speech enhancements to particular categories of speech, and is shown for each of the listeners, and on average for each noise type in table 9.2.

The average results for each class are plotted in fig9.4:

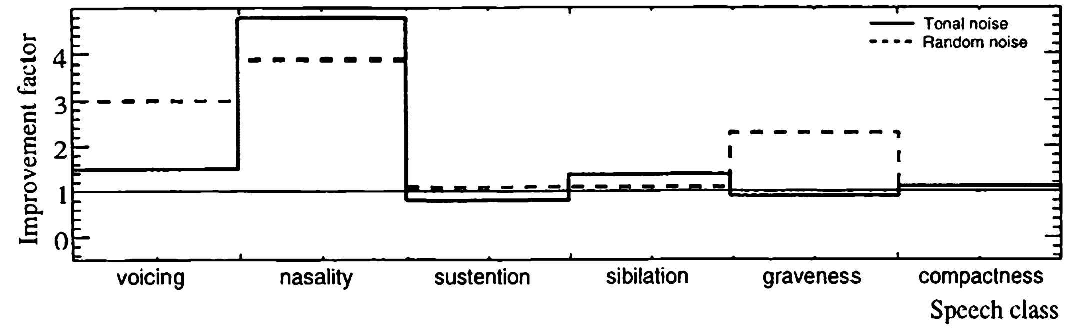


Figure 9.4: Average improvement factor in measured intelligibility rate caused by speech enhancement in each of the six speech feature classes.

	simulated vehicle interior noise					
listener	*voicing*	*nasality*	*sustention*	*sibilation*	*graveness*	*compactness*
aislam	0.2	2.4	2.0	0.6	3.5	1.8
ben	inf	6.7	1.3	1.0	11.0	1.6
dave	6.5	7.0	1.6	0.9	3.2	1.5
daveh	3.1	7.2	1.2	1.1	4.0	1.4
derekc	2.7	4.2	1.0	1.1	3.8	1.2
frank	2.7	3.4	1.0	1.0	2.4	1.1
hardw	2.6	4.0	1.0	1.1	2.8	1.1
harp	2.5	4.0	1.0	1.0	2.7	0.9
kirk	2.7	4.6	1.1	1.1	1.8	0.9
klaus	2.6	3.6	1.0	1.1	2.0	1.1
krishna	2.6	3.3	0.9	1.2	1.7	1.0
ong	2.7	3.4	1.0	1.2	1.5	1.0
robg	2.7	3.7	1.1	1.2	1.6	1.0
robj	2.6	3.4	1.0	1.2	1.6	1.0
salousm	2.6	3.4	1.1	1.1	1.6	1.0
sandhu	2.5	3.4	1.1	1.1	1.8	1.1
thai	2.3	3.4	1.2	1.2	1.8	1.0
temple	2.2	3.7	1.3	1.2	1.9	1.1
terry	2.2	3.7	1.3	1.2	1.9	1.1
zentani	2.2	3.7	1.3	1.1	1.8	1.1
average	3.0	3.9	1.1	1.1	2.3	1.1

	simplified simulated siren noise					
listener	*voicing*	*nasality*	*sustention*	*sibilation*	*graveness*	*compactness*
aislam	0.5	11.0	0.8	0.8	0.6	1.1
ben	0.5	3.9	1.0	0.7	0.6	1.1
dave	0.8	3.5	1.1	1.5	0.9	1.0
daveh	0.9	3.4	1.0	1.3	1.0	1.1
derekc	1.3	5.3	0.9	1.2	1.1	1.1
frank	1.3	6.0	0.8	1.3	1.2	1.0
hardw	1.5	5.9	0.7	1.3	1.0	1.2
harp	1.8	5.9	0.7	1.5	1.0	1.2
kirk	2.1	7.7	0.7	1.8	1.2	1.2
klaus	1.9	5.6	0.8	1.8	1.2	1.2
krishna	2.1	4.7	0.8	1.7	1.0	1.1
ong	2.1	4.9	0.8	1.7	1.1	1.1
robg	2.2	4.3	0.8	1.7	0.9	1.1
robj	2.1	4.1	0.9	1.6	0.9	1.1
salousm	2.3	4.4	0.8	1.5	1.0	1.1
sandhu	2.2	4.2	0.8	1.6	1.0	1.1
thai	2.0	4.7	0.9	1.6	0.9	1.1
temple	1.9	4.5	0.9	1.5	0.9	1.1
terry	1.9	4.6	0.9	1.5	1.0	1.2
zentani	1.8	4.4	1.0	1.6	0.9	1.2
average	1.5	4.8	0.8	1.4	0.9	1.1

Table 9.2: Improvement factor in speech intelligibility in the six feature classes between results for unenhanced and enhanced words for each of the listeners.

The DRT test results demonstrate that the speech enhancement scheme does in fact improve the intelligibility of speech under the tested conditions by an average factor of 1·9 (obtained by averaging each improvement in each class and for both types of noise).

The values in the six speech classes in table 9.2 and fig9.4 relate to the improvement in intelligibility in that class due to speech enhancement. The measure in each class itself is an indication of the ability of listeners to distinguish the presence or absence of the given feature.

When the different speech features are examined more closely, the enhancements are shown to slightly reduce the intelligibility of graveness and sustention in tonal noise. If such an enhancement system were to be used for unconstrained speech, in real situations, it is likely that further classes would be added to the speech classification, perhaps called *grave* and *sustended.* Speech falling in these classes, in tonal noise, would be subject to either no processing (which, judging from these results, would improve their intelligibility over the present 'enhanced' condition), or would be subject to another, yet undetermined, type of enhancement. This is an example of the further optimisations possible through acting on the results analysis presented here.

To summarise, the nasality class shows the greatest intelligibility increase, followed by voicing. The voiced speech feature class indicates that a choice of words was presented to the listener, one of which was voiced, and one of which was unvoiced. The unvoiced speech is enhanced through selective amplification, and voiced speech is generally enhanced through LSP adjustment, both of which thus appear to have been successful in enhancing intelligibility.

10 Summary

10.1 Discussion of results

Results from chapter 9 have shown that the novel LSP-based enhancements that were proposed can improve the recognition of vowel sounds in simulated vehicle noise. Testing of the enhancement system when integrated into a commercial CELP codec, showed that the combination of LSP and selective amplification enhancements with an automated enhancement controller can improve the intelligibility of words in vehicle type noise.

The actual improvement noted in the test does not relate easily to subjective experience. In most cases, the listeners were not able to recognise any improvement in intelligibility of individual words despite the fact that the test results indicate that enhancement is clearly present.

A given improvement in the ability of a listener to recognise a constrained set of vowel sounds does not necessarily mean that a similar improvement would be noted for words. And similarly, a given improvement in word intelligibility does not indicate a corresponding increase in sentence intelligibility (describing such a system operating under more realistic conditions).

The DRT test result is expected to relate more closely to realistic conditions than the C-V-C test result, in that the word set is less constrained, and relies upon the listener recognising a larger set of phonemes. There is evidence that given improvements in speech transmission conditions, such as through enhancement, cause a larger improvement in word recognition than in syllable recognition (this can be demonstrated through an inspection of the slopes of the 'nonsense syllables' and 'words in sentences' curves of figA5.1 in section A5.1.2. Prior to saturation at around 3dB SNR, the latter has a steeper slope [112]. This may be interpreted, with extreme caution, as the graph is not being used entirely in context, as indicating that a given improvement in conditions yields a larger recognition score increase for words than for syllables).

The relative improvements in intelligibility for word recognition over phoneme recognition may be expected to extrapolate to sentence intelligibility, or even the ability to communicate concepts. This is a natural consequence of the signal processing occurring within the human brain utilising additional factors to improve intelligibility. These include context (in terms of subject and also in terms of grammatical rules), redundancy and repetition, all absent for syllable communication. Word recognition tests provide the brain with significant extra clues involving the sequence of phonemes comprising that word in that certain combinations are more likely to occur, and certain of the possible combinations of phonemes found in the English language are unused.

In order to predict the speech enhancement to be expected from a system operating on unconstrained speech under conditions similar to the DRT test, and not accounting for non-linear effects caused by signal processing in the brain, it is possible to relate the relative occurrences of speech feature classes in continuous speech to the differing degree of enhancement found for each of those classes.

Using the test data of the speech classification tests in section 8.3.2, the relative occurrences of each of the six speech feature classes were found (see section A6.6), and this proportion multiplied by the average speech enhancement degree in each class for vehicle interior and tonal noise DRT tests (assuming an equal mixture of both) as shown in fig10.1. The average degree of speech enhancement expected for similar signal-to-noise ratio conditions for unconstrained speech is thus 2·1, with this value expected to increase if non-linear effects are considered.

class	voicing	nasality	sustention	sibilation	graveness	compactness
average enhancement:	2.25	4.35	0.95	1.25	1.55	1.1
relative frequency:	0.137	0.228	0.188	0.216	0.132	0.099
effect × frequency:	0.308	0.992	0.179	0.270	0.205	0.109
total:	2.1					

Table 10.1: Derivation of combined speech enhancement measure for unconstrained speech based upon effectiveness in each speech feature class multiplied by relative frequency of each class (as described in section A6.6).

10.2 Runtime system

The speech enhancement system proposed in chapter 7 has been tested as described in chapters 8 and 9. Results substantiate the assumptions that selective amplification, LSP-based formant shifting, and LSP-based formant widening can be used to improve the intelligibility of speech. Speech analysis, classification and speech detection methods proposed in chapter 7 incorporated a novel LSP-based speech measure. This has been investigated and found to be capable of classifying phonemes (section 8.3.1), to be similar in value to the well established zero-crossing rate measure, but more efficient, and less prone to interference from noise (section 8.3.2).

A hearing model was constructed from published material, subjectively tested and used to define a speech intelligibility measure (section 8.4). This was used in an expert system to choose type and degree of enhancement for a series of DRT tests (section 9.6). Results indicated that significant speech intelligibility enhancement occurred under the tested conditions.

The structure of the CELP coder modified with speech enhancements is thus known to be capable of enhancing speech intelligibility. The nature of such a system is that many of the system parameters and structures could probably be further adjusted to improve performance, to reduce errors or reduce complexity.

Were a speech enhancing CELP coder used in realistic situations, performance would be unknown. The system would be constructed as outlined in chapter 7, and used as shown in fig10.1.

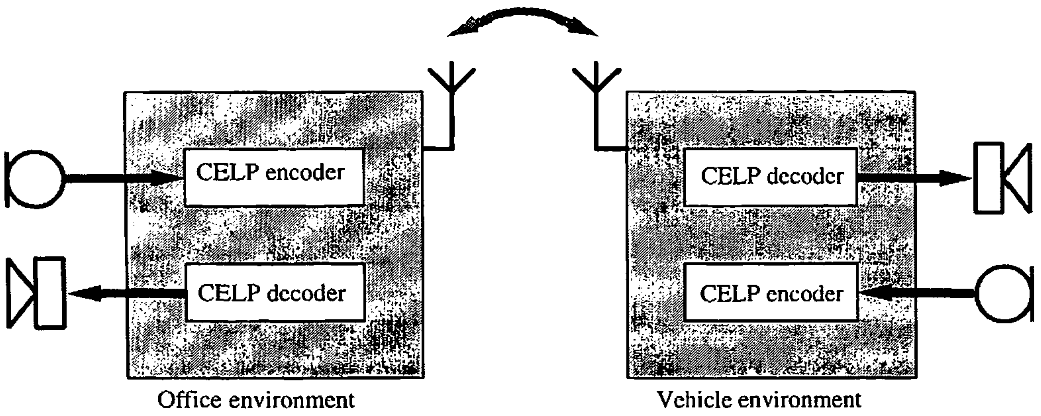


Figure 10.1: Illustration of a CELP communications system.

The intelligibility threshold of section 8.4 (the value above which the intelligibility measure must rise in order to start speech enhancement) was initially set to zero, effectively equal to the decision of the average listener model. Tests revealed that this value should be increased slightly in normal usage (section 8.4), however this is effectively applying a correction between the average listener model and the experiences of real listeners.

The intelligibility threshold value can be used as a correction for non-average hearing listeners. For example, if a user experiences speech which is unintelligible to him or her, then an offset adjustment can be made to soften the enhancement criteria. For highly-trained radio users operating a speech-enhancing CELP system, the offset may safely be increased.

Such a correction would be applicable only to listeners with reduced sensitivity but otherwise normal hearing: if the hearing loss was highly variable with frequency then the frequency response of the listener model would no longer fit the hearing frequency response of the user.

10.3 Practical considerations

The speech enhancements are designed to integrate closely with the existing structure and functions of a CELP codec (specifically, the TETRA codec [120]), and through this and an efficient implementation of algorithms, not to impose a significant complexity overhead when implemented.

The estimated number of signal processor operations per second required to implement the speech enhancement system has been calculated. The target signal processor is considered to be a generic device with all individual arithmetic operations and comparisons executing in a single instruction cycle, as does a single multiply-accumulate instruction.

Various authors have commented on the complexity of existing CELP implementations, discussed in more detail in appendix 2. These include the following (for CELP encoder unless specified otherwise):

- 12MIPS on a DSP56001 for 9.6kbit/s VSELP [35].
- 75% of the processing power of a DSP32 for multipulse LPC coder (approximately 9.5MFLOPS) [12]
- 100MFLOPS on a Cray-1 for standard 4.8kbit/s CELP [98]
- 1.2MFLOPS using a sparse codebook and frequency-domain codebook search method

[58]

- 10.6MIPS for ITU G.728 standard 16kbit/s CELP [16]
- 14MIPS for clipped overlapping codebook CELP at 4.0 to 9.6kbit/s [50]
- 80% of a DSP32C for 6.8kbit/s CELP (approximately 10MFLOPS) [52]

The TETRA codec, around which the enhancements are specifically aimed is a commercial coder, closely guarded by those organisations with access to it. Due to the highly competitive marketplace, those organisations refrain from publishing details of their implementations. However a reasonable figure for an efficient CELP implementation appears to require processing power in the region of 10MIPS.

Table 10.2 lists the estimated processing requirements of the additional algorithms required to implement speech enhancement when integrated with a CELP coder. The quoted figures have been derived in appendix 7.

Function	**IPS**
LPC → spectrum[1]	2x73260
perceptual weighting[1]	2x11655
intelligibility measure	1765
formant detection[2]	14420
speech detection	500
speech classification	633
expert system	266
formant shift[3]	1820
formant sharpen[3]	800
selective amplification[3]	33
Total (worst case):	189234

[1] performed once each for noise and speech analysis paths.
[2] not performed for non-speech or fricative frames.
[3] only one of these can be chosen for any particular frame

Table 10.2: Estimated processing requirement in instructions per second for speech enhancement operations (from appendix 7).

Even if the calculations have been underestimated by a factor of three, which is not uncommon is such estimations, then the entire processing requirement still totals less than 0.6MIPS. The speech enhancement additions to the CELP codec thus require a processing budget increase of only around 6% for the least complex of the CELP implementations listed previously.

11 Conclusion

Previous chapters have described the work conducted towards defining a speech enhancement system designed to modify the speech being decoded from a speech compression system. The type and degree of speech modification was designed to be dependent upon the acoustic background noise in the environment of the listener, the type of speech being decoded, and a model of an average human listener.

Investigation considered the nature of speech and the hearing process, the structure of the speech coder and the acoustic noise types likely to be interfering with the listeners understanding. The structure of the speech coder and the constraints of an adaptive system, determined that the speech enhancements should reside within the speech compression decoding system, and utilise the existing speech compression analysis parameters, where possible, for reasons of efficiency.

The widespread use of line spectral pairs for quantization within speech coding systems prompted this research to focus on the interpretation and alteration of such parameters, and has resulted in the definition of two novel speech enhancement schemes, and a speech classification method. Tests later explored these schemes and found that they are capable of significantly improving the intelligibility of speech. A further adaptation of an existing high distortion speech enhancement scheme led to the lower distortion but effective selective amplification enhancement.

In order to further specify likely interfering noise types and communications parameters, an example application was constructed: the speech compression system was assumed to be a CELP codec, and a target situation of a police vehicle environment was introduced. In this situation, the system operates to improve the intelligibility of the speech decoded and replayed to the vehicle occupant, when the acoustic noise levels within the vehicle become so high as to render communications difficult. This knowledge allowed the proposed enhancement algorithms to utilise raw speech parameters derived from the CELP coder, and to integrate the speech modification algorithms within the CELP decoder.

Existing speech analysis methods were combined with a novel LSP-based analysis scheme to define speech detection and classification algorithms. These operate in conjunction with a speech intelligibility detector constructed from an amalgamation of previously published spectral weighting methods, and an expert system applying hand-optimised decision rules to select and modify the speech enhancement methods used on particular frames of speech. The LSP-based analysis scheme was tested automatically using a phonetically labelled speech corpus, and found to classify at least as well, be more robust in noise, and be more efficient than a common standard method for deriving similar analyses.

After further successful component testing, the entire speech and noise analysis algorithm, listener model, expert decision system and speech enhancement methods were explored using standard multi-listener DRT intelligibility tests, and found to be capable of improving the intelligibility of speech which is severely masked by background acoustic noise in the environment of the listener. Enhancement applies to both phoneme and word sounds in specific examples of both narrow and wide-band noise. Further system optimization based upon the results from these tests could improve the enhancement still further.

A method has thus been demonstrated of enhancing the speech output from a standard CELP coder, with maximum integration with existing CELP components, and therefore low additional complexity, and the resulting high degree of efficiency.

The system could equally well be applied in any situation where high levels of interfering acoustic background noise are found, and speech compression algorithms are employed. Such conditions apply to most mobile radios, many public address or announcement systems, and more importantly, to mobile telephones.

The forthcoming generation of mobile telephones employ speech compression algorithms that are suitable for modification with the speech enhancement system described in this thesis. The enhancement system is inherently adaptive to its environment, and is only activated when it is required, needing no user intervention, a prerequisite for non-technical users.

A patent application [65] has been made for the novel techniques described in this thesis, and in addition to the planned deployment in police and other public service vehicle radio systems, licensing for use in future mobile telephone products is likely.

Further advancement of the line spectral pair adjustment and measurement techniques will require investigation into continuous speech, and to the application of the methods to other areas of speech processing such as speech recognition, speaker recognition, language recognition, voice alteration and speech synthesis.

Appendix 1: Speech and hearing

A1.1 Speech amplitude by phoneme

Speech can be broken up into individual sound units called phonemes, defined using the international phonic alphabet (IPA) [131]. In speech, phonemes are spoken with different amplitudes with average value as shown in table A1.1:

Phoneme	classification	relative intensity (dB)
f*or*d	vowel	28.3
c*ar*d	vowel	27.8
m*u*d	vowel	27.1
p*a*d	vowel	26.9
g*oo*d	vowel	26.6
h*ea*d	vowel	25.4
tr*ue*	vowel	24.9
h*i*m	vowel	24.1
t*ea*m	vowel	23.4
*r*oll	glide	23.2
*l*uck	glide	20.0
*sh*ip	voiceless fricative	19.0
sa*ng*	nasal	18.6
*m*ad	nasal	17.2
*ch*urch	affricative	16.2
*n*ight	nasal	15.6
*j*ack	affricative	13.6
*az*ure	voiced fricative	13.0
*z*oo	voiceless fricative	12.0
*s*ix	voiceless fricative	12.0
*t*ap	voiced fricative	11.8
*g*et	voiced plosives	11.8
*k*ick	voiceless plosive	11.1
*v*an	voiced fricative	10.8
*t*hat	voiced fricative	10.4
*b*ig	voiced plosive	8.5
*d*og	voiced plosive	8.5
*p*eep	voiceless plosive	7.8
*f*og	voiceless fricative	7.0
*th*ought	voiceless fricative	0.0

Table A1.1: Relative intensity of components of speech, from [113].

Examination of the table will indicate that without exception, vowels are spoken with more power, and that the range of intensity for all sounds, 28dB, is very large. Remember also that values are time-averaged.

A1.2 Speech formants

Speech is usually defined in terms of a pitch contour and formant frequencies [31]. Formants are resonant frequencies of the vocal tract which appear in the speech spectrum as peaks, shown in figA2.1.

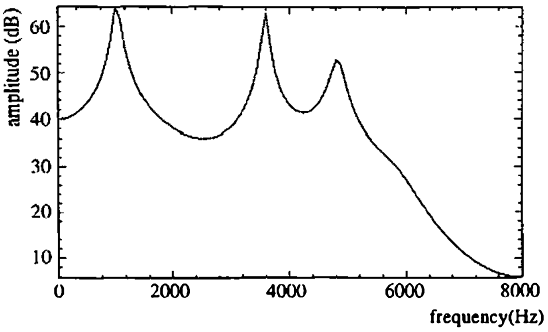


Figure A1.1: Speech spectrum showing three distinct formants. Calculated from linear prediction coefficients test vectors tabulated in (94).

Klatt [72], and other authors [36] have described formants as the single most important criterion in speech communication. Although many formants will be present in a typical speech spectrum, only the first three or so (named F1, F2, F3) contribute significantly to intelligibility or quality of speech. In fact, Fl contains most of the energy but F2 and F3, between them, contribute more to speech intelligibility [115].

The pitch contour (often called f0 - note the lower case notation) is the parameter that describes the tone of the voice (the perceived frequency), and is in effect the fundamental vocal frequency. Again pitch frequencies contain energy but contribute little to intelligibility for English and other European languages [84].

A1.3 A-weighting and equal loudness

Human subjects do not judge differing frequency signals of equal amplitude to be equal in loudness [9][28]. FigA2.2 shows typical equal loudness contours, measured in phons, where a curve of *n* phons sounds equally loud to a subject as an *n*dBA tone at 1kHz.

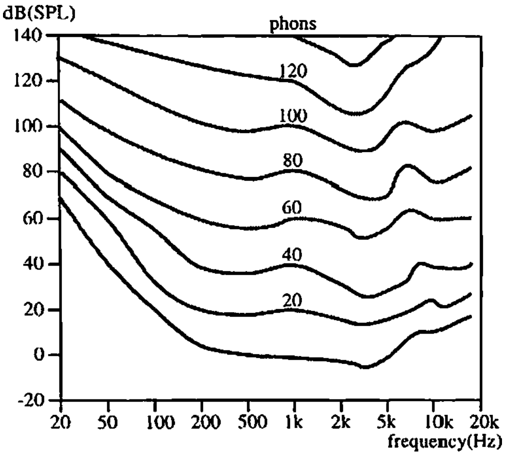


Figure A1.2: Equal-loudness contours (constructed after of figures given in (24)(111)(128)).

For speech and hearing purposes, voice power, background noise and other sound levels are usually measured in dBA, where the signal is *A-weighted* before being quantified. This is the application of a frequency weighting based on the 40-phon equal loudness contour for hearing to the signal (refer to figA2.2), now incorporated as an ISO standard. Thus all frequency components in the signal are weighted so that they make a contribution to the overall figure dependent upon their perceived loudness, rather than upon their actual intensity. Although this scheme appears reasonable, it takes no account of the ability of particular frequencies to disturb speech communications to different degrees (as the importance of frequencies to speech does not match the equal-loudness contour), or the absolute loudness of the signal (the 40 phon curve only applies to a signal of 40dB$_{SPL}$ at 1kHz). Other common measures are the ISO *B*- and *C-weighting* curves based on the shapes of the 70 and 100 phon curves respectively.

The curves of figA1.2 are the result of a number of factors, one of which is the filtering induced by the pinna, *orthotelephonic gain* [32]. The frequency distribution impinging on the eardrum differs when inner-ear headphones are used as opposed to loudspeakers, as the pinna provides around 5dB gain at 2kHz, 10dB of gain at 4kHz and 2dB gain at 8kHz [48]. The filtering effect of the pinna below 500Hz is negligible [127].

A1.4 The Bark scale

The hearing process is often considered to derive from a bandpass-filter like processing of the input sound into a number of critical bands [10][99][70]. Each critical band filter has a similar shape but the bandwidth and weighting applied to each filter depend upon frequency. The amplitude weighting with respect to frequency is considered in section A1.3 as the equal-loudness response.

It still remains however that frequency selectivity and masking effects depend on the bandwidth of each critical band. For this purpose, table A2.1 has been determined to quantify how the relative sizes of critical bands vary with frequency.

The non-linear Bark frequency scale [102] is derived from the critical band filter bandwidths, ensuring that a unit change in Bark value is reflected by a perceived unit change in frequency effect by listeners. Thus the bark scale is a psychoacoustic frequency scale.

Critical band (Bark)	Lower cutoff frequency (Hz)
1	300
2	410
3	510
4	630
5	770
6	920
7	1080
8	1270
9	1480
10	1720
11	2000
12	2320
13	2700
14	3150
15	3700

Table A1.2: Critical band scale and corresponding frequency in Hz (102).

Appendix 2: CELP coding

A2.1 CELP as a collection of algorithms

The CELP coder is considered here to be a collection of disparate algorithms. Each algorithm imparts certain of its characteristics to the synthetic speech produced at the output of the coder.

In part, the evolutionary nature of the modern CELP algorithms have contributed to the fairly loose-fitting nature of the algorithms used. Some of the algorithms included within CELP are listed below:

① Autocorrelation analysis (which eventually yields linear prediction coefficients).
✍ Pitch (long-term) analysis, which determines the period and amplitude of pitch spikes, and the parameters to be used in the pitch (LTP) filter.
③ Linear predictive filtering to yield LPC coefficients, and after conversion, LSP parameters.
④ Long term predictive (pitch) filtering.
⑤ Spcctral sharpening, under the guise of a perceptual weighting filter (PEWF).
⑥ Mean-squared-error calculation.

A2.2 CELP algorithms

Some of the many algorithms that together comprise CELP are presented below for reference:

A2.2.1 LTP filter

In the inner loop of a CELP coder, the pitch synthesis filter operates on a codevector, c, to produce a 'spiky codeword', x:

$$x(n) = c(n) + \beta x(n - M) \qquad (A2.1)$$

β is the pitch scaling factor (the strength of the pitch component) and M corresponds to the pitch period. Multiple or fractional pitch representations are occasionally used to increase the (rather poor) quality of the simple pitch filter. A three-tap pitch filter would be;

$$x(n) = c(n) + \beta_1 x(n - M - 1) + \beta_2 x(n - M) + \beta_3 x(n - M + 1) \qquad (A2.2)$$

It can be seen that the pitch (LTP) filter requires its past output values, and as M may be shorter than the length of a subframe, or possibly longer than a frame, this is an important feature of any potential realisation.

A2.2.2 LPC filter

Below is the equation of the LPC synthesis filter, a simple all-pole IIR filter:

$$y(n) = x(n) + \sum_{p=0}^{P-1} a(p)y(n - p) \qquad (A2.3)$$

output y is the synthetic subframe, input x is the filter excitation vector (the 'spiky codeword' from the LTP filter), a are the linear prediction coefficients and P is the filter order.

A2.2.3 Perceptual error weighting filter

The PEWF filter uses the LPC coefficients to sharpen the formant regions, and attenuate the frequencies surrounding these regions [34][38]. This has the effect of making the subsequent matching of the formant regions of more effect during the mean-squared-error calculation. As the formant regions are more relevant to human perception of speech, this is called a perceptual weighting filter.

The z-transform form of the equation illustrates the bandwidth expansion function of the filter, where $\zeta_1 < \zeta_2 \leqslant 1$:

$$W(z) = \frac{1 - H(z/\zeta_1)}{1 - H(z/\zeta_2)} \qquad (A2.4)$$

$H(z)$ is the standard LPC synthesis filter, and the ζ's are bandwidth expansion parameters. As

$$H(z) = \sum_{k=1}^{P} a_k z^{-k} \qquad (A2.5)$$

then

$$H(z/\zeta) = \sum_{k=1}^{P} \zeta^{k} a_{k} z^{-k} \qquad (A2.6)$$

In difference equation form, this filter can then be realised as;

$$y[n] = x[n] + \sum_{k=1}^{P} a_{k} \left\{ \zeta_{2}^{k} \cdot y[n-k] - \zeta_{1}^{k} \cdot x[n-k] \right\} \qquad (A2.7)$$

Certain CELP coders employ this filter operating on the decoded output [2][16][52][95] as a speech quality enhancement.

A2.2.4 Pitch extraction

Although there are many ways of deriving pitch parameters [1][17][43][52][103][133], the described method is probably the most common. It relies on minimising the mean-squared error between inverse LPC filtered input speech (the residual) and the value predicted using the pitch prediction formula;

If E is the mean squared error,

$$E(M, \beta) = \sum_{n=0}^{N-1} \{e(n) - e'(n)\}^{2} \qquad (A2.8)$$

then

$$E(M, \beta) = \sum_{n-0}^{N-1} \{e(n) - \beta e(n-M)\}^{2} \qquad (A2.9)$$

N is the analysis window sizc, usually a subframe, e is the residual and e' is the predicted residual. β is the pitch scaling parameter and M is the pitch delay or period. To find the optimal β, differentiate the expression and set to zero;

$$\frac{\delta E}{\delta \beta} = \sum_{n=0}^{N-1} \left\{ 2\beta e^{2}(n-M) - 2e(n)e(n-M) \right\} = 0 \qquad (A2.10)$$

so

$$\beta_{optimum} = \frac{\sum_{n=0}^{N-1} e(n)e(n-M)}{\sum_{n=0}^{N-1} e^{2}(n-M)} \qquad (A2.11)$$

If this is substituted back into the original equation, the value of M giving the optimum β is

$$E_{opt}(M) = \sum_{n=0}^{N-1} e^2(n) - E'_{opt}(M) \qquad (A2.12)$$

as only the second part of the equation varies with respect to M, it must be maximised in order to minimise the error. Thus the following must be determined with respect to each permissible value of M, and the value at which a maximum occurs, stored:

$$E'_{opt}(M) = \frac{\left[\sum_{n=0}^{N-1} e(n)\, e(n-M)\right]^2}{\sum_{n=0}^{N-1} e^2(n-M)} \qquad (A2.13)$$

An interpretation of the above process is that the pitch delay that, averaged over a whole subframe, allows the best prediction of that subframe, is chosen as the final parameter.

Once the delay has been found, the pitch scaling factor, β, is chosen as the optimal scaling factor averaged over the subframe using:

$$\beta = \frac{\sum_{n=0}^{N-1} e(n)\, e(n-M)}{\sum_{n=0}^{N-1} e^2(n-M)} \qquad (A2.14)$$

in practice, this method of pitch extraction often produces multiples of the pitch period, and thus some method of constraining the rate of change of the pitch period is used (in actual speech, the pitch period does not vary quickly). One example is used in the G.728 coder which constrains the rate of change of the pitch period to ±6 samples unless the relative strength of the new pitch value outside this range is 2.5 times as great as the old value [14].

Other methods of extracting the pitch period include [1];

- The average magnitude difference function (AMDF), which instead of using autocorrelation, calculates the magnitude difference between the residual signal and a delayed version of itself. The delay at which a minimum in the function occurs corresponds to the pitch period.
- The cepstrum technique which computes the inverse Fourier transform of the log power spectrum of the speech, and then searches for maximum values within the pitch range.
- The maximum likelihood technique, using statistical analysis on an assumed periodic signal corrupted by white noise.

- Time domain techniques that rely on measurement of waveform characteristics such as the well known parallel processing or Gold-Rabiner technique.

A2.2.5 Reflection coefficient calculation

Although reflection coefficients can be converted easily to and from LPC coefficients, the most commonly used method of calculation is directly from the input speech, with LPC parameters derived from the reflection coefficients. This method is based on autocorrelation of the speech input wave.

Assuming that the speech waveform over the frame of interest can be approximated as a linear combination of the past p samples (P is the filter order) given by:

$$x'[n] = a_1x[n-1] + a_2x[n-2] + a_3x[n-3] + \ldots + a_Px[n-P] \qquad (A2.15)$$

the prediction error, e, is given by:

$$e[n] = x[n] - x'[n] \qquad (A2.16)$$

it is possible to derive the a parameters such that they produce a minimum error over the analysis period. The least mean squared error is then given as:

$$E = \sum_n e^2[n] = \sum_n \left\{ x[n] - \sum_{k=1}^{P} a_kx[n-k] \right\}^2 \qquad (A2.17)$$

To minimise the error we then differentiate E with respect to each coefficient, and equate to zero:

$$\frac{\delta E}{\delta a_j} = -2\sum_n x[n-j]\left\{ x[n] - \sum_{k=1}^{P} a_kx[n-k] \right\} = 0 \qquad (A2.18)$$

And thus a set of linear equations having P unknowns is produced:

$$\sum_{k=1}^{P} a_k \sum_n x[n-j]x[n-k] = \sum_n x[n]x[n-j] \qquad (A2.19)$$

Where $j = 1 \ldots P$

There are now two methods of solving the set of equations, the covariance and autocorrelation methods. The former corresponds to splitting the speech into segments using a rectangular window and minimising the error only over each segment of length N. The latter method assumes that the given signal is stationary with finite energy and the range of summation is infinite, thus speech must be windowed prior to analysis. Covariance analysis does not require a soft window function and is therefore more accurate for narrow frames than autocorrelation analysis, however it does not always result in a stable LPC filter, whereas autocorrelation analysis does. In practice the majority of speech coders use autocorrelation analysis:

Given that the following relationship exists for the infinite summation:

$$\sum_{n=-\infty}^{\infty} x[n-j]x[n-k] \equiv \sum_{n=-\infty}^{\infty} x[n-j+1]x[n-k+1] \equiv \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] \tag{A2.20}$$

the equations can now be re-formulated as:

$$\sum_{k=1}^{P} a_k \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] = \sum_{n=-\infty}^{\infty} x[n]x[n-j] \tag{A2.21}$$

but now making use of the autocorrelation function:

$$R(k) = \sum_{n=-\infty}^{\infty} x[n]x[n+k] \tag{A2.22}$$

the relationship can now be written in matrix form as:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(P-1) \\ R(1) & R(0) & R(1) & \dots & R(P-2) \\ R(2) & R(1) & R(0) & \dots & R(P-3) \\ \vdots & \vdots & \vdots & & \vdots \\ R(P-1) & R(P-2) & R(P-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(P) \end{bmatrix} \tag{A2.23}$$

In practice, a window, usually Hamming, is applied to the input speech prior to calculating the autocorrelation functions, the autocorrelation results are all divided by $R(0)$ to give normalized autocorrelation coefficients, $r(i)$.

This function can then be solved using a variety of techniques, such as Durbin-Levinson-Itakura, or the Le Roux method. The former solution is very efficient but requires complicated control [90] whereas the Le Roux method is a slightly less efficient recursive formula that is most often used:

$$\begin{aligned} k_{n+1} &= \frac{e^n_{n+1}}{e^n_0} \quad for\, n = 0 \dots P \\ e_0^{n+1} &= e_0^n - k_{n+1}e^n_{n+1} = e_0^n(1 - k^2_{n+1}) \\ e_i^{n+1} &= e_i^n - k_{n+1}e^n_{n+1-i} \quad for\, i = n \dots P \end{aligned} \tag{A2.24}$$

With the initial conditions $e_i^0 = R(i) \qquad for\, i = 1 \ldots P$

The values of k derived from these relationships are the *reflection coefficients* (named by the model of the predictive filter that they describe; a number of lossless, joined tubular segments characterised by their backward reflected energy), sometimes referred to as partial correlation (PARCOR) coefficients.

A2.2.6 LPC coefficient calculation

The LPC coefficients, required for the LPC synthesis filter (and also for the analysis filter used prior to LTP parameter extraction), the perceptual weighting filter and any formant emphasis required in a post-processing scheme. To convert from reflection coefficients into LPC parameters:

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \qquad with\, 1 \leqslant j \leqslant i-1 \qquad for\, i = 1 \ldots P \qquad (A2.25)$$

and the conversion from LPC parameters into reflection coefficients is accomplished by:

$$k_i = a_i^{(i)} \qquad and \qquad a_j^{(i-1)} = \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2} \qquad 1 \leqslant j \leqslant i-1 \qquad (A2.26)$$

with i decreasing from p to 1 with initial condition of $a_j^{(P)} = a_j$ for j's between 1 and P [61].

A2.2.7 LSP representation

Line spectral pairs result from a mathematical transformation of the linear prediction parameters. The LSP representation has advantages for transmission purposes because the effects of quantization are uniform across the frequency spectrum, do not cause instability, and when LSPs are quantized, they result in lower speech degradation as compared to LPC coefficients quantized to an equivalent degree [37][48]. LSPs can also be interpolated and scaled effectively [90].

Line spectral pairs are discussed further in appendix 3.

A2.2.8 Codebook search loop

The codebook search loop contains codeword extraction, amplification, LTP and LPC filtering, perceptual error weighting and mean-square-error calculation. Using a matrix notation, the unweighted MSE measure between synthetic subframe, *y*, and real speech subframe, *s*, obtained from the *j*th codeword, ***c*** is:

$$E_j = \|s - y_j\|^2 \tag{A2.27}$$

Expanding *y*:

$$E_j = \|s - gHc_j\|^2 \tag{A2.28}$$

where g is the gain and ***H*** is the matrix form of the combined LTP and LPC filters. If we then differentiate this with respect to the gain and set to zero to find the optimum value of *g:*

$$\frac{\delta E_j}{\delta g} = -2sHc_j + 2g\|Hc_j\|^2 = 0 \tag{A2.29}$$

Thus:

$$g = \frac{sHc_j}{\|Hc_j\|^2} \tag{A2.30}$$

And substituting this back in to find the optimum error value:

$$E_j = \|s\|^2 - \frac{(sHc_j)^2}{\|Hc_j\|^2} \tag{A2.31}$$

Since only the second term changes with respect to *j* over a search of one complete codebook, the first term can be ignored: the value of *E* which is minimum (and thus maximum value of the second term) occurs for the best codebook index.

Reverting to difference equation notation, the expression is:

$$E(j) = \frac{\left[\sum_{n=0}^{N-1} s(n)y_j(n)\right]^2}{\sum_{n=0}^{N-1} y_j^2(n)} \tag{A2.32}$$

Note that the above calculation is performed once per codeword for each subframe (in real terms, that is around 140,000 times per second), which is why CELP can be so processor intensive.

A2.3 Alternative representation

It is often advantageous to represent the filtering operation by the use of matrices. This may reveal opportunities for efficiency savings, or allow operations that are not obvious when using a z-transform or similar approach. A review of matrix representation of filters followed by its relevance to CELP follows:

A2.3.1 Matrix representation

The impulse response of the LPC and LTP filters is represented by the matrix H, which is a square matrix of size equal to the frame length. The matrix is constructed as a lower diagonal Toeplitz matrix of the combined impulse responses of the two filters [119].

If the impulse response of the filters is given as $\begin{bmatrix} 1 & a_0 & a_1 & a_2 & a_3 \end{bmatrix}$ then the lower diagonal Toeplitz matrix representing that filter will be:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ a_0 & 1 & 0 & 0 & 0 \\ a_1 & a_0 & 1 & 0 & 0 \\ a_2 & a_1 & a_0 & 1 & 0 \\ a_3 & a_2 & a_1 & a_0 & 1 \end{bmatrix} \qquad (A2.33)$$

Note that the LPC and LTP syntheses filters are actually IIR filters. Although the impulse response of an FIR filter is given by the coefficients, the case is slightly more complicated for IIR filters, no account of memory stored in the filter is represented by the matrix. Thus the filter memory must also be stored as a vector prior to any arithmetic.

To determine the impulse response of the required filters, construct a matrix as shown in eqnA2.33 using the IIR coefficients zero padded to the length of the frame, making the lower left hand corner of the square matrix contain zeros. This matrix is then inverted - a relatively simple procedure due to its structure, and the resultant lower triangular Toeplitz matrix holds the impulse response. See [64] for some detail on matrix form computational complexity.

The action of filtering is now performed by matrix multiplication, and the addition of the filter memory vector to the output. An alternative way of describing this process is, the addition of the zero-state response to the zero-input response.

The zero-input response of a filter accurately quantises the filter memory. Feeding a zero block of data (of size given by the frame size under consideration) into an IIR filter yields the zero-input response. If an IIR filter with zero memory (or a matrix multiplication) yields a result that is added to its zero-input response then the combined result is equal to a filtering operation by that same IIR filter with memory.

A2.3.2 Simplified CELP

Using the matrix representation from section A2.3.1 we can see that the task for choosing a correct codeword is to maximise the measure:

$$\tau_j = \frac{(sHc_j)^2}{\|Hc_j\|^2} = \frac{(sHc_j)^2}{c_j^T H^T H c_j} \qquad (A2.34)$$

This would be calculated for each codeword described by the index, j. Using matrix notation, we can actually pre-calculate some of the multiplications on a subframe-by-subframe basis rather than for each codeword using $\Psi = H^T s$ and $\Phi = H^T H$.

The former relationship is called backwards filtering, or reverse time series filtering. The error measure becomes:

$$\tau_j = \frac{(\Psi^T c_j)^2}{c_j^T \Phi c_j} \qquad (A2.35)$$

This only provides a simplification under certain circumstances, such as in ACELP. Precalculations for each subframe require N^2 operations for Ψ and N^3 operations for Φ. Inner loop subframe comparisons now require $C*(N^2+N+2)$ operations [68].

In the case of ACELP, where almost all of the input frame samples are of zero value, it is possible to use prior knowledge of the positions of the non-zero values to simplify the measure of equation A2.35 [57]. If the codeword has only R pulses that are non-zero, in locations described by m and with sign described by b, the mean square measure is now:

$$\tau_j = \frac{\left[\sum_{i=0}^{R-1} b_i \Psi(m_i)\right]^2}{\sum_{i=0}^{R-1} \Phi(m_i, m_i) + 2\sum_{i=0}^{R-2}\sum_{j=i+1}^{R-1} b_i b_j \Phi(m_i, m_j)} \qquad (A2.36)$$

$$b_i = \begin{cases} 1 & \mathit{if\ i\ is\ even} \\ -1 & \mathit{if\ i\ is\ odd} \end{cases} \qquad (A2.37)$$

Ψ and Φ require N^2 and N^3· calculations per subframe respectively. In addition to this, the inner loop subframe comparisons now only require R+1 multiplications, R^2+2*R additions and one division for each codeword, or about C*(R^2+3*R+2) operations. Using realistic CELP parameters, this is a reduction in processing over the standard form from around 180 to 38 million operations per second [68]. In fact, further reduction is possible because, when incrementing the code index, only one pulse position changes (assuming the code index to pulse position relation is well designed) and thus only subtraction of old value and addition of the new are required prior to division.

Appendix 3: Line Spectral Pairs

A3.1 Generation of LSPs from LPC coefficients

The process of converting from LPC coefficients (*a's*) to LSP is explained below, however, for a rigorous proof of the conversion process, refer to [91][92][93].

Firstly define polynomials to represent two extreme conditions relating to LPC coefficients [15][48]. A symmetric polynomial, A_k, is formed by adding the time-reversed and forward LPC coefficients:

$$A_k = a_k - a_{(p+1-k)} + A_{k-1} \qquad (A3.1)$$

for k between 1 and p, with initial condition $A_0 = 1$.

Similarly, an antisymmetric polynomial is formed by subtracting the time-reversed and forward LPC coefficients:

$$B_k = a_k + a_{(p+1-k)} - B_{k-1} \qquad (A3.2)$$

for k again between 1 and p, with initial condition $B_0 = 1$.

It is then necessary to determine the complex roots of each equation, which will lie on the unit circle in the z-plane if the original LPC filter was stable. It has been shown [106] that this condition also results in the roots of *A* and *B* alternating around the unit circle, and in fact, any alternating pairs of line spectral frequencies, in turn represent a stable set of linear prediction coefficients.

Roots can be found by a number of methods including Newton-Raphson, evaluation at intervals around the unit circle in the z-plane (looking for sign changes in either real or imaginary components, indicating the function crossing the origin - the angle when it crosses representing the line spectral frequencies) followed by zooming-in on areas of sign change for increased resolution, FFT, or any other more computationally efficient algorithms (perhaps using Chebychev polynomials [37][48]).

Once the complex roots, θ_k, have been found, the line spectral frequencies are easily determined from the formula given in eqnA3.3:

$$\omega_k = tan^{-1}\left(\frac{Re\{\theta_k\}}{Im\{\theta_k\}}\right) \tag{A3.3}$$

Listing A3.1 shows a routine written in ‘C’ like language for RLaB, the mathematical development tool, to derive LSP coefficients from LPC values:

```
% This file defines a function, lpc_lsp, which derives line-spectral
% frequencies for the given linear prediction coefficients, a.
%
% © Ian McLoughlin, 10 November 1995
%
% Note: format of input argument should be;
% [1,a1,a2,a3,a4], with an initial '1' and an even number of coeffs.
%
%-----------------------------------------------------------------
require(roots);
lpc_lsp=function(a)
{
        local(p,k,A,B,r1,r2,theta1,theta2,theta);
        global(pi);
      p=length(a);
      %derive the coefficients for P'(z) and Q'(z)
        A[1]=1;
        B[1]=1;
        for (k in 2:p)
        {
                A[k]=(a[k] - a[p+2-k]) + A[k-1];
                B[k]=(a[k] + a[p+2-k]) - B[k-1];
        }

        r1=roots(A);
        r2=roots(B);
        for (k in 1:p-1)
        {
                if (real(r1[k]) < 0)
                {
                        theta1[k]=pi-abs(atan(imag(r1[k])/real(r1[k])));
                else
                        theta1[k]=abs(atan(imag(r1[k])/real(r1[k])));
                }
                if (real(r2[k]) < 0)
                {
                        theta2[k]=pi-abs(atan(imag(r2[k])/real(r2[k])));
                else
                        theta2[k]=abs(atan(imag(r2[k])/real(r2[k])));
                }
        }
        p=p-1;
        for (k in 1: p/2)
        {
                theta[k]=theta1[k*2];
                theta[k+(p/2)]=theta2[k*2];
        }

        theta=sort(theta).val;          %Sort into ascending order
        return theta;                   %return the line-spectral frequencies
}
```

Listing A3.1: RLaB code routine to convert from LPC to LSP coefficients.

A3.2 Generation of LPC coefficients from LSPs

Conversion from LSPs to LPCs is a mathematically efficient process. The proof is again to be found in [83]. It is advantageous to begin the process in the cosine domain, where q is the array of ordered line spectral frequencies (ordered such that the lowest is first).

$$q_k = cos(\omega_k) \qquad (A3.4)$$

Where ω is a line spectral frequency expressed in radians..

The following recursive equations are now solved:

```
for $i = 1 \ldots p$
    $f_1(i) = -2f_1(i-1)q_{2i-1} + 2f_1(i-2)$
    for $j = i - 1 \ldots 1$
        $f_1(j) = f_1(j) - 2f(j-1)q_{2i-1} + f_1(j-2)$        (A3.5)
```

with the initial conditions $f_1(0) = 1, \quad f_1(-1) = 0.$

The coefficients $f_2(i)$ are calculated similarly, but with q_{2i-1} replaced by q_{2i}.

These values are the used in a second set of equations:

$$f'_1(i) = f_1(i) + f_1(i-1) \qquad (A3.6)$$

$$f'_2(i) = f_2(i) - f_2(i-1) \qquad (A3.7)$$

Which are then used to form the LPC coefficients from:

$$a_i = \frac{1}{2}f'_1(i) + \frac{1}{2}f'_2(i) \qquad for\,i = 1 \ldots 5 \qquad (A3.8)$$

$$a_i = \frac{1}{2}f'_1(i-5) - \frac{1}{2}f'_2(i-5) \qquad for\,i = 5 \ldots 10 \qquad (A3.9)$$

Listing A3.2 gives the RLaB routine required to convert a set of line spectral pair parameters into a set of linear prediction coefficients.

```
% This file defines a function, lsp_lpc, which derives linear pred-
% iction coefficients for the given line spectral pairs, w
%
% © Ian McLoughlin, 10 November 1995
%
% Note: output format has an initial 1 followed by an even number
% of coefficients;  [1,a1,a2,a3,a4]
%
%-------------------------------------------------------------------
lsp_lpc=function(w)
{
        local(k,q,n,f1,f2,f1b,f2b, a2);
        p=length(w);

        for (k in 1:p)
        {
                q[k]=cos(w[k]);
        }

        f1[10]=1;
        f1[9]=0;
        for (n in 1:p/2)
        {
                f1[10+n]=-2*q[2*n-1]*f1[10+n-1] + 2*f1[10+n-2];

                for (k in n-1:1:-1)
                {
                        f1[10+k]=f1[10+k] - 2*q[2*n-1]*f1[10+k-1] + f1[10+k-2];
                }
        }

        f2[10]=1;
        f2[9]=0;
        for (n in 1:p/2)
        {
                f2[10+n]=-2*q[2*n]*f2[10+n-1] + 2*f2[10+n-2];

                for (k in n-1:1:-1)
                {
                        f2[10+k]=f2[10+k] - 2*q[2*n]*f2[10+k-1] + f2[10+k-2];
                }
        }

        f1b[1]=f1[11]+1;
        f2b[1]=f2[11]-1;

        for (n in 2:p/2)
        {
                f1b[n] = f1[10+n] + f1[10+n-1];
                f2b[n] = f2[10+n] - f2[10+n-1];
        }

        for (n in 1:p/2)
        {
                a2[n]       = 0.5*( f1b[n] + f2b[n] );
                a2[n + p/2] = 0.5*( f1b[(p/2)-n+1] - f2b[(p/2)-n+1] );
        }

        return([1,a2]);
};
```

Listing A3.2: RLaB code routine to convert from LSP to LPC coefficients.

A3.3 LSP simulation

FigA3.1 shows the LSP simulation process used to derive the information required for a plot of the linear prediction spectrum with LSP values overlaid:

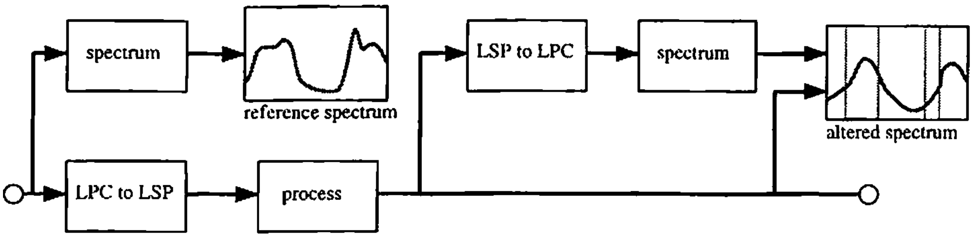


Figure A3.1: Test process for investigating LSPs.

The test process begins with a reference speech frame, represented as a set of LPC coefficients describing a reference spectrum. Routines then extract LPC coefficients from a given frame of data. These LPC coefficients are then converted to line spectral pairs, processed, and new LPC coefficients and spectrum determined. Thus the effects on the underlying spectrum of the process under test can be visualised. This basic procedure has been used to conduct various tests, including those presented in chapter 5.

Conversion of LPC coefficients to line spectral pairs and vice versa is accomplished using the algorithms presented in sections A3.1 and A3.2 respectively.

The LPC filter frequency response is calculated by substituting $z = e^{-j\theta}$ into the linear prediction equation (eqn A2.3 in section A2.2.2) with θ being swept from 0 to π. EqnsA3.10 and A3.11 show, respectively, the real and imaginary components of the linear predictor function:

$$\begin{aligned} Re\{A(\theta)\} = {} & a_1 + a_2nr + a_3(nr^2 - ni^2) + a_4(nr^3 - 3nr.ni^2) + \\ & a_5(nr^4 + ni^4 - 6nr^2ni^2) + a_6(nr^5 + 5nr.ni^4 - 10nr^3ni^2) + \\ & a_7(nr^6 - ni^6 + 15nr^2ni^4 - 15nr^4ni^2) + \\ & a_8(nr^7 + 35nr^3ni^4 - 7nr.ni^6 - 21nr^5ni^2) + \\ & a_9(nr^8 + ni^8 + 70nr^4ni^4 - 28nr^2ni^6 - 28nr^6ni^2) + \\ & a_{10}(nr^9 + 9nr.ni^8 - 84nr^3ni^6 + 126nr^5ni^4 - 36nr^7ni^2) + \\ & a_{11}(nr^{10} - ni^{10} - 45nr^8ni^2 + 45nr^2ni^8 + 210nr^6ni^4 - 210nr^4ni^6) \end{aligned} \qquad (A3.10)$$

$$Im\{A(\theta)\} = a_2ni + a_32nr.ni + a_4(3nr^2ni - ni^3) + a_5(4nr^3ni - 4nr.ni^3) +$$

$$\begin{aligned}&a_6(ni^5 + 5nr^4ni - 10nr^2ni^3) + a_7(6nr.ni^5 + 6nr^5ni - 20nr^3ni^3) + \\ &a_8(7nr^6ni - ni^7 - 35nr^4ni^3 + 21nr^2ni^5) + \\ &a_9(8nr^7ni - 56nr^5ni^3 + 56nr^3ni^5 - 8nr.ni^7) + \\ &a_{10}(ni^9 + 126nr^4ni^5 + 9nr^8ni - 84nr^6ni^3 - 36nr^2ni^7) + \\ &a_{11}(10nr^9ni - 120nr^7ni^3 + 252nr^5ni^5 - 120nr^3ni^7 + 10nr.ni^9)\end{aligned} \tag{A3.11}$$

where $nr = sin\theta$ and $ni = cos\theta$.

Once real and imaginary components have been found, the magnitude is calculated, giving the amplitude response of the system at the current frequency:

$$mag(\theta) = 10log_{10}\left\{tan\left(\frac{Re\{A(\theta)\}}{Im\{A(\theta)\}}\right)\right\} \tag{A3.12}$$

This provides the frequency response of the linear prediction analysis filter $A(z)$, with the synthesis filter being defined as the inverse of this. Thus figures which plot the LPC filter spectrum, actually plot the inverse of eqnA3.12.

Appendix 4: Implementation of the classifier

A4.1 Formant detection

This section described formant detection method tests, and the implementation of each of these methods.

A4.1.1 Differentiation of LPC spectrum

Operating on the linear prediction spectrum, *s* (obtained as shown in section A3.3), this analysis method calculates the first and second derivatives:

$$s_i' = s_i - s_{i-1} \qquad (A4.1)$$

$$s_i'' = s_i' - s_{i-1}' \qquad (A4.2)$$

$$for\ i = 1 \rightarrow window\ length$$

then checks for gradient changes due to maxima:

$$if\ sign(s_i') \neq sign(s_{i-1}')\ and$$

$$if\ \{\tfrac{1}{2}(s_i'' + s_{i-1}'')\} < 0\ then\ i\ is\ a\ formant\ frequency \qquad (A4.3)$$

the averaging process, when searching the second derivative in eqnA4.3, is required because the gradient change occurred somewhere between index *i* and *i-1*.

Detected formants are then located between the index positions where eqnA4.3 is satisfied. The formant amplitudes are calculated as the average amplitude of the spectrum at the two index points between which the gradient changes.

A4.1.2 Solving LPC polynomial

The LPC polynomial, eqnA2.3 in section A2.2.2, describes a resonant circuit modelling the vocal tract. If this polynomial is solved, the complex conjugate roots resulting from this solution give the resonant frequencies.

A numerical method is used to solve the roots (simulations have used the MATLAB built-in 'roots' function; as this method has not been selected for implementation, further details are not required. However conversion of LPC coefficients to LSP parameters also requires polynomial solving, see section A3.1).

The angular frequency of the resulting roots gives the formant frequencies, and the magnitude gives a measure of the formant bandwidth.

A4.1.3 Spectral peak picking

Prior knowledge of likely formant position (fig2.2, section 2.2.1) allows for maxima to be determined within small ranges, however spectral tilt will cause confusion. For example, the low frequency bias in a speech frame will often cause the low-frequency end of the spectrum to have much higher amplitude than the high end. Furthermore, the maximum frequency in the band within which F1 is usually located will often not be due to F1, but due to low-frequency bias (therefore the maximum amplitude will be the lowest frequency in that band).

A more general solution to the problem is to manually search for gradient changes in a process similar to that of section A4.1.1. As the index i is swept across the spectrum, s, then if s_i is found to be greater than s_{i-1}, then the subsequent samples are tested to determine if these are greater or less than s_i. If the first non-equal sample is less, then i is a formant location, otherwise it is not.

A further constraint is then applied to reject any candidate formants that are located within 300Hz of lower formants. This figure having been found to provide good results through testing, to reduce the occurrences of double-peak confusion.

A4.1.4 LSP-based formant detection

The *p* line spectral pair parameters are analysed to determine the three closest pairs, presumably corresponding to to the three strongest formants. The formant frequency is considered to be at the mid point between each of these three pairs.

A4.2 Noise detection

In order to detect a given noise, it is advantageous to first define it. This section initially describes three types of siren noise (two-tone, wailer and yelper), and presents simulation equations for these. The equations were developed through analysis and inspection of short-time spectrograms, autocorrelograms and FFT spectra. The siren noises are shown in sections A4.2.1, A4.2.2 and A4.2.3.

Section A4.2.4 considers methods of siren noise detection, and section A4.2.5 describes the implementation of a chosen siren detection method.

A4.2.1 Two-tone

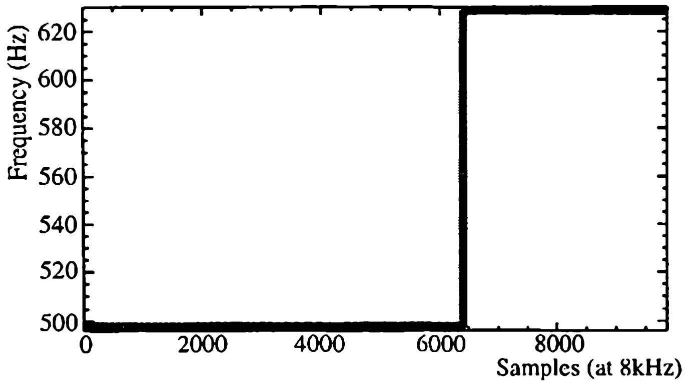


Figure A4.1: Temporal variation in two-tone siren frequency (one period).

Measurements of the low-period frequency, from correlation, inspection of power spectrum and spectrograms ranged from 460 to just under 500Hz, with the high frequency period at between 599 and 620Hz. Spectrogram plots allowed easy identification of the period of high and low sections.

A4.2.2 Wailer

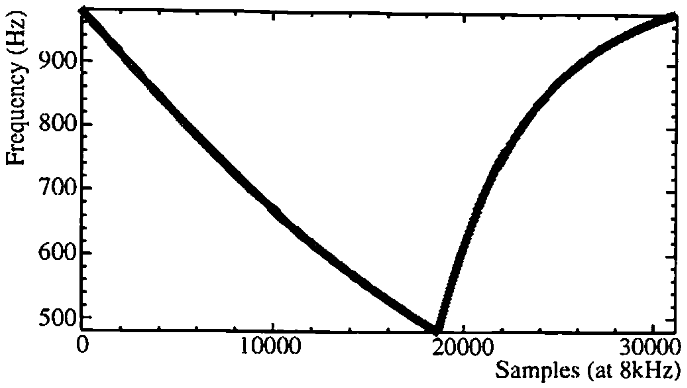


Figure A4.2: Temporal variation in wailer siren frequency (one period).

On the assumption that analogue electronics produced the siren, exponential functions were chosen to model the sound. Parametric curve fitting revealed an exception to the exponential model for the upward rising frequency part of the waveform. The decreasing frequency part of the curve is given by eqnA4.4:

$$f = 980 \times e^{-t.3.814 \times 10^{-5}} \qquad (A4.4)$$

where the sample index t runs from 0 to 18714. The increasing frequency part is modelled by eqnA4.5:

$$f = \{980 - 500 \times 1.6^{-t.4{,}982 \times 10^{-4}} + i.2.164 \times 10^{-3}\} \qquad (A4.5)$$

with the sample index t running from 0 to 12476.

A4.2.3 Yelper

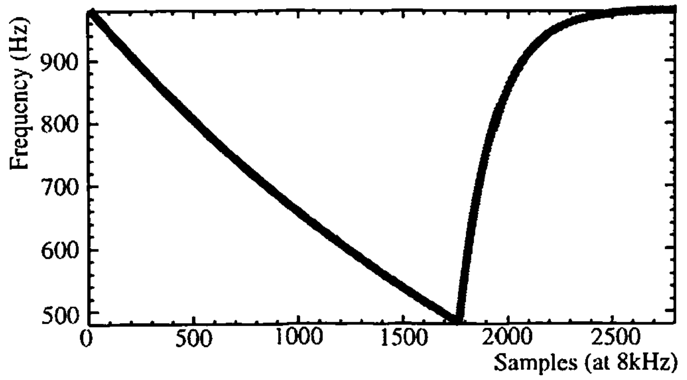


Figure A4.3: Temporal variation in yelper siren frequency (one period).

Again working on the assumption that analogue electronics produced the siren, exponential functions were chosen to model the sound. The decreasing frequency part of the curve is given by eqnA4.6:

$$f = 980 \times e^{-t.4.059 \times 10^{-4}} \tag{A4.6}$$

with the sample index t running from 0 to 1760. The increasing frequency part of the siren is modelled by eqnA4.7:

$$f = \{980 - 500 \times e^{-t.5.981 \times 10^{-3}}\} \tag{A4.7}$$

with the sample index t running from 0 to 1040.

A4.2.4 Methods of siren detection

The initial method of siren detection that was tested involved storing a large number of past spectral arrays. Each array containing the power spectrum of successive analysis frames. Each type of siren noise was correlated against the past spectral arrays by considering each possible non-repeating siren phase delay, and summing the amplitudes of the frequency bins of the spectra of past frames within which the siren frequency of each tested phase delay would reside. A perfect match would result in spectral maxima from each past analysis spectrum being summed. The resultant value from each tested phase delay required normalization, and comparison with other phase delay values.

The maximum phase delay position is the most likely phase of the tested siren. Comparing these values for each siren yields the most likely siren type present. It was hoped that adaptive thresholding of this likelihood would then allow the presence or absence of sirens to be stated.

The scheme outlined was unsuccessful for a number of reasons. Firstly, the mean and variance of the frequency locations of the sounds comprising each sirens are different, and thus even when no sirens are present, the type of siren mostly located in frequency regions of high spectral energy yields the highest score. The large variance in noise frequency distribution means that normalization to account for this would have to be both adaptive and would have to ignore siren sounds. In particular, spectral tilt, such as high levels of low frequency noise confused results.

Secondly, the narrowness of the frequencies from which sirens are constructed mean that even small phase variations (for yelper and wailer sirens) and small variations in the absolute frequency detected for the two-tone siren mean that when such siren noise is present, the correlation process may not be summing spectral peaks, but perhaps the frequency bins adjacent to the spectral peaks. Normalization then produces a small measure result.

Despite the poor response of this technique, adapting the process slightly to calculate the amplitude mean and variance for each frequency bin in the past arrays could detect tonal noise. The technique was then extended to also calculate the mean and variance for variable frequency sounds. For this, not only was the calculation performed for each fixed frequency, but was performed between a given frequency and, respectively, each frequency in the previous spectrum that could be related to the current frequency through the siren noise equations (then to the previous spectral array and hence continuing through the history arrays). Each possible frequency path was normalized with respect to its length, and were compared in strength. This method suffered the same phase misalignment and spectral tilt disadvantages as the previous scheme.

To counter possible phase misalignment problems, the methods were adjusted to sum frequency bins in a small region (of ±30Hz) around the expected model frequency. Whilst performance was slightly improved, errors due to variation in power across the spectrum increased.

The degree of computation required to perform the two detection schemes for three sirens, over many analysis frames at every possible phase delay were considered excessive, as were memory storage requirements for the history of 50 or more 200-sample frequency arrays.

Simplified approaches were then developed, detecting frequency peaks in the noise spectrum, and storing these for past frames. Then the frequency tracks for each possible siren phase delay, for each siren, could be matched against the stored frequency peak positions using a mean squared error criterion.

A4.2.5 Siren noise detection

Given the criteria that LSP parameters are available for analysis, a spectral peak picking technique was developed using narrowest-LSP detection (such as was used for formant detection in section A4.1.4).

The three most prominent (closed paired) LSP-detected spectral peaks in the range of 200Hz to 1kHz were calculated for each speech analysis frame and stored in a history array. The frequency tracks for each possible siren phase were matched against the history array. During this calculation, if no spectral peaks are present in the history array within 200Hz of the expected location, then the position is ignored for calculation purposes. This is because the spectral peak detection algorithm will only pick out strong spectral peaks, not those obscured in high levels of noise, and so missing points are likely.

The LSP detection algorithm is unaffected by moderate levels of spectral tilt (ie. DC contamination or high levels of high-frequency noise), which prevented most of the methods of section A4.2.4 from performing adequately.

A4.3 Speech classification

Speech classification is performed according to the outcome of one or several speech analysis methods. The *n* resultant values form an *n*-dimensional feature vector which describes the speech. The speech classification is then performed by defining regions in *n*-dimensional space, with feature points falling in a particular region being classed together.

Efficiency dictates that classification regions are best bounded by straight lines (in two dimensions) or planes (in higher dimensions). This considerably simplifies the classification process.

Once the format of the feature vector is defined (ie. the speech analysis methods are known), automatic or manual methods may be employed to find equations for the classification boundaries. For manual optimisation, a number of plots can be made of sets of two of the measures in the feature vector for different speakers and different speech, and compromise decision boundaries drawn by hand.

The process of defining decision regions in the case of a two-dimensional feature vector formed by LSP-based and frame power measures is illustrated in figA4.4. The experiment used to derive this graph is that of section A6.5, with the phonemes deriving from an automatic speech analysis of a number of sentences from the TIMIT database [118].

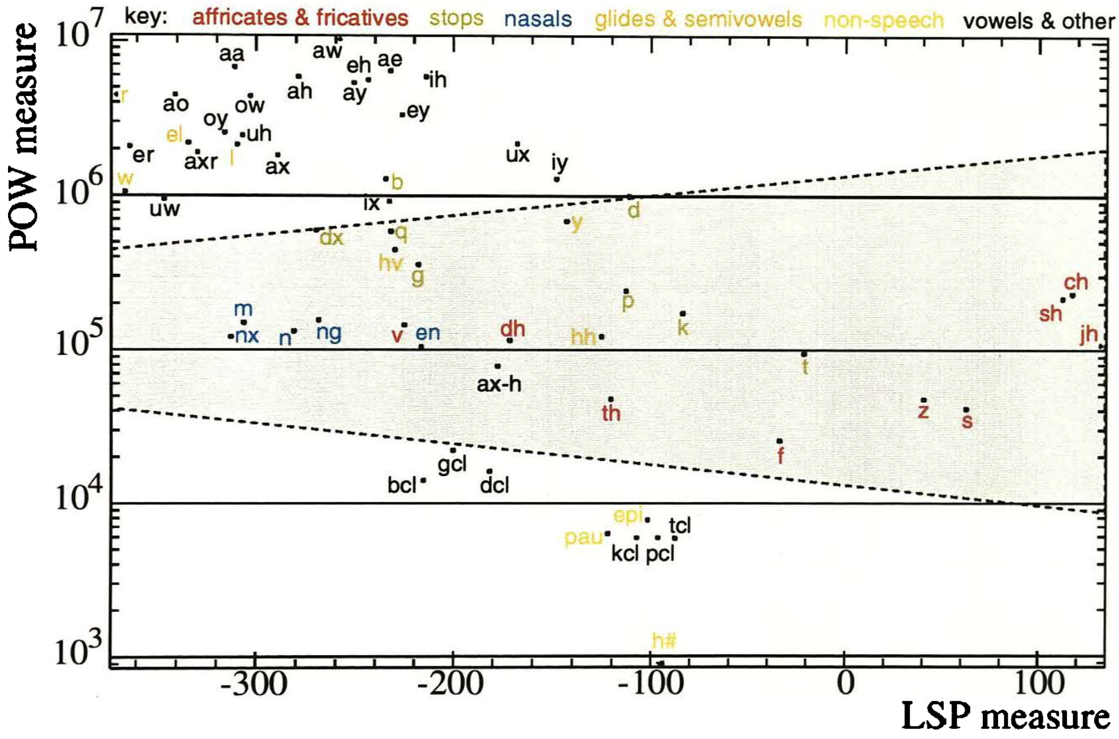


Figure A4.4: Example straight line classifications performed to determine example phoneme regions in a LSP measure against frame power measure space, with phoneme types shown.

Two straight lines have been drawn in figA4.4, defining a wedge-shaped internal region. Although this is an example classification region, it can be seen that the phonemes above the top line are predominantly vowels, and are all voiced. Those below the lower line are non-speech regions or closures (those periods in speech when the tongue or lips block the vocal tract and sound is momentarily paused.

The phonemes within the wedge-shaped region are all fricatives, affricatives and nasals, and include the unvoiced schwa *ax-h*, along with *hv*, *y* and *hh* (the ‘h’ in “ahead”, the ‘y’ in “yacht” and the ‘h’ in “hay” respectively). These are all low energy unvoiced phonemes, most likely to require speech enhancement by selective amplification when heard in noise.

In fact the method of plotting out many of these graphs to produce compromise speech classifications is extremely time consuming. The more efficient adopted method is to begin with a plot of figA4.4 to define approximate classification regions, and then to process a large number of speech recordings, comparing the speech classification output to the waveform (and phonetic transcription, if available), and time-domain plots of the feature vector measures.

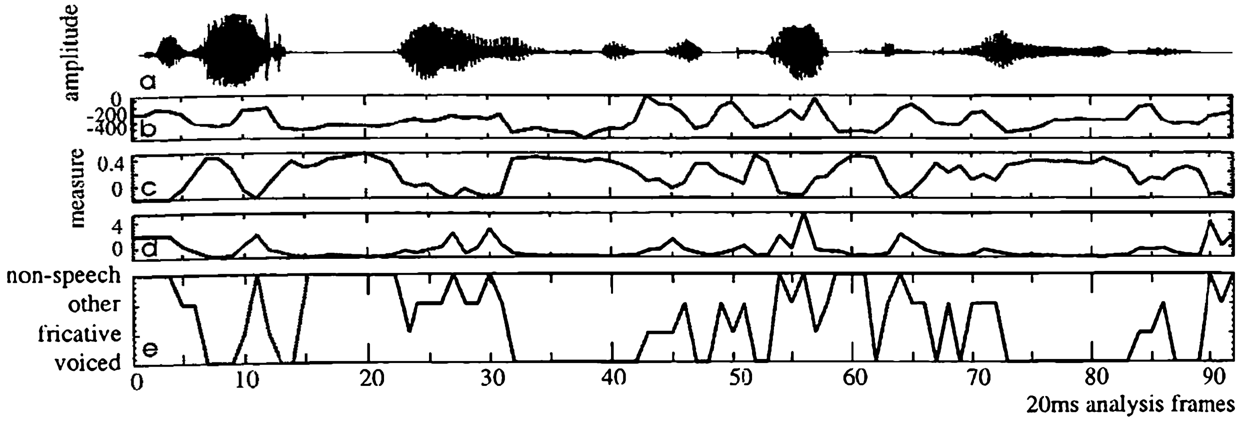


Figure A4.5: (a) a typical speech waveform shown with (b) LSP, (c) pitch strength and (d) gain measure. (e) shows the results of speech classification performed using those measures.

FigA4.5 is an example of the plot uscd to fine-tune decision regions. Straightforward thresholds applied to individual measures may be shown by drawing horizontal lines on the graphs at the appropriate value. Adjustments are made to the classification rules based upon any mismatch between the waveform contents (known through listening or through an existing phonetic transcription) and the automatic classification output.

Appendix 5: Intelligibility

A5.1 Speech understanding

Speech understanding is a complex issue, involving hearing, speech context, importance, audibility and perception. Some of these issues are considered in the following subsections.

A5.1.1 Measurements of intelligibility

In order to determine and compare the ways in which various factors influence the understanding of speech, it is necessary to create a standard measurement. In fact two main standards and many variations of them exist. An *intelligibility* test measures the ability of listeners to correctly identify words, phrases or sentences, whereas an *articulation* test measures the ability of listeners to correctly identify individual phonemes (vowels and consonants in monosyllabic or polysyllabic real or artificial words as described in [55]).

A5.1.2 Contextual information, redundancy and vocabulary size

Everyday experience indicates that contextual information plays an important role in the understanding of speech, often compensating for an extreme lack of original information. For example the sentence "He likes to xxxx brandy", can easily be understood even though a complete word is missing ("drink").

The construction of sentences is such that the importance of missing words is almost impossible to predict. The missing word "stop" differs in both importance and predictability in the two sentences "She waited in a queue at the bus xxxx" and "As the car sped towards him he shouted 'xxxx'!".

Contextual information may be regarded as being provided by surrounding words which constrain the possibilities of the enclosed word through grammatical rules or subject matter, or on a smaller scale as the surrounding syllables may constrain the choice of a missing syllable (as certain combinations do not appear, or are not common in the English language). Vocabulary size reduction also causes a similar constraint. It is natural for humans to reduce vocabulary size when communications is impaired, eloquence is not common under noisy conditions.

Redundancy, in which information is imparted in more ways than would normally be necessary, has effects similar to contextual constraint. Redundancy may be provided in the form of over-complex sentences in which the role of each single word is limited, by context, to a very small set of choices. In this way the extra information given in a sentence should enable complete understanding even if some words were lost. A simple example would be the repetition of important words, such as saying "It is a green car. The green car is turning ...." instead of the more concise "There is a green car turning". In the latter sentence, the two pieces of information "green" and "car" are spoken only once. Obviously, redundancy will improve communications effectiveness by at least 3dB.

Redundancy may also be achieved by the use of longer phrases of words of description, such as the use of "alpha bravo foxtrot" instead of "ABF". This reduces the chance that a short distortion would obliterate the transmission, or cause confusion to an entire quantum of information.

A measurement of the effects of contextual information on understanding is extremely difficult to quantify and is highly subjective. It is also entirely dependent upon the testing method, and the method of information removal necessary to carry out the tests.

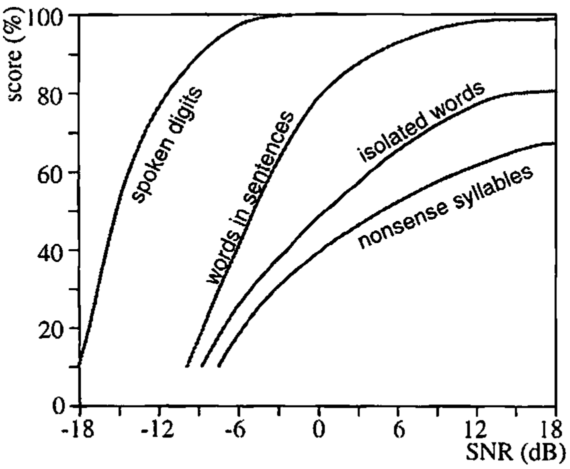


Figure A5.1: Effect of contextual information on speech. Constructed from examination of a figure presented in (112).

However some comparisons may be useful as shown in figA5.1, which plots the percentage of correctly identified digits, syllables or words spoken in the presence of the given degree of background noise [110]. Although the shorter digit words are relatively more likely to be corrupted by noise than the predominantly longer unconstrained words, the extremely limited set of possible choices involved (as the listeners knew they were only listening for digits) means that even a corrupted digit may be guessed with some accuracy.

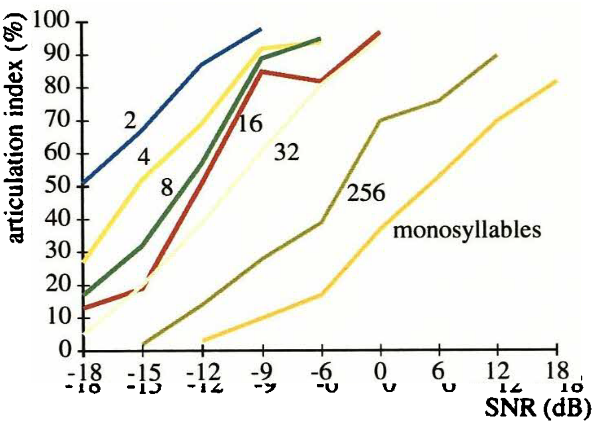


Figure A5.2: Effect of vocabulary size on intelligibility, Constructed from tabular data reported (112), and similar to a graph in (54).

FigA5.2 indicates the effects of changing the size of the speech vocabulary (where listeners are given the indicated number of choices) on intelligibility. This shows the very large improvement in recognition when vocabulary size is constrained either artificially or by context, for example reducing vocabulary size from 256 to 16 at -9 dB signal to noise level results in almost four times as many words being recognised. It should be remembered that the articulation index is a measure of the recognition rate of individual phonemes, not words.

This section has demonstrated the importance of context and redundancy in speech communication systems, and why these must thus be regarded as major factors in the effectiveness of such systems.

A5.1.3 Background noise

The effects of background noise play an important part in speech communication, but it is as important to consider the type of noise present as it is to consider its amplitude.

FigsA5.1 and A5.2 illustrate the effects of signal-to-noise level on speech understanding in the presence of evenly-distributed noise, where even quite small changes in this ratio may cause large changes in the degree of intelligibility.

Cases where standard A-weighted signal-to-noise ratio measurements are ineffective are relatively common. In motor vehicles, noise measured in dBA within the vehicle cabin may reach 80dBA [127]. In contrast, noise at infrasonic frequencies (2 to 32Hz - not considered by the A-weighted measure) can easily reach 120dB. Apart from causing a considerable masking effect on lower speech frequencies, this amplitude may cause hearing damage.

A second case where dBA measures do not apply is where audio frequencies are not actually present, but are perceived as residuals by being induced from higher frequency harmonics. The most famous example of this is related in [127], where householders near New York's JFK airport who complained about excessive rumble from aeroplanes prompted tests for noise pollution. The reported findings indicated that very little low-frequency noise was present, however there were several high frequency harmonics which induced the perception of rumble that householders complained about [127].

In general, for effective communication, around 6dB of signal should be present above the noise floor. Special consideration should be given to the formant frequency regions in speech [27], if not for the entire important frequency range of 800Hz to 3kHz, within which they predominantly lie.

Appendix 6: Test results

A6.1 LSP effectiveness tests

The results of the LSP effectiveness tests reported in section 9.3 are as follows:

			person:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
				4/12/96	4/12/96	4/12/96	5/12/96	5/12/96	6/12/96	6/12/96	6/12/96	16/1/97	14/1/97	14/1/97	16/1/96	22/1/97	3/12/96
Number	Vowel	Enhancement	Volume	R.G.	N.L.M.	H.Z.J.	R.O.J.	D.C.	S.X.Q.	O.E.P.	K.K.K.	J.P.D	J.A.Z.	J.O.	K.Y.M.	M.S.	I.V.M.
1	o	[illegible]	[illegible]	a	x	x	x	a	a	a	a	a	o	o	o	x	a
2	o	2	1.3	o	a	x	o	a	o	x	a	o	o	x	a	o	a
3	a	3	1.4	o	a	x	a	o	a	x	a	a	a	o	o	o	o
4	o	2	1.4	o	x	x	o	a	o	o	a	a	o	x	a	o	a
5	o	[illegible]	1.4	o	x	x	o	o	a	a	a	a	o	a	o	x	a
6	a	2	[illegible]	o	o	x	o	o	o	a	a	o	a	a	a	o	a
7	a	2	[illegible]	o	x	x	a	a	o	o	x	o	o	x	o	o	a
8	o	1	[illegible]	o	x	x	o	a	a	a	a	a	a	a	o	a	a
9	a	[illegible]	[illegible]	o	x	x	o	o	a	o	a	o	o	a	o	o	o
10	[illegible]	[illegible]	[illegible]	o	o	x	a	a	o	a	a	o	o	a	o	o	a
11	a	2	1.4	o	a	o	o	o	a	a	a	o	o	a	a	a	a
12	a	3	1.2	o	a	a	a	a	o	a	a	o	o	a	a	o	a
13	o	2	1.1	o	x	x	o	o	a	a	o	o	o	a	a	o	a
14	o	3	1.3	o	o	o	o	a	a	a	a	o	a	a	a	o	o
15	a	3	1.2	o	o	o	a	a	a	x	a	o	a	a	o	o	o
16	o	3	1.1	o	o	o	o	a	a	a	o	o	o	a	o	o	o
17	o	2	1.4	o	o	a	a	a	a	a	a	o	a	o	o	a	a
18	a	2	1.3	o	o	o	o	a	a	a	a	a	a	a	o	x	a
19	a	1	1.1	o	x	o	a	o	a	o	a	o	o	a	o	o	a
20	o	2	1.2	a	o	a	o	o	a	a	a	o	o	x	o	n	o
21	a	2	1.3	a	x	o	o	a	o	a	a	o	o	a	o	o	a
22	a	2	1.2	o	a	a	o	a	a	o	a	o	o	a	o	a	a
23	a	1	1.4	o	o	a	o	a	a	a	a	a	a	o	o	o	a
24	o	2	1.4	o	x	o	o	a	a	a	a	o	a	a	o	x	o
25	a	2	1.1	a	o	a	o	o	a	a	a	o	o	a	a	a	a
26	o	1	1.3	o	x	o	o	a	a	o	a	o	o	a	a	o	a
27	o	3	1.1	u	o	o	o	a	o	o	o	o	a	a	o	o	o
28	o	2	1.1	o	o	a	o	a	o	a	o	o	a	a	o	x	o
29	o	1	1.1	a	o	x	o	a	a	a	a	o	o	a	o	o	o
30	a	1	1.1	o	o	o	o	a	a	o	a	o	a	o	o	a	a
31	o	3	1.2	a	a	a	o	a	o	a	o	o	a	a	a	o	o
32	o	2	1.2	o	a	a	o	a	a	a	o	o	o	a	a	x	o
33	a	2	1.2	o	a	a	o	a	o	a	o	o	o	a	o	a	a
34	o	3	1.2	o	x	o	o	a	a	a	o	o	o	o	a	o	o
35	o	1	1.2	o	o	a	o	o	o	a	a	o	a	a	a	x	a
36	a	3	1.2	a	o	o	o	a	o	a	o	o	a	a	a	o	a
37	a	3	1.3	a	a	a	o	a	a	a	a	o	o	o	o	o	a
38	o	3	1.3	o	o	o	o	a	a	a	o	o	o	a	a	o	o
39	a	3	1.1	a	o	a	o	a	a	a	a	o	a	a	o	o	a
40	a	1	1.3	o	o	a	x	o	a	a	o	a	o	a	a	o	a
41	o	2	1.2	o	a	a	o	a	a	a	o	o	a	a	o	o	a
42	a	2	1.1	a	a	a	o	a	o	a	a	o	a	a	a	x	a
43	a	1	1.4	a	a	a	x	a	o	o	a	o	o	a	o	x	a
44	a	3	1.1	a	a	a	o	a	a	a	o	o	o	a	a	o	a
45	a	1	1.3	o	a	a	o	a	a	a	a	o	a	x	o	o	a
46	o	1	1.1	a	x	a	o	a	o	a	a	o	a	o	a	o	o
47	a	1	1.2	a	a	a	o	o	o	a	a	a	o	o	o	o	a
48	o	1	1.2	a	o	a	o	a	o	a	a	o	o	a	o	x	a
49	a	2	1.4	a	o	a	o	a	a	o	a	o	a	a	a	o	a
50	a	3	1.4	a	o	o	o	o	a	a	a	a	a	a	o	o	a
51	o	3	1.4	a	x	a	o	a	o	a	a	o	o	a	o	o	o
52	o	1	1.1	o	a	a	o	a	a	a	a	o	a	a	a	a	a
53	o	2	1.1	o	o	a	o	a	o	o	a	o	o	a	o	o	o
54	o	1	1.2	a	x	a	o	a	o	o	o	o	o	a	o	o	o
55	o	2	1.3	o	a	a	a	a	o	a	o	o	a	a	o	o	o
56	o	3	1.4	o	o	a	o	a	o	a	o	o	o	a	a	o	o
57	a	1	1.2	a	o	a	o	o	a	o	o	o	o	o	o	o	a
58	a	1	1.1	a	o	a	o	a	a	o	a	a	o	o	a	o	a
59	o	1	1.4	o	o	a	o	a	a	a	o	o	a	o	o	o	o
60	a	3	1.3	o	o	a	o	a	a	o	a	o	o	a	a	x	a
61	o	3	1.3	a	o	a	o	a	a	a	a	o	o	a	o	a	o
62	o	2	1.3	a	o	o	a	a	o	a	a	o	o	a	a	o	o
63	a	1	1.4	o	o	x	o	a	a	o	x	o	o	x	o	o	a
64	a	3	1.4	o	o	a	x	o	o	a	a	o	o	a	a	a	a
65	a	1	1.2	a	x	a	a	o	a	o	a	o	o	a	a	o	a
66	o	3	1.2	o	x	a	o	o	o	a	o	o	o	a	o	o	o
67	a	3	1.1	o	a	a	o	o	o	o	o	a	o	a	a	o	a
68	o	1	1.3	a	a	x	a	o	a	a	a	o	a	o	o	a	o
69	o	3	1.4	o	x	a	o	a	a	o	o	o	o	a	o	a	o
70	a	2	1.1	o	o	a	o	a	a	o	o	o	o	a	a	o	a
71	a	3	1.3	a	a	a	x	a	o	o	a	o	a	a	o	o	a
72	o	3	1.1	o	o	o	o	o	a	o	o	a	o	o	o	o	o
73	o	1	1.3	a	a	a	x	o	a	a	a	o	a	o	o	a	o
74	o	2	1.3	o	o	a	o	a	o	o	o	o	a	o	o	o	o
75	a	3	1.4	o	x	a	o	a	o	a	o	o	o	o	o	o	a
76	o	2	1.4	o	o	a	o	a	a	a	a	o	o	o	o	o	u
77	o	1	1.4	o	a	x	o	o	a	a	a	o	a	a	o	a	a
78	a	2	1.4	a	o	a	o	a	a	o	o	o	o	a	o	o	a
79	a	2	1.3	o	x	a	o	a	o	a	a	o	a	a	o	o	o
80	o	1	1.4	o	a	a	a	a	a	a	a	o	a	a	o	a	a
81	a	2	1.2	o	o	a	o	a	a	a	o	o	o	o	o	o	a
82	a	1	1.3	a	a	a	a	o	o	o	a	a	a	a	a	a	a

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rates of recognition	raw:	54%	33%	42%	50%	42%	58%	25%	46%	46%	33%	38%	46%	33%	75%
	shifted:	63%	46%	46%	46%	50%	54%	42%	58%	54%	42%	58%	63%	50%	88%
	widen:	63%	50%	58%	58%	50%	46%	50%	71%	58%	63%	50%	54%	50%	100%
	correct:	43	31	35	37	34	38	28	42	38	33	35	39	32	63
Improvements	shifted:	1.15	1.38	1.10	0.92	1.20	0.93	1.67	1.27	1.18	1.25	1.56	1.36	1.50	1.17
	widened:	1.15	1.50	1.40	1.17	1.20	0.79	2.00	1.55	1.27	1.88	1.33	1.18	1.50	1.33

Averages

	rate	factor
correct:	52.38%	
raw:	44.35%	
shifted:	54.17%	1.26
widened:	58.63%	1.37

sample sizes

total:	1148
each enh:	383
subjects:	14

Table A6.1: Results of intelligibility testing for LSP processed vowel sounds listened to in simulated vehicle noise.

A6.2 Formant sharpening/broadening

Table A6.2 gives the raw intelligibility results obtained for each listening trial in a number of tests designed to asses the effects of various degrees of LSP widening and narrowing.

condition:	1	2	3	4	5	6	7	8	9	10	11
% widened	400	270	240	210	180	150	120	90	60	30	0
trial 1	100	100	100	100	100	83.3	66.7	100	66.7	83.3	66.7
trial 2	66.7	83.3	83.3	50	100	83.3	100	100	66.7	83.3	66.7
trial 3	83.3	100	83.3	83.3	83.3	83.3	83.3	50	66.7	50	66.7
trial 4	100	100	66.7	50	66.7	100	66.7	83.3	83.3	83.3	83.3
trial 5	83.3	83.3	83.3	100	100	83.3	66.7	83.3	83.3	66.7	50
trial 6	66.7	50	83.3	83.3	66.7	50	83.3	33.3	66.7	50	33.3
trial 7	100	100	100	83.3	83.3	83.3	66.7	66.7	83.3	83.3	50
trial 8	83.3	66.7	100	83.3	83.3	66.7	66.7	66.7	50	33.3	33.3
trial 9	66.7	83.3	83.3	100	83.3	83.3	83.3	100	100	83.3	100
trial 10	66.7	83.3	83.3	83.3	50	66.7	66.7	83.3	50	50	66.7
trial 11	100	83.3	83.3	66.7	83.3	83.3	83.3	83.3	50	50	50
trial 12	100	100	100	100	83.3	100	83.3	66.7	100	50	33.3
trial 13	100	66.7	100	83.3	83.3	100	83.3	50	66.7	50	83.3
trial 14	83.3	100	83.3	66.7	83.3	83.3	66.7	66.7	83.3	83.3	66.7
Average	85.7	85.7	88.1	80.9	82.1	82.1	76.2	73.8	72.6	64.3	60.7

Table A6.2: Individual and average correct response rates for 14 listening trials tested against 11 degrees of formant widened and narrowed vowels in added noise.

A6.3 DRT test results

The DRT test results for 20 trials hearing both enhanced and unenhanced words in two types of noise are given in table A6.3. Listeners are identified through the filename of their results, and values are percentages of the words correctly identified in the given categories. In other words, no correction has yet been made for guessing answers.

result listener	overall score *(% correct)*	**simulated interior noise**		**simulated siren noise**	
		unenhanced	*enhanced*	*unenhanced*	*enhanced*
aislam	65	55	69	65	71
ben	66	54	69	66	77
dave	68	59	65	70	78
daveh	67	71	61	66	69
derekc	65	63	64	62	71
frank	61	48	62	66	67
hardw	60	59	56	61	63
harp	62	53	61	64	71
kirk	57	55	55	59	61
klaus	68	59	76	70	66
krishna	62	60	61	61	67
ong	60	50	61	62	67
robg	66	56	67	64	79
robj	65	58	65	62	76
salousm	64	48	66	71	70
sandhu	67	56	66	66	80
thai	63	50	67	62	73
temple	71	60	71	72	82
terry	67	59	66	67	75
zentani	59	50	63	62	63
average	64	56	65	65	68

Table A6.3: Listeners (identified by result filename), overall correct response average and response averages for unenhanced and enhanced speech in both types of noise. Boxed results indicate an intelligibility reductions caused through processing.

A6.4 DRT test contents

Table A6.4 shows the consonant taxonomy used in construction of the DRT test.

Features	m	n	v	ð	z	ʒ	ʒ̂	b	d	g	w	r	l	j	f	θ	s	ʃ	ʃ	p	t	k	h
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
Nasality	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sustention	-	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+
Sibilation	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
Graveness	+	-	+	-	-	O	O	+	-	O	+	-	O	O	+	-	-	O	O	+	-	O	O
Compactness	-	-	-	-	-	+	+	-	-	+	-	-	O	+	-	-	-	+	+	-	-	+	+

Table A6.4: Consonants used in the DRT test, and their relation to the six word feature categories (from (125)). + indicates present, - absent and O inapplicable.

Table A6.5 gives the 96 DRT stimulus words in feature categories. Within each feature list, words on the left are positive within that category, and words on the right are negative [3]. Note that the word 'calf' may rhyme with 'gaff' when spoken by the American English speakers for whom the DRT test was designed, but does not necessarily rhyme for native English speakers, so this word is replaced by 'caff' as is usual when speakers are British [77].

Voicing		**Nasality**		**Sustention**	
voiced	*unvoiced*	*naral*	*oral*	*sustained*	*interrupted*
veal	feal	meat	heat	vee	bee
bean	peen	need	deed	sheen	cheat
gin	chin	mitt	bit	vill	bill
dint	tint	nip	dip	thick	tick
zoo	sue	moot	boot	foo	pooh
dune	tune	news	dues	shoes	choose
voal	foal	moan	bone	those	doze
goat	coat	note	dote	though	dough
zed	said	mend	bend	then	den
dense	tense	neck	deck	fence	pence
vast	fast	mad	bad	than	dan
gaff	calf [gaff]	nab	dab	shad	chad
vault	fault	moss	boss	thong	tong
daunt	taunt	gnaw	daw	shaw	chaw
jock	chock	mom	bomb	von	bon
bond	pond	knock	dock	vox	box
Sibilation		**Graveness**		**Compactness**	
sibilated	*unsibilated*	*grave*	*acute*	*compact*	*diffuse*
zee	thee	weed	reed	yield	wield
cheep	keep	peak	teak	key	tea
jilt	gilt	bid	did	hit	fit
sing	thing	fin	thin	gill	dill
juice	goose	moon	noon	coop	poop
chew	coo	pool	tool	you	rue
joe	go	bowl	dole	ghost	boast
sole	thole	fore	thor	show	so
jest	guest	met	net	keg	peg
chair	care	pent	tent	yen	wren
jab	gab	bank	dank	gat	bat
sank	thank	fad	thad	shag	sag
jaws	gauze	fought	thought	yawl	wall
saw	thaw	bong	dong	caught	taught
jot	got	wad	rod	hop	top
chop	cop	pot	tot	got	dot

Table A6.5: The 96 alternative words comprising the DRT test arranged by feature difference.

A6.5 LSP measure evaluation tests

Spearman ranking correlation values were obtained for different speech measures for a large number of phoneme sounds with differing levels of added white noise. Table A6.6 shows the noise conditions tested:

Noise condition:	0	1	2	3	4	5
Noise amplitude scaling:	0%FSD	5%FSD	10%FSD	20%FSD	40%FSD	80%FSD
Relative noise amplitude:	no noise	9dB$_{SNR}$	6dB$_{SNR}$	3dB$_{SNR}$	0dB$_{SNR}$	-3dB$_{SNR}$

Table A6.6: Noise conditions for Spearman rank calculations, with the corresponding noise amplitude in percentage of full scale value, and the noise amplitude relative to speech amplitude (fixed at 40%FSD).

The Spearman sample coefficient for each of the tested conditions is shown in table A6.7, the measures correlated were LSP measure, zero-crossing rate, frame power and average magnitude difference function (referred to as LSP, ZCR, POW and AMDF respectively).

For each phoneme, in each type of noise, arrays were created containing the average value of each of the measures. In order to calculate the Spearman coefficient, r_S, from two measure arrays m_1 and m_2, the arrays were first ranked as in eqnsA6.1 and A6.2:

$$k_1[i] = rank(m_1[i]) \qquad (A6.1)$$

$$k_2[i] = rank(m_2[i]) \qquad (A6.2)$$

$$for\, i = 1 \ldots (length(m_n) = n)$$

where the length of m, n, is the number of separate phonemes (here 58) and m_1 and m_2 are the two measures being correlated.

The Spearman coefficient for the n phonemes is then:

$$r_S = 1 - \frac{6\sum_{i=1}^{n}\{k_1[i] - k_2[i]\}^2}{n(n^2 - 1)} \qquad (A6.3)$$

A large absolute value of coefficient indicates strong correlation, with the sign indicating proportionality or reverse proportionality. For a large sample size (over 20), it may be assumed that the results are normally distributed, and actual significance, if required may be estimated from normal variate values [29].

	Noise 0		
	LSP	**ZCR**	**POW**
LSP	□	0.8098	-0.6434
POW	-0.6434	-0.545	□
AMDF	-0.6548	-0.5719	0.9919
	Noise 1		
	LSP	**ZCR**	**POW**
LSP	□	0.9257	-0.9232
POW	-0.9232	-0.8592	□
AMDF	-0.9305	--0.8472	0.9951
	Noise 2		
	LSP	**ZCR**	**POW**
LSP	□	0.789	-0.9247
POW	-0.9247	-0.7133	□
AMDF	-0.9057	-0.6497	0.9859
	Noise 3		
	LSP	**ZCR**	**POW**
LSP	□	0.626	-0.863
POW	-0.863	-0.4686	□
AMDF	-0.7387	-0.2843	0.9419
	Noise 4		
	LSP	**ZCR**	**POW**
LSP	□	0.3373	-0.6077
POW	-0.6077	-0.0554	□
AMDF	-0.422	0.3337	0.8702
	Noise 5		
	LSP	**ZCR**	**POW**
LSP	□	0.1339	-0.2114
POW	-0.2114	0.5433	□
AMDF	-00.0583	0.7009	0.9099

Table A6.7: Spearman ranking coefficients for the tested conditions.

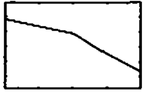
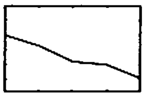
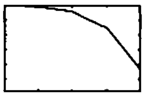
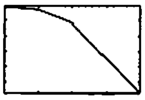
	Noise 1	Noise 2	Noise 3	Noise 4	Noise 5	Deterioration
LSP	0.7667	0.6734	0.5804	0.3043	0.0729	
ZCR	0.5938	0.4598	0.2454	0.1989	0.018	
POW	0.9929	0.9772	0.9186	0.6899	0.1426	
AMDF	0.9825	0.9408	0.7728	0.3024	-0.1578	

Table A6.8: Spearman ranking coefficients for noise degraded measures correlated with respect to the measures for noise-free speech, and a plot of the deterioration in each measure (x-axis is noise level, y-axis is Spearman coefficient, as figA6.1).

The deterioration plots in table A6.8 show the effect of increasing noise level on the performance of each measure with respect to its performance for noise condition 1. These plots are shown combined in figA6.1.

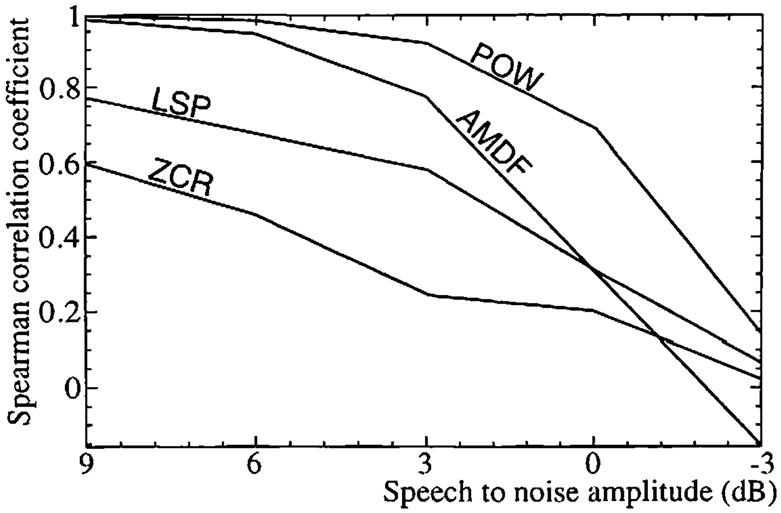


Figure A6.1: Spearman coefficient against speech to noise amplitude ratio, showing the deterioration in each measure as the proportion of noise is increased.

Subsections A6.5.1 to A6.5.3 plot the correlations for noise 0, noise 1 and noise 4 conditions respectively between various of the measures. In the plots, the degree of correlation between the measures can be seen as the resemblance of the distribution to a straight line at 45° for proportionality or 135° for inverse proportionality.

Each point on the distribution graphs corresponds to a single phoneme - located by its average measure values over all of the tested speech. The "sc" value in the headings is the absolute value of the Spearman ranking coefficient for the given plot and given conditions.

A6.5.1 Noise 0 correlation plots

In noise-free speech analysis, the LSP measure correlated well with the ZCR measure, as did the POW and the AMDF measure.

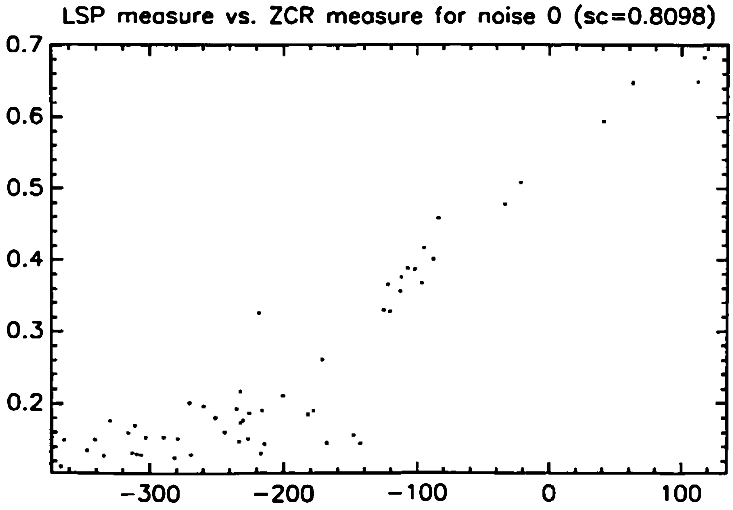


Figure A6.2: Correlation between LSP and ZCR measures.

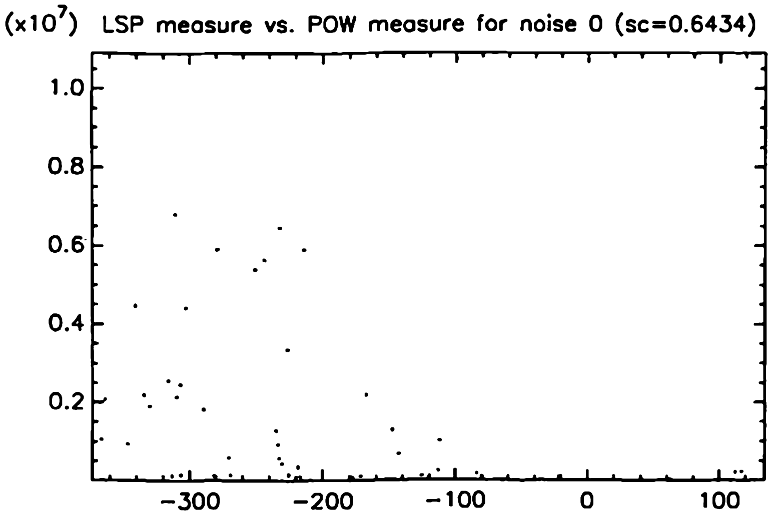


Figure A6.3: Correlation between LSP and POW measures.

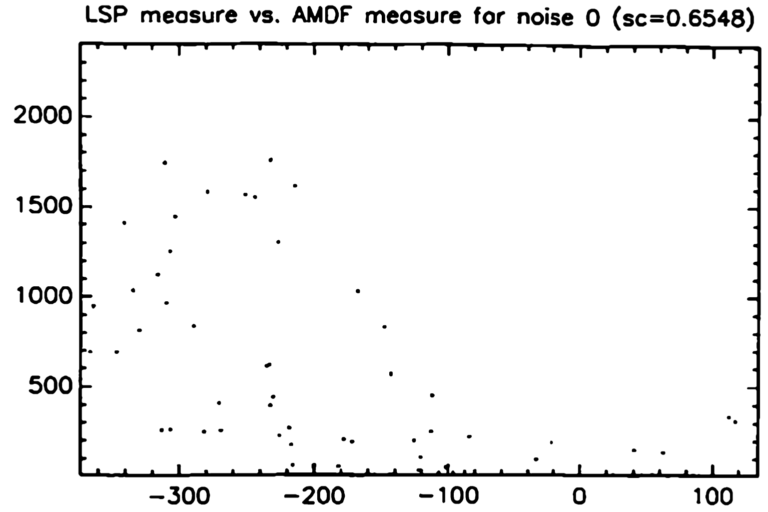


Figure A6.4: Correlation between LSP and AMDF measures.

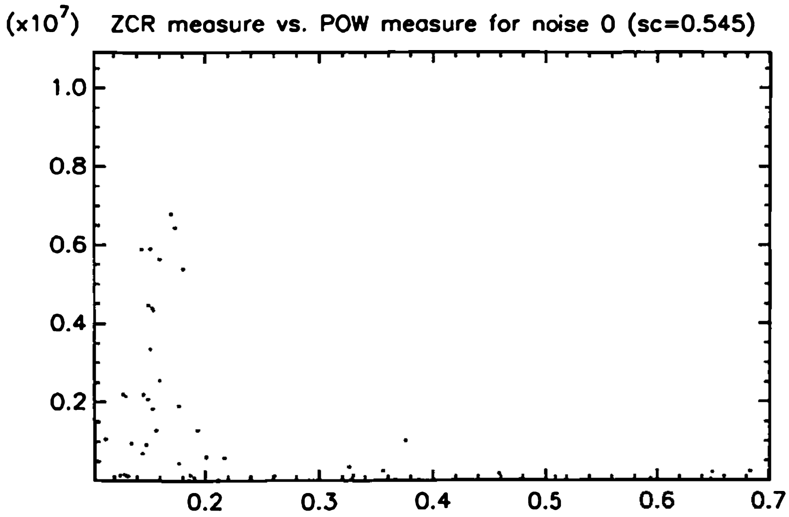


Figure A6.5: Correlation between ZCR and POW measures.

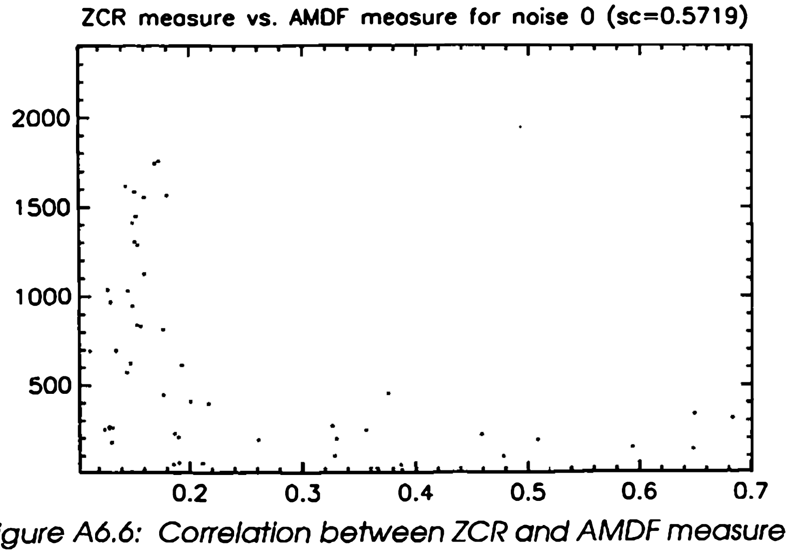


Figure A6.6: Correlation between ZCR and AMDF measures.

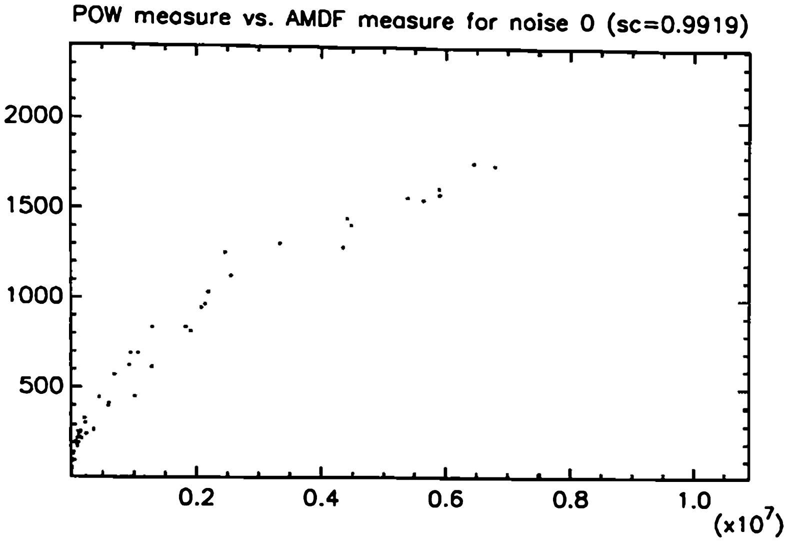


Figure A6.7: Correlation between POW and AMDF measures.

A6.5.2 Noise 1 correlation plots

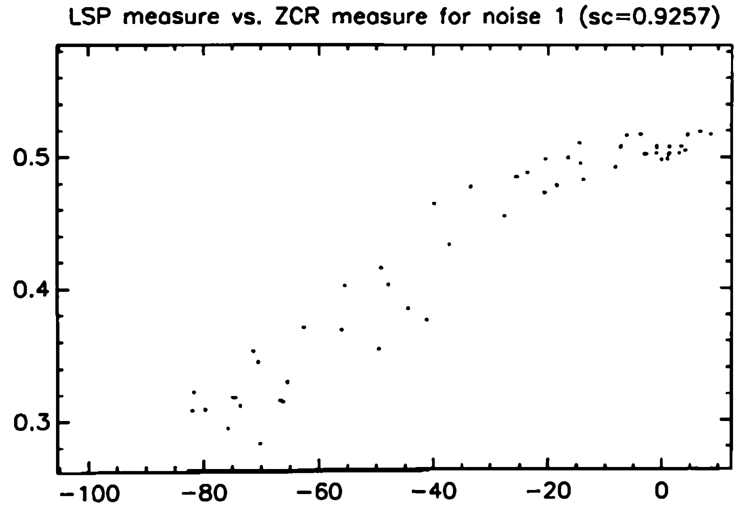


Figure A6.8: Correlation between LSP and ZCR measures.

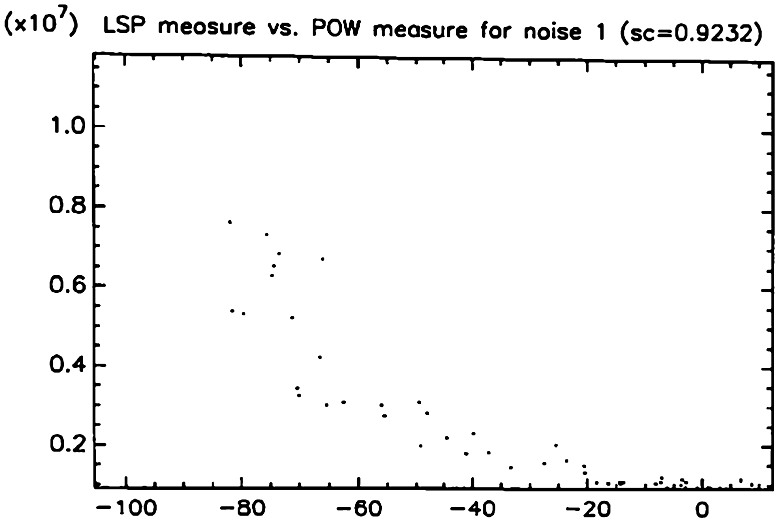


Figure A6.9: Correlation between LSP and POW measures.

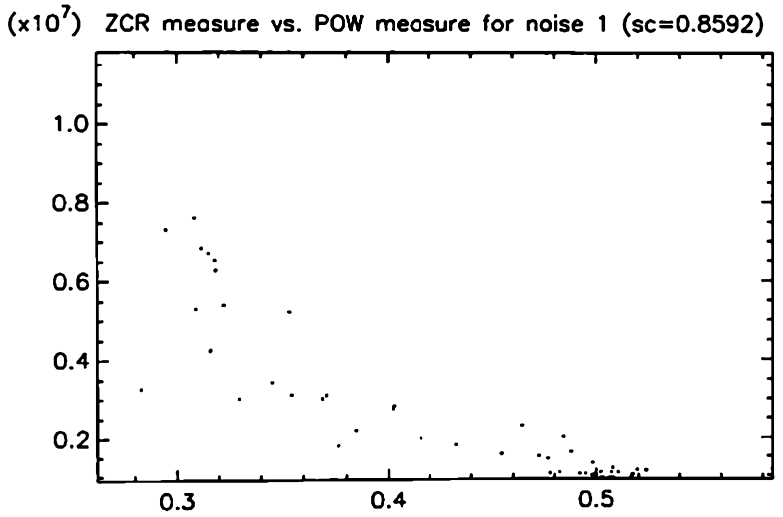


Figure A6.10: Correlation between ZCR and POW measures.

A6.5.3 Noise 4 correlation plots

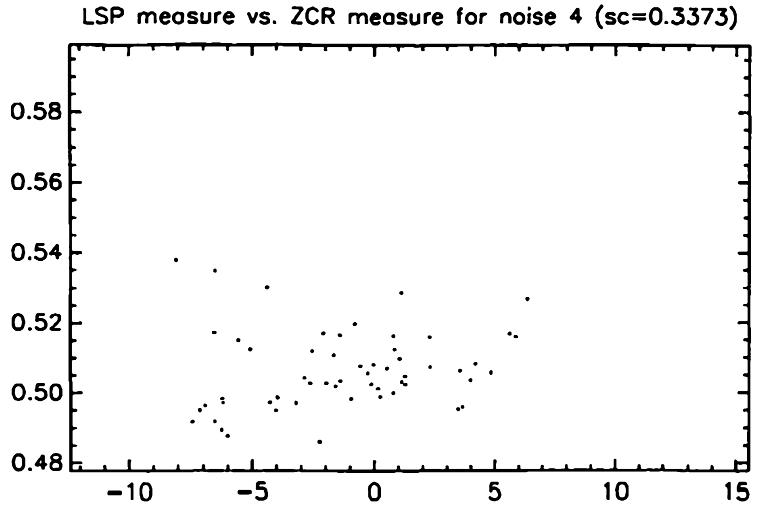


Figure A6.11: Correlation between LSP and ZCR measures.

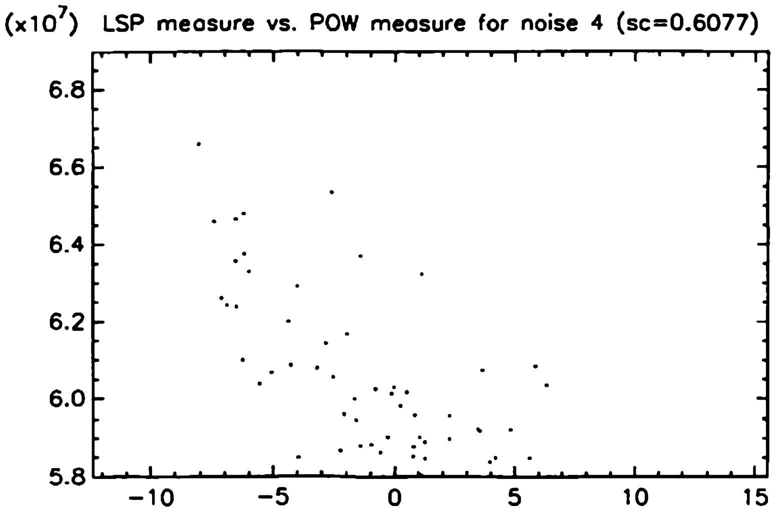


Figure A6.12: Correlation between LSP and POW measures.

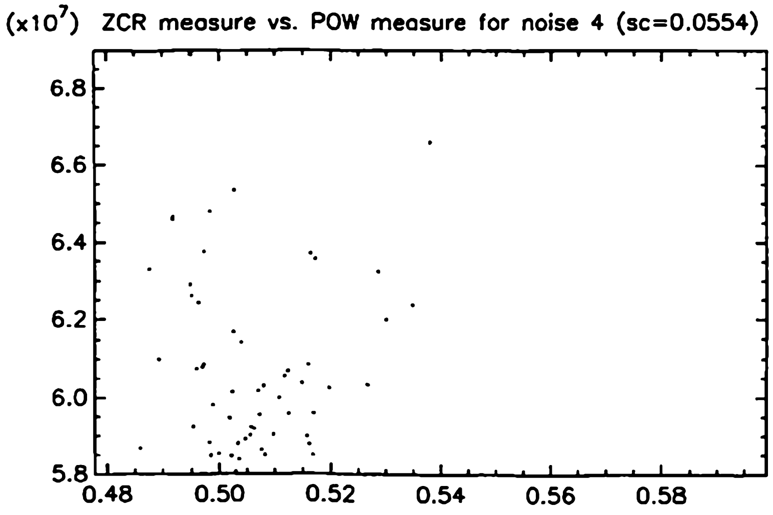


Figure A6.13: Correlation between ZCR and POW measures.

A6.5.4 Measure variance between noise-free and noise 1 condition

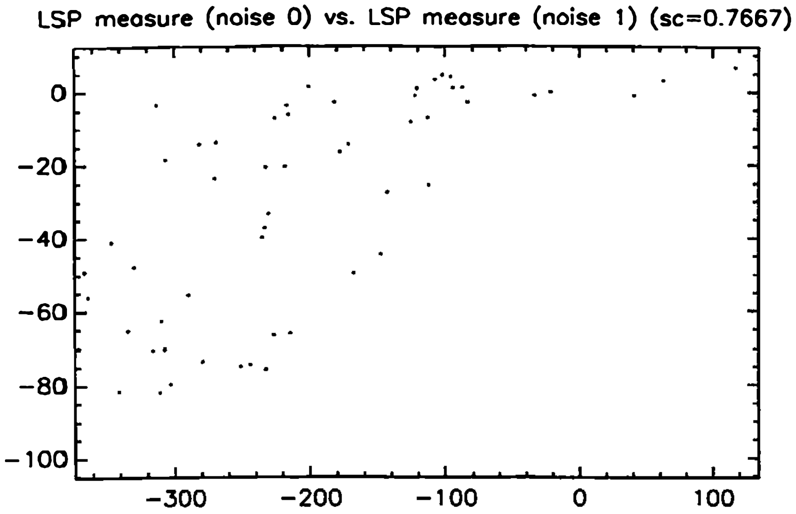


Figure A6.14: Correlation of LSP measure from noise-free speech with that from noise 1.

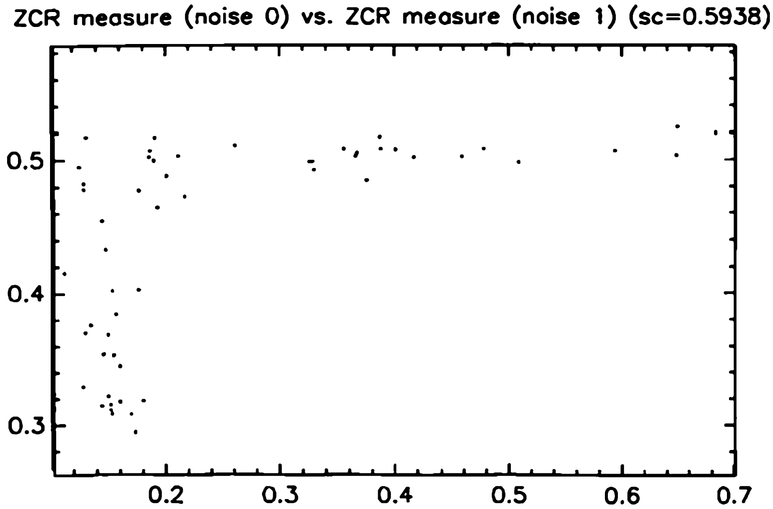


Figure A6.15: Correlation of ZCR measure from noise-free speech with that from noise 1.

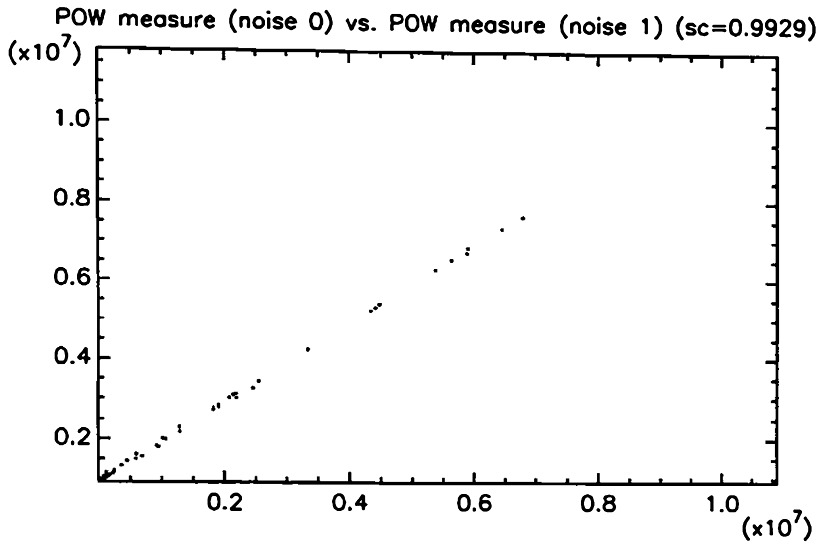


Figure A6.16: Correlation of POW measure from noise-free speech with that from noise 1.

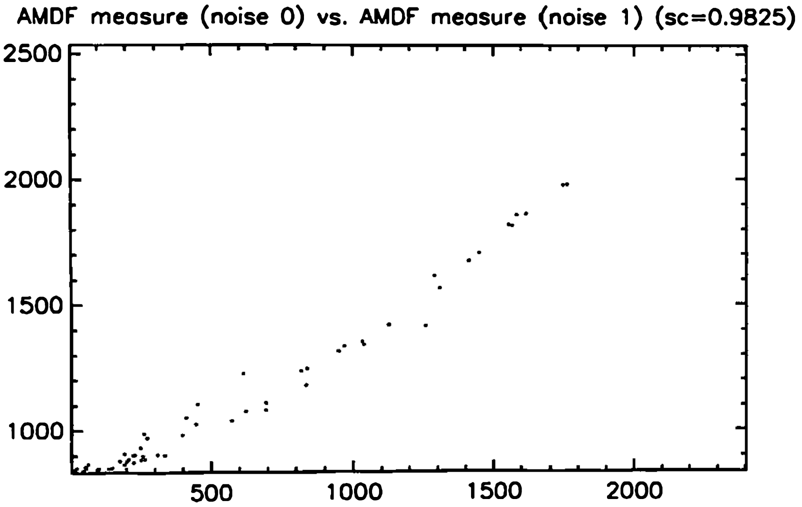


Figure A6.17: Correlation of AMDF measure from noise-free speech with that from noise 1.

A6.5.5 Measure variance between noise-free and noise 2 conditions

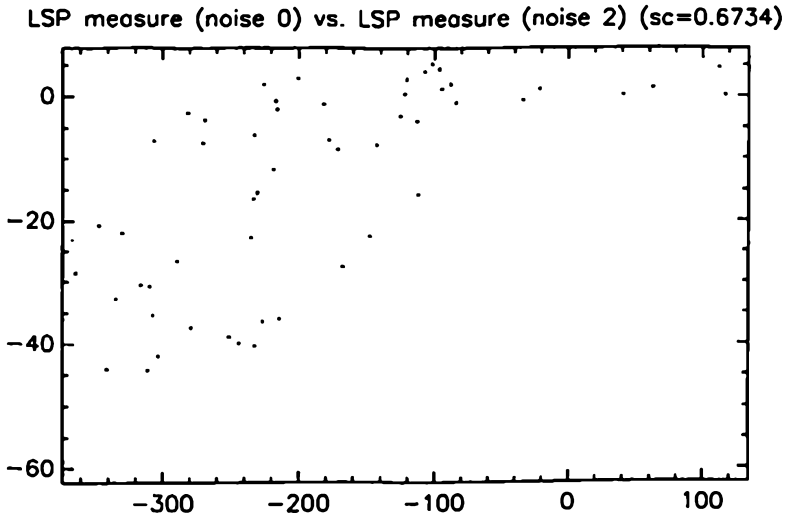


Figure A6.18: Correlation of LSP measure from noise-free speech with that from noise 2.

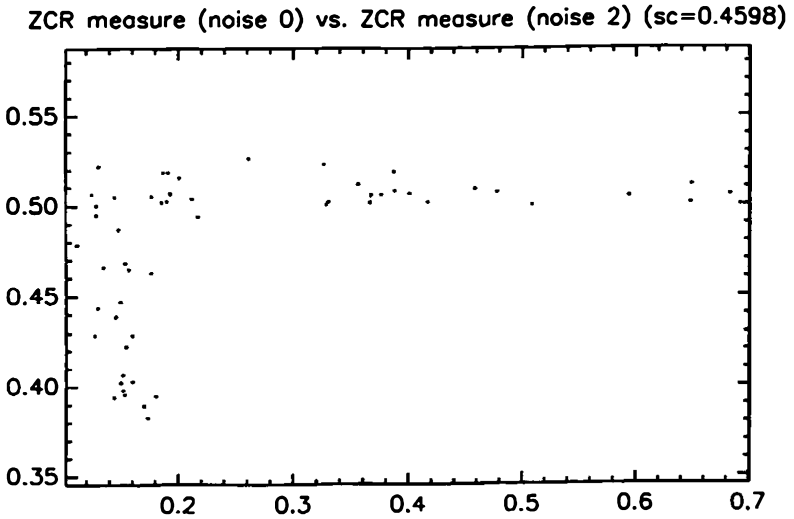


Figure A6.19: Correlation of ZCR measure from noise-free speech with that from noise 2.

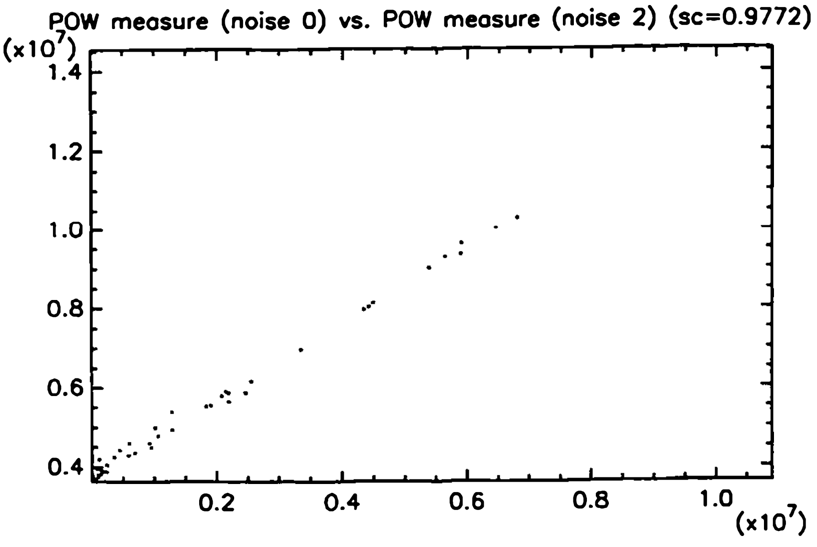


Figure A6.20: Correlation of POW measure from noise-free speech with that from noise 2.

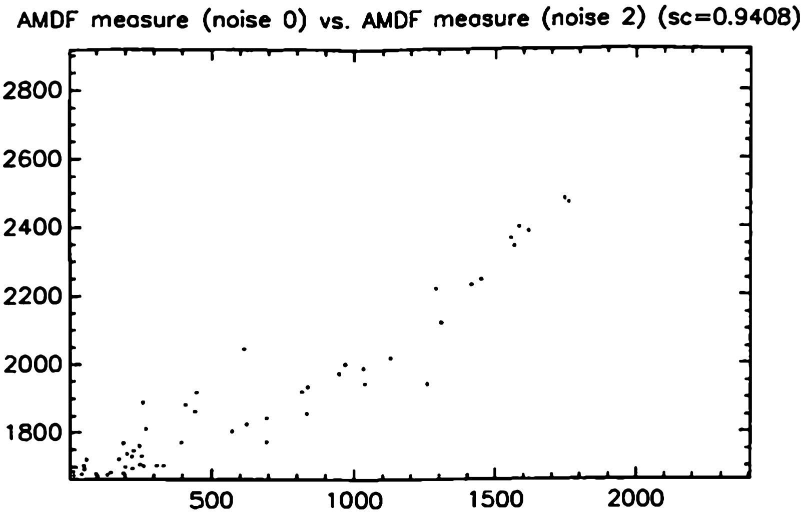


Figure A6.21: Correlation of AMDF measure from noise-free speech with that from noise 2.

A6.5.6 Measure variance between noise-free and noise 4 condition

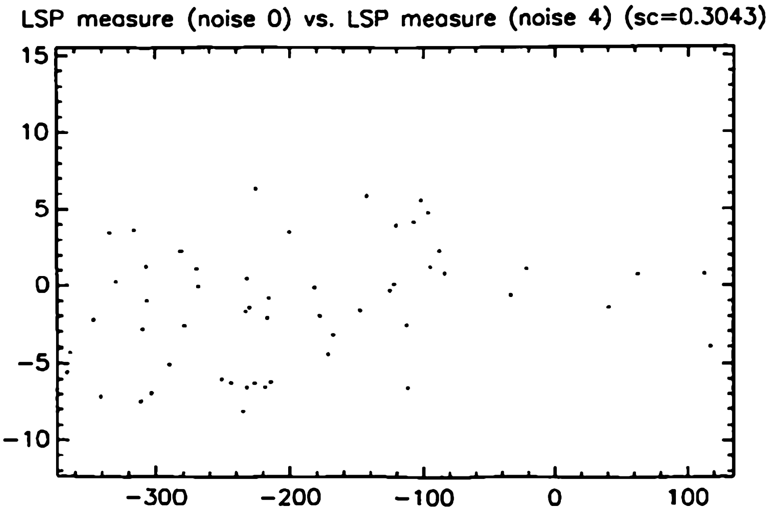


Figure A6.22: Correlation of LSP measure from noise-free speech with that from noise 4.

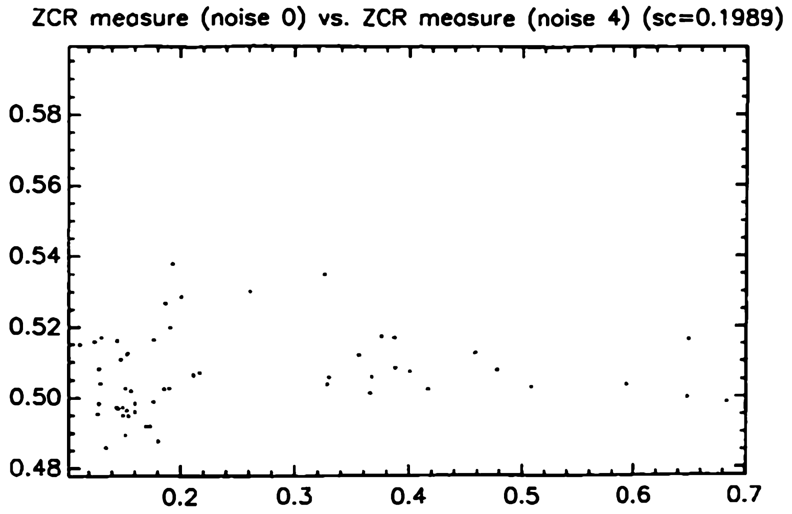


Figure A6.23: Correlation of ZCR measure from noise-free speech with that from noise 4.

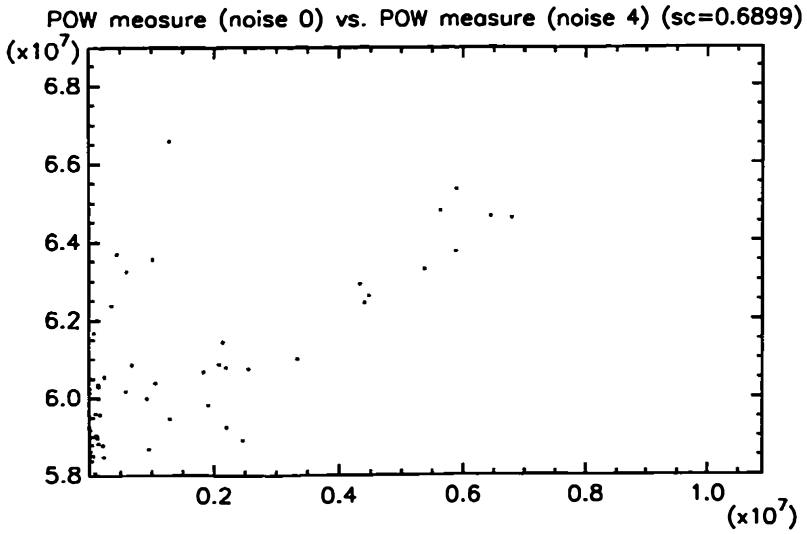


Figure A6.24: Correlation of POW measure from noise-free speech with that from noise 4.

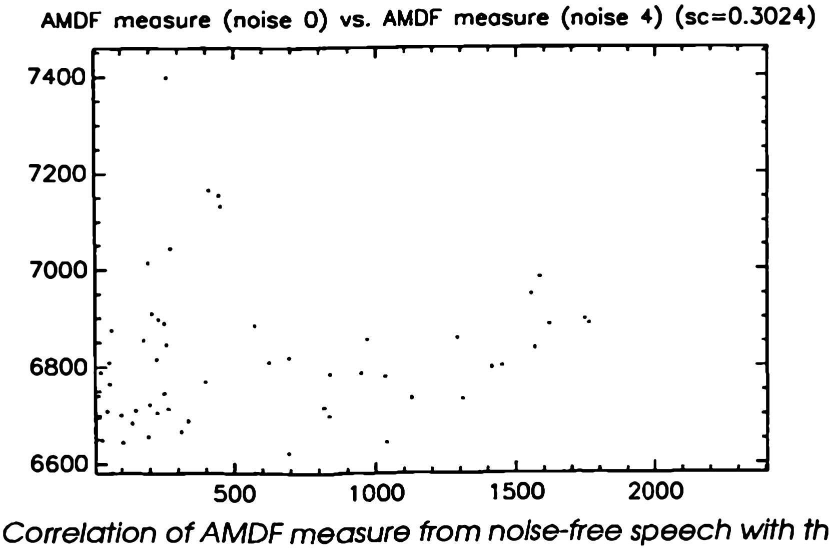


Figure A6.25: Correlation of AMDF measure from noise-free speech with that from noise 4.

A6.6 Extrapolation from DRT results to predict intelligibility

The DRT test results of section A6.3 indicated the average improvement in speech intelligibility obtained for all listeners in each of six different groups of speech features (section A6.4). Values have been found for both the tonal and wideband noise tests.

In order to construct a single composite intelligibility measure for the tests, it is necessary to firstly determine the relative frequency of the six feature classes in unconstrained speech, secondly to choose the relative frequency of tonal and wideband interfering noise, and thirdly to use these values as weights to the intelligibility increase value in each speech feature category.

The types of phonemes comprising each of the six speech feature categories are known (and are shown in table A6.4 in section A6.4), where the symbols given in the table are from the international phonetic alphabet.

Speech classification tests performed using the TIMIT database [118] compiled a cumulate count of the phonetic transcription of a number of recorded sentences of speech (actually for the results in section A6.5). The occurrences of each phoneme in this test can be grouped to determine the proportion of occurrences for each of the six DRT test speech feature classes.

Unlike the DRT test taxonomy of table A6.4, the TIMIT database uses a substantially different proprietary phonetic classification - requiring the construction of mappings between the two formants. The TIMIT mapping for each of the phonemes comprising the six speech feature categories, and the number of occurrences of each phoneme is shown in table A6.9, where the average number of occurrences of each feature class is calculated.

voicing		nasality		sustention		sibilation		graveness		compactness	
m	99	m	99	v	27	z	105.1	m	99	zh	0
n	117.7	n	117.7	dh	26.5	zh	0	v	27	g	18.8
v	27	ng	25.4	z	105.1	s	245.5	b	10.6	y	29.9
dh	26.5	**86.7**		zh	0	sh	92.3	w	46.6	sh	92.3
z	105.1			hh	25.3	jh	25.7	f	75.6	k	59
zh	0			w	46.6	ch	24.7	p	42.4	hh	25.3
b	10.6			r	127.7	**82.2**		**50.2**		**37.5**	
d	14.1			l	106.2						
g	18.8			y	29.9						
w	46.6			f	75.6						
r	127.7			th	21.7						
l	106.2			s	245.5						
y	29.9			sh	92.3						
jh	25.7			**71.5**							
ch	24.7										
52											

Table A6.9: DRT speech feature categories, the TIMIT transcribed phonemes comprising them, and the number of occurrences of each phoneme in the tested subset of the TIMIT database. Average number of occurrences have been calculated in each speech feature category.

The average number of occurrences of each speech feature category in the TIMIT sentences from table A6.9 were summed to provide a quotient used to determine the relative frequency of each category, shown in table A6.10.

voicing	nasality	sustention	sibilation	graveness	compactness
0.137	0.228	0.188	0.216	0.132	0.099

Table A6.10: Relative frequency of each speech feature category

A7.1 Speech enhancement additions to CELP

Figure A7.1 shows a block diagram of the detail within the vehicular CELP system. Note that the functions inside the dotted area are those additional to a standard CELP system for the purpose of speech enhancement.

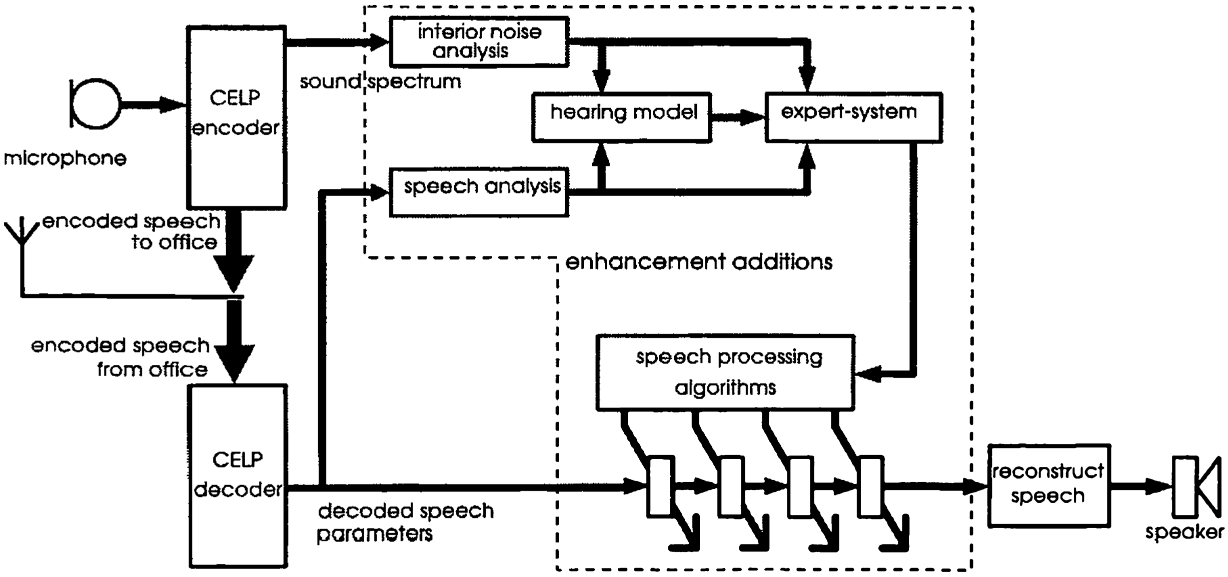


Figure A7.1: Speech enhancing CELP block diagram.

The various blocks presented in figure A7.1 are now explored in more detail, and described in terms of the required instructions per second (ips). The TETRA CELP coder operates on 240 sample window sizes at a sampling rate of 8kHz (and thus 33.3 30ms frames per second).

A7.1.1 Interior noise analysis

The enhancement system requires an analysis of the noise present in the environment of the listener in order to make ‘intelligent’ decisions regarding the type and degree of alteration to make to the speakers speech.

Noise analysis would generally consist of time domain elements to track the noise amplitude and frequency domain elements to determine the spectrum of the noise. In addition, a method is needed to determine whether the current noise parameters in fact contain the speech of the listener (in which case they would be discarded and the previous frames noise analysis used instead).

Rather than conduct explicit noise analysis, it is possible to utilise the powerful analysis capabilities of the existing CELP uplink (ie the CELP encoder located in the vehicle).

Amplitude tracking would consider the CELP gain parameter for each analysis frame, and construct a measure based on past frames, probably requiring one multiply-accumulate operation per frame, or 33 ips.

Speech detection would require a comparison of the CELP parameters for each frame, with 11 additions for LSP measure construction, and up to 4 comparisons for parameter interpretation, giving a total of 15*33.3 = 500 ips.

Finally, the noise spectrum must be derived for perceptual weighting and comparison with the speech spectrum. Thus would actually be derived from the raw LPC parameters (rather than the more complex route from LSP parameters), to yield a 200 element spectral array. The process requires 10 multiply-accumulates for the LPC polynomial, repeated 200 times for each spectral frequency. An overhead for indexing and converting finding angular frequencies may require 1*200 operations (if not implemented using a look-up table). Thus these calculations in each frame would be 2200 operations, or 73260 ips.

A7.1.2 Speech analysis

Similar to the interior noise analysis, the CELP decoder is provided with spectral, amplitude and pitch information by the downlink. Construction of a spectral array would also require 73260 ips, and speech classification, like speech detection needs the LSP measure construction and rather more comparisons (up to 8), or 633 ips.

A7.1.3 Hearing model

The hearing model applies a perceptual comparison to the speech and noise spectra in order to ascertain the audibility (and thus provide information on the intelligibility) of the currently decoded speech in the current background noise for an average listener.

Given that spectra have been provided to the model, the equal loudness pre-emphasis is based on a precalculated array, and the log calculation is performed by table look-up, the warping from frequency to Bark scale and the critical band convolution (to derive a 40-point Bark-domain weighted spectral array) together require around 350 operations per frame for both the speech and noise spectra, or 23310 ips.

A7.1.4 Formant detection

If speech is classified as not being in the *fricative* or *non-speech* classes, then formant detection must be used to determine the position of the first three spectral peaks. Differentiation of the 200 point LPC-derived spectrum requires 199 subtractions (and comparisons to determine zero-crossings). in the regions of each zero-crossing, another 4 subtractions for second differentiation, 3 additions, to average the peak value and a comparison are performed. For a 10th order LPC spectrum, there may typically be 5 such points. This leads to 199+199+(4x5)+(3x5) = 433 operations per frame, or 14420 ips.

A7.1.5 Intelligibility measure

The speech intelligibility measure compares the perceptually weighted speech spectrum, and the perceptually weighted noise spectrum in the region of the three most prominent speech spectral peaks. Comparison takes the form of averaging the weighted speech to weighted noise amplitude values in the ranges around each formant peak. This can be performed using 18 additions and 4 multiply operations per frame, a total of 733 ips.

For unintelligible voiced speech, the intelligibility measure investigates the regions around the formant positions to determine whether formant shifting can improve intelligibility. For this purpose, it conducts the same 22 operations, using the same formant data, but using noise data from shifted locations. In addition to the basic 22 operations, an array of correction data must be added in to the shifted values to account for the perceptual change in formant amplitude should the frequency be shifted. This correction data is based upon the perceptual weighting obtained for purely white noise, and increases the 22 operations to 31 per frame. For a worst case analysis, considering that every frame is unintelligible voiced speech then this adds another 1032 instructions, giving 1765 ips.

A7.1.6 Expert system

The expert system must use intelligibility results from the hearing model, and with consideration to the speech and noise spectra, and the available types of enhancement, direct such enhancements. The expert system decides which (if any) speech alterations to apply, and if necessary regulates the degree of adjustment.

The most efficient means of implementing such a system is through a series of binary decisions regarding intelligibility, speech type, noise type, absolute amplitude and formant shift applicability. There are thus a maximum of 5 levels to the decision tree, meaning only 5 comparisons per frame.

In addition, the selective amplification scheme requires the use of a gain multiplier which must be tracked and updated frame-by-frame (1 multiply-accumulate) and protected from overflow/underflow (1 decision and one load), totalling another 3 operations per frame.

The expert system thus requires around 266 ips.

A7.1.7 Speech adjustment

Selective amplification causes modification of the amplitude of certain frames of speech. The actual amplification process is already performed by the CELP decoder using the CELP gain parameter. For selective amplification, only single additional multiplication, performed on the CELP gain parameter, is required (33 ips).

For formant shifting by LSP adjustment, firstly the LSPs describing each formant are found through 9 subtractions and comparisons (for a 10th order, 10 LSP system), and then up to three formants, or the entire set of LSPs are moved. For a non-Bark based shift, this requires a maximum of 10 multiple-accumulate operations, but for the Bark-based shift, the bark calculation would be performed using a look-up table (as it involves log and sinh calculations). There would be 10 lookups, 10 additions and divisions, followed by another 10 lookups. With per-frame setup calculations, this may then total 1820 ips.

Formant narrowing or widening would simply require 9 subtractions and comparisons to locate formant-related LSPs, and then 6 multiply accumulates to adjust each line of the three pairs, totalling approximately 800 ips.

Note that only one type of speech enhancement would be chosen per frame, and thus the worst case for enhancement complexity is the Bark-based shift requirement of 1820 ips.

Appendix 8: Publications

Draft copies of some of the publications referred to in the thesis are reproduced in this appendix. Being draft copies, some small differences will be evident when compared to the final published document.

A8.1 Electronics Letters [67]

LSP analysis and processing for speech coders

Indexing terms: CELP, Speech Enhancement, LSP

Abstract: Linear prediction parameters within CELP coders are commonly represented by line spectral pairs (LSP), giving stable filters and efficient coding. However, LSP manipulation can also alter the frequencies of the represented signals. We use computationally efficient LSP manipulation to enhance the intelligibility of speech degraded by acoustic interference.

Introduction: Line spectral pairs are a mathematical transformation of the linear prediction parameters generated within a CELP coder [1]. They have risen to prominence because they are guaranteed to be stable even after quantization [2], and may be quantized with fewer bits than representations such as reflection coefficients or log-area ratios, whilst maintaining speech quality [3].

LSP values are based in the frequency domain as shown in fig. 1 where the linear prediction spectrum obtained from analysing a typical segment of speech is plotted. Lines drawn at the LSP frequencies derived from the linear prediction parameters are overlaid on this.

Interpretation of LSPs: Spectral peaks are usually bracketed quite closely by LSP line pairs, with the degree of closeness being dependent upon the sharpness of the spectral peak, and its amplitude. Note that in fig. 1, the three largest spectral peaks correspond with the 3 narrowest pairs of LSPs, and that this ordering is analogous to that given by the bandwidths of the three spectral peaks.

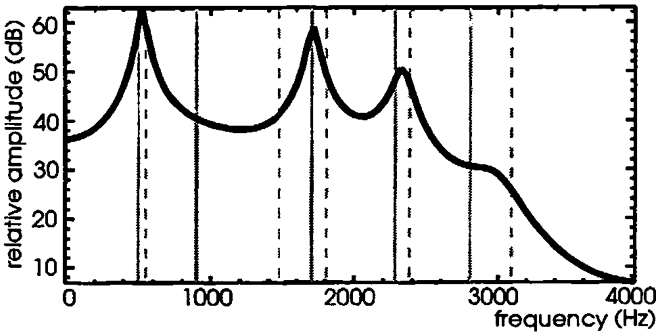


Figure 1. LPC spectral plot of a speech frame with LSPs overlaid (odd shown as solid and even as dotted lines).

Speech alteration through LSP modification: The effects of moving lines may be determined by comparing the original linear prediction spectrum with that obtained from the linear prediction parameters derived from an altered set of lines.

Fig. 2 compares an original spectrum with that obtained by altering four lines by plotting the newly created spectrum and the difference between this and the original spectrum. In fig. 2, lines that are moved closer together (B) have resulted in a higher amplitude, sharper peak between them. Lines that are further apart (A) have resulted in a wider, lower amplitude spectral peak between them. The effects on the underlying spectrum of modifying a line, are predominantly confined to the immediate frequency region [4].

Considering that the original spectrum in fig.2 was derived from speech, and that the three spectral peaks represent formants, the effect of the LSP operations has been to either 'sharpen' an individual formant (reducing formant bandwidth) or widen it (increasing formant bandwidth) and alter relative amplitudes.

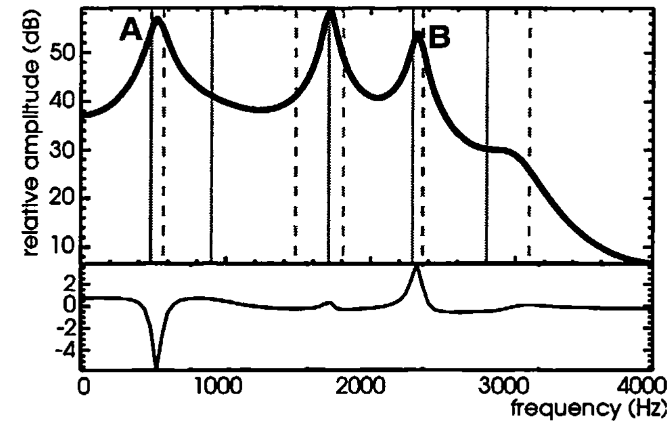


Figure 2. LSP-modified speech frame with overlaid LSPs (top) and a plot of the change between original spectrum of fig. 1 and the modified spectrum (bottom).

Speech enhancement by formant spreading: Altering the positions of LSPs has been shown to change the underlying spectrum. In the case of vowels, moderately widening the lines corresponding to spectral peaks (*increasing* the bandwidths of the formants) has been found to improve intelligibility.

Fig. 3 demonstrates the effect on the LPC-derived spectrum of widening the three pairs of lines that correspond to the three spectral peaks. It can be seen that the peaks have become broader but that the spectral valleys between the peaks have also increased in amplitude.

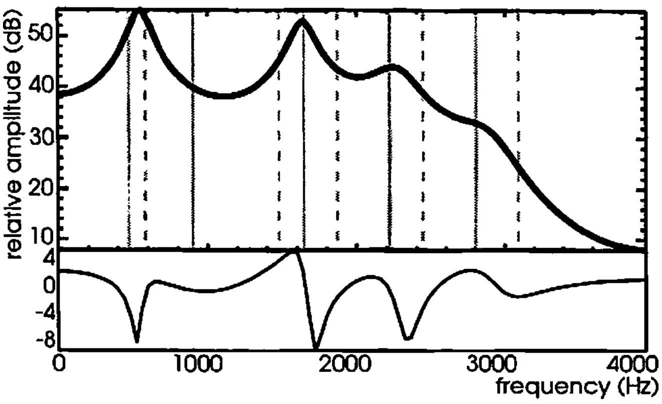


Figure 3. 20% formant widened speech frame with overlaid LSPs (top) and a plot of the difference between this and the original spectrum (bottom).

LSP alteration such as described here carries a minimal processing overhead compared to alternative methods. For example the action of narrowing or widening formant bandwidths may alternatively be performed with an adaptive filter (such as the perceptual error weighting filter of [5]) which has the form of a p-order IIR filter. If operating over a window of N samples, this would entail $2Np$ multiplications per window. For the case of LSP alteration, only $p-1$ comparisons and $p+5$ multiply-accumulate instructions are required to isolate and alter the three main formants in similar ways.

Testing: intelligibility tests were constructed using a set of LSP-modified voiced syllables, processed to widen or narrow the three spectral peaks corresponding to the three narrowest LSP pairs (preselected as corresponding to the three main formants). Normalization was applied to the processed speech which was then interposed between framing syllables and mixed with noise before presentation to the listener. White noise filtered to create a spectral representation of vehicle interior noise [6] was used.

Alternative forced choice questions were applied to a number of listeners within an anechoic chamber. Testing determined mixing SNR values required to span the range between just-intelligible and just-unintelligible (for the given syllables for average listeners), and determined the degree of formant alteration that generated the highest average increase in intelligibility.

Other authors have applied complex normalization functions to formant bandwidth-altering algorithms [7], however such functions may themselves improve intelligibility in such tests irrespective of the formant-altering function. We thus normalized processed speech to be equivalent either by-amplitude or by-power to non-processed speech.

Results: Preliminary testing identified that the optimum degree of LSP alteration of those tested was to induce a widening in the lines describing each formant of 20% when SNR was set so that average recognition rate for all listeners was 52.4%.

Results obtained from testing on 18 volunteers were analysed. 17 of the 18 subjects found the processed speech more intelligible, with the minimum improvement in syllable recognition rate (to a 95% confidence level) found to be 21%.

Conclusion: It has been shown that the transformation of LSPs can provide a useful means of manipulating the frequency spectrum of speech waveforms. An algorithm for the enhancement of speech intelligibility in the presence of acoustic noise has been shown to be effective by subjective testing. Such a scheme is computationally much more efficient than alternatives, such as adaptive filtering, if LSPs are available.

I.V.McLoughlin, R. J. Chance
School of Electronic and Electrical Engineering,
The University of Birmingham, Edgbaston,
Birmingham, UK, B15 2TT
i.v.mcloughlin@bham.ac.uk, r.j.chance@bham.ac.uk

References

[1] Goalic A, Saoudi S, "An intrinsically reliable and fast algorithm to compute the line spectrum pairs (LSP) in low bit-rate CELP coding", ICASSP, pp. 728-731, 1995.

[2] Saito S, Nakata K, *Fundamentals of speech signal processing*, Academic Press, 1985, chapter 9.

[3] Kabal P, Ramachandran RP, "The computation of line spectral frequencies using Chebyshev polynomials", IEEE Trans. A.S.S.P. Vol. ASSP-34, no.8, pp. 1419-1425, 1986.

[4] Paliwal KK "On the use of line spectral frequency parameters for speech recognition", Digital Signal Processing, Vol. 2, pp. 80-87, 1992.

[5] Chen JH, Cox RV, Lin Y, Jayant N, Melcher MJ, "A low delay CELP coder for the CCITT 16kb/s speech coding standard", IEEE J. Selec. Areas Comms, volume 10, no.5, pp. 830-849, 1992.

[6] Tempest W, *The Noise Handbook*, Academic Press, 1985, chapter 9.

[7] Schaub A, Straub P, "Spectral sharpening for speech enhancement/noise reduction", ICASSP, pp. 993-996, 1991.

A8.2 DSP '97 [66]

Analysis and modification of LSPs for speech intelligibility enhancement

I.V.McLoughlin, R.J.Chance
School of Electronic and Electrical Engineering
The University of Birmingham
Edgbaston, Birmingham
B15 2TT, U.K.
i.v.mcloughlin@bham.ac.uk, r.j.chance@bham.ac.uk

CELP coders commonly use line spectral pairs (LSP) to represent linear prediction parameters, giving stable filters and efficient coding. LSP ordering is thus related to the spectral properties of the underlying signal such that analysis of LSP positions can reveal useful information about the frequency distribution of that signal. In addition, LSP manipulation can alter frequencies within the represented signal. This paper describes computationally efficient LSP-based methods of speech analysis and speech modification, and the application of these methods to enhance the intelligibility of speech degraded by acoustic interference.

1 Introduction

Line spectral pairs (LSP) are a mathematical transformation of linear prediction parameters as generated and used within many speech compression systems, such as CELP coders [1]. Their usage originates from their stability even after quantization [2], and that they may be quantized with fewer bits than representations such as reflection coefficients or log-area ratios, with less reduction in speech quality [3].

LSPs describe the two resonance conditions arising from an interconnected tube model of the human vocal tract (which is a consequence of the linear prediction representation [4]). These conditions relate to the modelled vocal tract being either fully open or fully closed at the glottis, with the consequent resonance frequencies being the line spectral frequencies. In reality, as the glottis is opened and closed rapidly, resonances occur at frequencies somewhere between the two extremes. The LSP representation thus has a significant physical basis.

LSPs are resident in the frequency domain as shown in fig. 1 which shows lines drawn at the LSP frequencies derived from the linear prediction parameters, overlaid on the linear prediction spectrum. The linear prediction parameters were obtained from performing a 10th order linear predictive analysis on a 20ms frame of voiced speech.

2 Interpretation of LSPs

Spectral peaks are usually bracketed quite closely by LSP line pairs, with degree of closeness being determined by the sharpness of the underlying spectral peak, and its amplitude. Note that in fig. 1, the three tallest spectral peaks correspond to the 3 narrowest pairs of LSPs (between lines 1 and 2, 5 and 6, 7 and 8 respectively), and that the separation of each line pair relates to the bandwidth of the corresponding spectral peak.

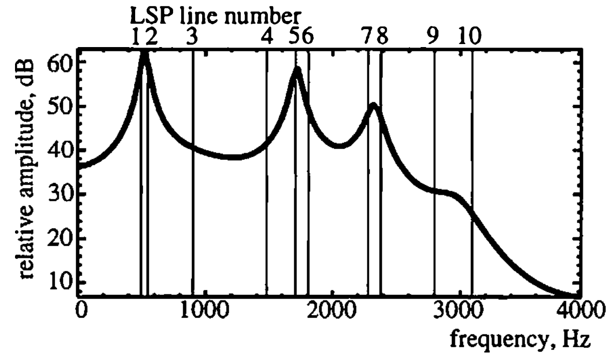


Figure 1. Plot of LPC spectrum with LSPs overlaid.

3 LSP-based processing

Within a typical CELP speech codec system, LSP data is derived from an LPC analysis of a frame of speech. The LSPs are quantized and transmitted from encoder to decoder, where they are reconstructed and used within a synthesis filter, as shown in fig. 2.

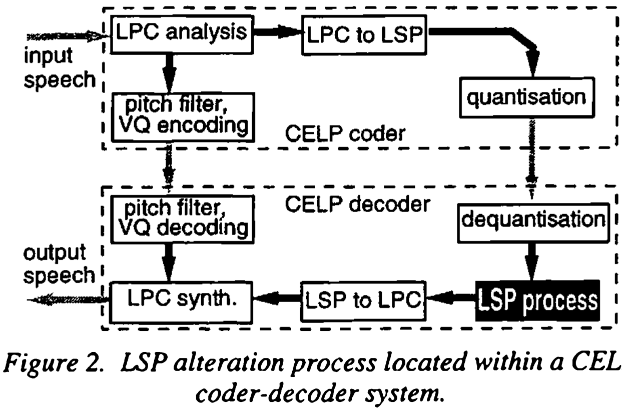


Figure 2. LSP alteration process located within a CELP coder-decoder system.

In the diagram of fig.2, the LSP process block illustrates the functional location of an LSP-based analysis and modification functions within the CELP codec. The process block receives LSP data as transmitted from the coder, can analyse this data and then alter it before the data is used to generate LPC coefficients and then reconstruct speech.

4 LSP analysis

As LSP position relates to the frequency spectrum of the underlying signal, it is reasonable to expect that analysis of LSP data can reveal facts about the spectrum.

Analysis methods may be divided into instantaneous and continuous procedures. The former being used to interpret a single frame of speech and the latter being used to interpret speech features over time.

Possible the most useful instantaneous LSP analyis procedure is the detection of speech frequency resonances; the spectral peaks or formants of fig.1 are an example. The measure simply compares the separation in frequency between each consecutive pair of lines, with narrower separations meaning larger spectral peaks. For speech analysis purposes, the centre frequency between the three closest pairs of lines corresponds well with the frequencies of the three most prominent formants as determined by other methods such as solving the LPC coefficient equation or differentiating the spectrum.

Continuous LSP analysis methods include determining overall spectral 'tilt' and degree of voicing. Spectral 'tilt' can be measured by comparing the distribution of LSP values to those obtained for a reference spectrum (such as the flat spectrum of white noise), and this measure can be used, for example, to detect periods of fricative speech. Measurement of the presence, position and amplitude (by LSP separation) of spectral peaks can indicate the degree of voicing present in a given period of speech.

5 The effects of LSP adjustment

The effects of moving lines may be determined by comparing the original linear prediction spectrum with that obtained from the linear prediction parameters derived from an altered set of lines. This is equivalent to comparing the spectrum of one frame of input speech and one frame of output speech from fig. 2.

Compare the spectrum shown in fig. 3 to the original spectrum of fig. 1. In fig. 3, the separation of line pair {1:2} has been increased, resulting in a wider, lower amplitude spectral peak between them. The separation of line pair {5:6} has been decreased, and the lines have also been translated upward in frequency, causing a sharper peak between them, at a higher frequency.

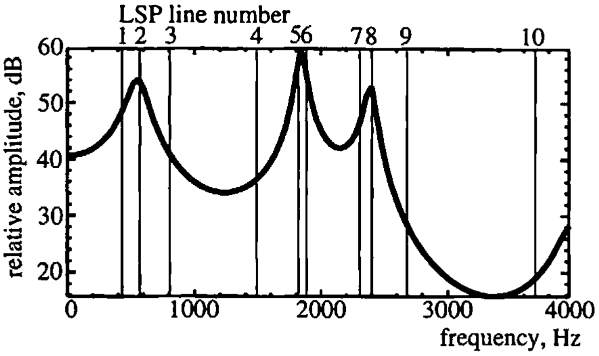


Figure 3. LPC spectrum derived from an altered set of LSPs (overlaid and numbered).

Angular frequencies of 0 and π both demonstrate the properties of virtual lines, which can be demonstrated where line 10 has been moved closer to 4kHz. A spectral peak has formed between the line and the virtual line located at the angular frequency of π.

Paliwal [5] reports that the effects on the underlying spectrum of modifying a line, are predominantly confined to the immediate frequency region. Note however that amplitude changes in one region will cause some compensatory power redistribution in other regions.

The original LPC data from which the spectral plots of fig.1 and 3 were calculated, was derived from a recording of speech, so that the three spectral peaks represent formants. LSP operations have thus been demonstrated to alter formant bandwidths and shift the position of formants.

The LSP operations to derive fig. 3 are performed as follows. If ω_i represent the LSP frequencies where $i = 1 \ldots p$ (with the order p being 10) and ω_i' the altered frequencies, then narrowing line pair {1:2} by degree α would be achieved by:

$$\omega_1' = \omega_1 + \alpha(\omega_2 - \omega_1) \qquad (1)$$

$$\omega_2' = \omega_2 - \alpha(\omega_2 - \omega_1) \qquad (2)$$

and shifting line k by degree γ may be achieved by:

$$\omega_k' = \omega_k + \omega_k(\gamma - 1)(\pi - \omega_k)/\pi \qquad (3)$$

When altering the frequency of lines it is important to avoid the formation of unintentional resonances by narrowing the gap between two previously separated lines. This problem may be obviated by either intelligently selecting and moving lines or by moving the entire set of LSPs. In the latter case, movement of lines 1 and 10 closer to angular frequencies of 0 and π may induce a resonance (see line 10 in fig. 3). Eqn. 3, designed for upward shifting, progressively limits the degree of formant shift as a frequency of π is neared.

This method of adjusting line pairs consequentially alters the frequency relationship between the formants being shifted, and so degrades the perceived quality of the underlying speech. To reduce quality degradation, a perceptual basis for line shifting was introduced, whereby frequencies were altered by a constant bark. If b_k is the bark corresponding to frequency ω_k then the line shifted by degree δ would be given by;

$$\omega_k' = 600 sinh\{(b_k + \delta)/6\} \qquad (4)$$

In such a case, a hard limit is applied to prevent LSP values approaching π. Fig. 4 illustrates the upward shift factor applied to LSPs by eqns. 3 and 4 with $\gamma = 1.5$ and $\delta = 1$.

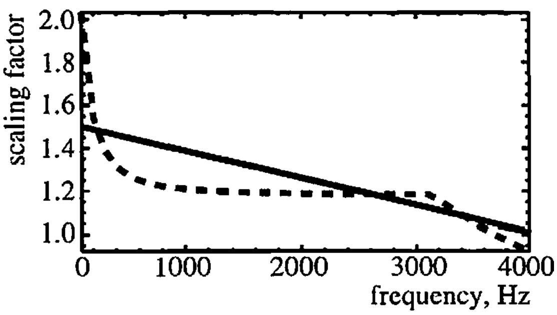


Figure 4. LSP upward shift for γ=1.5 derived from eqn. 3 (solid line) and δ=1 eqn. 4 (dotted line).

The LSP narrowing enhancement scheme requires two operations per line, or 6 operations for a 3-formant frame. Shifting using eqn.3 also requires two operations per line, or 20 operations to shift all lines in a frame. LSP shifting using eqn. 4 requires around 6 operations per line.

Similar bandwidth-altering and formant shifting effects can be produced using an adaptive filter such as that implemented by Schaub and Straub [6], however this requires at least $2Np$ operations per N-sample frame.

For a typical 10th-order analysis system with frame size of 240 samples, the LSP processes described here are respectively 800, 480 and 80 times more efficient, provided that LSP data is already available (which is the case for many CELP coders). If LSP data is not available however, it is likely that the overheads required to transform LPC coefficients to and from LSPs are greater than any efficiency gain.

5 Speech enhancement testing

In order to determine the effects of the two LSP processes described above, tests were conducted to measure the intelligibility of certain processed and unprocessed vowels, when replayed to listeners in an acoustically noisy environment.

Prospective methods of intelligibility enhancement under investigation were formant frequency movement and formant bandwidth alteration as described in section 4. In the tests, processing was applied to chosen vowels which were then normalized before being mixed with shaped noise and presented to listeners. The noise was created to model the acoustic interference within an average car interior [7].

Listeners were presented with a series of vowels delimited by unprocessed consonants (voiced plosives), and asked to choose which of two vowels they had heard. Overall amplitude was set to a comfortable level by each listener, seated within an anechoic chamber, and being presented with example questions prior to the commencement of testing. The relative amplitude of each sound was confined to a narrow range such that approximately 50% of vowels were correctly identified for average listeners, with every phoneme being individually normalized by amplitude.

Each vowel was interposed between each of three different consonant pairs, with instances of four different speech-noise amplitude ratios. Formant bandwidth altered vowels, formant shifted vowels, and a control (unprocessed) version of each vowel was included in a randomly-ordered question list, which was extended by repeating the final group of questions at the beginning to minimise learning effects.

Acoustic noise mixed with the speech presented to listeners consisted of a normally distributed signal shaped by a 20th order IIR filter to resemble car interior noise. The filter transfer function was created to replicate the graphs of average vehicle interior noise power given in [7]. The noise exhibited an extreme low-frequency bias, with the amplitude falling off significantly as frequency increases.

6 Choice of enhancement types

A formant shift was designed to move all formants upward in frequency by a small amount. As frequency increases, the power of the vehicle interior noise decreases, thus improving the signal-to-noise ratio at that frequency. For formant movements, the 'formant-to-noise' ratio is thus increased. Too great a formant shift results in speech quality degradation.

Subjective tests determined that an upward movement in formants using $\gamma = 1.5$ produced a noticeable intelligibility improvement but did not significantly reduce perceived quality.

Preliminary testing also determined the speech intelligibility alteration that could be expected from applying different degrees of formant bandwidth widening and narrowing.

It was found that a formant bandwidth change of $\alpha = -0.7$ typically resulted in an intelligibility improvement, and did not significantly reduce the subjective quality of the vowel sounds.

7 Results

18 volunteers were tested in the manner described in section 6. In 17/18 cases, formant bandwidth adjustment improved recognition rates, and in 16/18 cases, formant shifting improved recognition rates.

Table 1 lists both the average recognition rates for all listeners and the enhancement factor. Recognition rate is defined as the percentage of correct replies given by all of the listeners in the tests for the given conditions. The enhancement factor is defined as the average improvement in recognition rate over unenhanced speech, for speech altered using the relevant LSP process.

condition	recognition	improvement
all vowels	52.4%	-
unenhanced	44.4%	-
position	54.2%	1.22
bandwidth	58.6%	1.32

Table 1. Average recognition rates, and improvement factor attributed to the LSP-based enhancements.

Further statistical analysis reveals that, to a 95% confidence level, the formant shift improved intelligibility by 15% and that formant bandwidth adjustment improved intelligibility by 21%.

8 Conclusion

It has been shown that the transformation of LSPs can manipulate the frequency spectrum of speech waveforms in a useful fashion. Two algorithms for the enhancement of speech intelligibility in the presence of acoustic noise have been shown to be effective by subjective testing results. Both LSP processing schemes presented in this paper are significantly more efficient to compute than alternatives such as adaptive filtering.

Existing CELP coder implementations that utilise LSP data may integrate a speech-enhancement processing element if the target environment is acoustically noisy. Such an addition would be computationally simple to integrate into the CELP structure.

9 References

[1] Goalic A, Saoudi S, "An intrinsically reliable and fast algorithm to compute the line spectrum pairs (LSP) in low bit-rate CELP coding", ICASSP, pp. 728-731, 1995.

[2] Saito S, Nakata K, *Fundamentals of speech signal processing*, Academic Press, 1985, chapter 9.

[3] Kabal P, Ramachandran RP, "The computation of line spectral frequencies using Chebyshev polynomials", IEEE Trans. A.S.S.P. Vol. ASSP-34, no.8, pp. 1419-1425, 1986.

[4] Sugamura N, Itakura F, "Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP -", Speech Communication 5, pp. 199-215, 1986.

[5] Paliwal KK "On the use of line spectral frequency parameters for speech recognition", Digital Signal Processing, Vol. 2, pp. 80-87, 1992.

[6] Schaub A, Straub P, "Spectral sharpening for speech enhancement/noise reduction", ICASSP, pp. 993-996, 1991.

[7] Tempest W, *The Noise Handbook*, Academic Press, 1985, chapter 9.

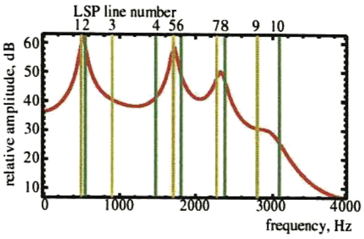
A8.3 DSP '97 [66] poster

Improving Speech Intelligibility by Line Spectral Pair Adjustment

- For replaying uncorrupted speech in a noisy environment
- Alters speech characteristics to improve intelligibility
- System is compatible with standard CELP-based speech coders

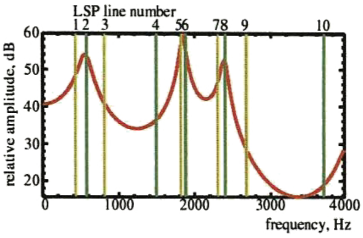
Line Spectral Pairs

- Are resident in the frequency domain
- Relate predictably to the underlying spectrum
- Can be readily altered and remain stable



LSP adjustement

- LSP alteration affects the underlying spectrum in the immediate neighbourhood of the adjusted lines
- Can reduce the amplitude and increase the bandwidth of spectral peaks by widening the separation of lines, see {1:2}
- Can increase the amplitude, decrease the bandwidth, see {5:6}.
- Can move the LSPs udward or downward infrequency to move the spectral peaks correspondingly, see lines {5:6}



- Moving adjacent LSPs closer together by degree α is accomplished by:

$$\omega'_{k+1} = \omega_{k+1} + \alpha(\omega_{k+2} - \omega_{k+1}) \quad ①$$

$$\omega'_{k+2} = \omega_{k+2} - \alpha(\omega_{k+2} - \omega_{k+1}) \quad ②$$

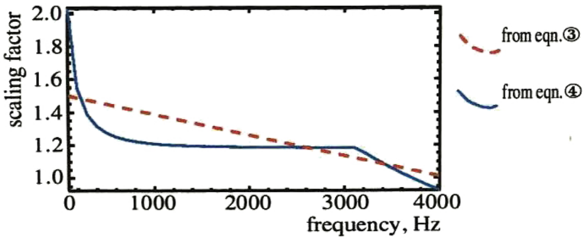
- And shifting line k by degree γ can be acheived using:

$$\omega'_k = \omega_k\{\pi + (\gamma - 1)(\pi - \omega_k)\} \quad ③$$

- To introduce a perceptual basis for the degree of LSP shifting, a bark-based scheme may be used:

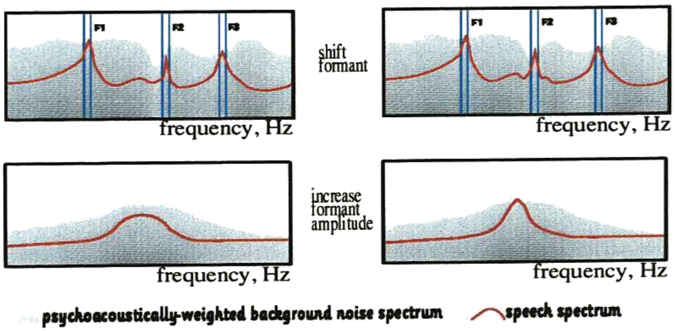
$$\omega'_k = 600 sinh\{(b_k + \delta)/6\} \quad ④$$

- Care must be taken with all schemes to prevent the positions of lines from approaching the extremes corresponding to DC and the Nyquist rate.
The following graph shows the degree of LSP shift resulting from both schemes when γ =2.5 and δ =1 respectively.



Enhancing Speech Intelligibility

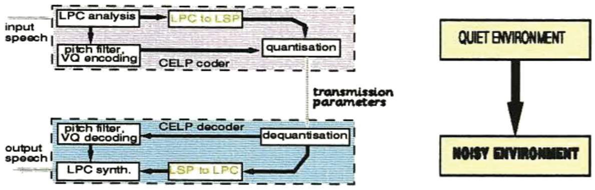
- Move formant frequencies away from interfering tones, or shift formants to frequency regions with lower noise amplitude levels
- Increase formant peak amplitudes to break through the level of interfering noise
- Increase formant bandwidth to reduce the effect of wideband noise (spread-spectrum speech transmission)



http://www.pobox.org.sg/~ivm

CELP Coder Structure

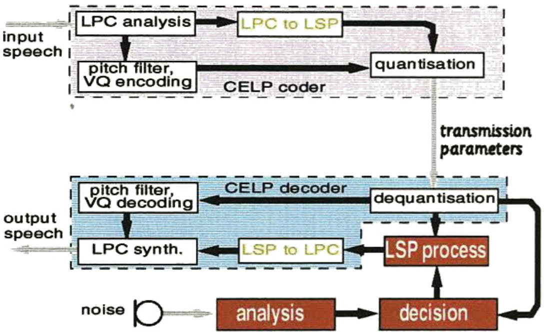
- Figure shows CELP coder-decoder structure structure, ignoring the codebook search loop



- CELP transmits LSP, pitch and gain information from coder to decoder

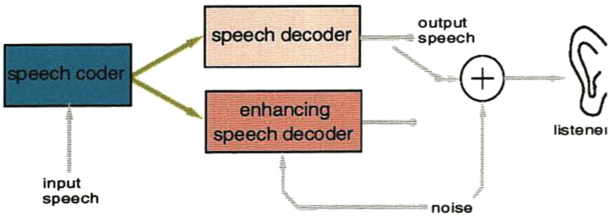
Speech Intelligibility-Enhancing CELP Coder Structure

- Adjusting LSP value alters the recreated output speech
- LSP adjustment is based upon current speech being decoded and background noise analysis



Testing procedure

- Process uncorrupted speech through speech coder and decoder structures
- Normalise the levels of the processed speech signals
- Mix noise and present to a group of listeners
- Measure intelligibility rate from listeners for processed and non-processed speech



- 18 listeners heard, in total, more than 1100 different C-V-C triplets
- Instances of control (unprocessed), LSP position shifted and LSP bandwidth-altered speech w replayed in random order to each of the listeners
- The acoustic background noise used was filtered pseudo-random noise with a spectrum correspo to average car interior noise

Results

- 17/18 listneners found formant bandwidth-adjusted speech to be more intellibible
- 16/18 listeners found formant shifted speech to be more intelligible
- Over all listeners, correct response rates were:
  average for non-processed speech: **44.4%**
  average for formant shifted speech: **54.2%**
  average for formant widened speech: **58.6%**
- To 95% confidence level, formant shifting improved recognition by 15%
  formant bandwidth adjustment improved recognition by 21%

Conclusion

- **LSP transformations can predictably and reliably alter speech characteristics**
- **Such transformations may be integrated into standard CELP-based speech coding systems**
- **LSP adjustement is a computationally efficient speech alteration method**
- **Targetted LSP-based transformations can significantly improve speech intelligibility under certain conditions**

References

1 Ainsworth WA, *Advances in Speech, Hearing and Language Processing*, JAI press Ltd, 1990.

2 Alcántera JI, Dooley GJ, Blamey PJ, Seligman PM, "Preliminary evaluation of a formant enhancement algorithm on the perception of speech in noise for normally hearing listeners", J. Audiology, vol. 33, no.1, pp. 15-27, 1994.

3 American National Standards Institute, "American national standard method for measuring the intelligibility of speech communication systems", standard S2.3-1989, Acoustical Society of America, 1990.

4 Atal BS, "Predictive coding of speech at low bit-rates", IEEE Trans. on Comunications, COM30, pp. 600- , 1982.

5 Atal BS, Hanauer SL, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoustical Soc, America. vol. 50, pp.637-655, 1971. Reprinted in Flanagan JL, Rabiner LR (Ed), *Speech Synthesis*, Dowden Hutchinson and Ross Inc., 1973.

6 Atal BS, Schroeder MR, "Predictive coding of speech signals and subjective error criteria", IEEE ASSP, ASSP-27, 1979.

7 Azirani A, Jeannes R, Faucon G, "Optimizing speech enhancment by exploiting masking properties of the human ear", ICASSP, pp. 800-803, 1995.

8 Bagshaw PC, Hiller SM, Jack MA, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching", from CSTR website, Uni. of Edinburgh, April 1997.

9 Beranek LL, "The design of speech communications systems", Proc. IRE, pp. 880-890, September 1947.

10 Blamey PJ, Dowell RC, Clark GM, "Acoustic parameters measured by a formant-estimating speech processor for a multiple-channel cochlear implant", J. Acoustical Soc. America, vol. 82, no.1, pp. 38-47, 1987.

11 Boll SF, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. ASSP, vol. 27, no.2, pp. 113-120, 1979.

12 Boyd I, Southcott CB, "A speech codec for the skyphone service", B.T. Technical Journal, vol. 6, no.2, pp. 50, 1988.

13 British Standards Institute, BS6086: "Measurement of Noise Inside Motor Vehicles", 1981.

14 CCITT (ITU), "Coding of speech at 16kbit/s using low-delay code excited linear prediction", Recommendation G.728, September 1992.

15 Chan C, "Computation of LSP parameters from reflection coefficients", Electronic Letters, vol. 27(19), pp. 1773-1774, 1991.

16 Chen JH, Cox RV, Lin Y, Jayant N, Melcher MJ, "A low delay CELP coder for the CCITT 16kb/s speech coding standard", IEEE J. Selec. Areas Comms, vol. 10, no.5, pp.

830-849, 1992.

17 Chen H, Wong WC, Ko CC, "Comparison of pitch prediction and adaptation algorithms in forward and backward adaptive CELP systems", IEE Proc.- I, vol. 140, 1993.

18 Cheng YM, O'Shaughnessy D, "Speech enhancement based conceptually on auditory evidence", ICASSP, pp. 961-963, 1991.

19 Cheng YM, O'Shaughnessy D, "Speech enhancement based conceptually on auditory evidence", IEEE Trans. Sig. Pro., vol. 39, no.9, pp. 1943-1954, 1991.

20 Chiesa A, "Experimental studies on noise inside cars", J. Sound Vib, vol. I, no.2, pp211-225, 1964.

21 Cracknell P, "Asphalt Bungle", Motoring & Leisure, the magazine of the Civil Service Motoring Association, pp.12-13, March 1995.

22 Crawford DH, Stewart RW, Toma E, "Digital signal processing strategies for active noise control", IEE Electronics & Comms Engineering Journal, vol. 9, no.2, pp81-89, April 1997.

23 Darwin CR, Gardner RB, "Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality", J. Acoustical Soc, America, vol. 79, pp.838-845, 1986.

24 Doelle LL, *Environmental Acoustics*, section 3.5, McGraw-Hill, 1972.

25 Dowling REP, Turner LF, "Modelling the detectability of changes in auditory signals", ICASSP, pp. 133-136, 1993.

26 Drake LA, Rutledge JC, Cohen J, "Wavelet analysis in recruitment of loudness compensation", IEEE Trans. Signal Processing, vol. 41, no. 12, pp.3306-3312, December 1993

27 Drullman R, "Speech intelligibility in noise: Relative contributions of speech elements above and below the noise level", J. Acoustical Soc, America, vol. 93(3), pp. 1796-1798, 1995.

28 Duncan-Luce R, *Sound and Hearing: a conceptual introduction*, Lawrence Erlbaum and Associates, 1993, part IV of book.

29 Erricker B.C., *Advanced General Statistics*, Hodder and Stoughton, 1971.

30 Erzin E, Cetin A, Yardimci Y, "Subband analysis for robust speech recognition in the presence of car noise", ICASSP, pp. 417-420, 1995.

31 Fant G, Tatham M (Ed), *Auditory Analysis and Perception of Speech*, Academic Press, 1975.

32 Ganong WF, *Review of Medical Physiology*, chapter 9, 9th edition, Lange Medical Publications, 1979.

33 Gao Y, Huang T, Haton JP, "Central auditory model for spectral processing", ICASSP, pp. 704-707, 1993.

34 Garland CR, Menez JE, Rosso MM, "Adaptive code excited predictive coding", IEEE Trans. Sig. Pro., vol. 40, no.6, pp. 1317-1326, 1992.

35 Gerson IA, Jasiuk MA, "Techniques for improving the performance of CELP type speech coders", ICASSP, pp205-208, 1991.

36 Glaskin M, "Earplugs let speech through", Sunday Times, pp.10, 17 December 1995.

37 Goalic A, Saoudi S, "An intrinsically reliable and fast algorithm to compute the line spectrum pairs (LSP) in low bit-rate CELP coding", ICASSP, pp. 728-731, 1995.

38 Guylain R, Kabal P, "Wideband CELP speech coding at 16kbits/sec", ICASSP, pp. 17-20, 1991.

39 Hardwick J, Yoo CD, Lim JS , "Speech enhancement using the dual excitation speech model", ICASSP, pp. 367-370, 1991.

40 Harland DG, "Rolling noise and vehicle noise", J. Sound Vib., volume 43, no.2, pp.305-315, 1975.

41 Hawley M, *Speech Intelligibility and Speaker Recognition*, Dowden, Hutchinson & Ross Inc, 1977.

42 Hermansky H, "Perceptual linear predictive (PLP) analysis of speech", J.Acoustical Soc. America, vol.87, no.4, pp1738-1752 April 1990.

43 Hess WJ, "Pitch determination - An example for the application of signal processing methods in the speech domain", Signal Processing Theories and Applications; Proc EURASIP, pp. 625-634, 1980.

44 "ISO/MPEG-Audio standard layers", Editorial pages of Studio Sound, pp40-41, July 1992.

45 Jamieson DG, Brennan RL, Cornelisse LE, "Evaluation of a speech enhancement strategy with normal-hearing and hearing impaired listeners", Ear and Hearing, vol. 16, no.3, pp. 274-286, 1995.

46 Jayant N, "Digital Coding of Wideband Audio", Tutorial #3, April 26, 1993, presented during ICASSP 1993.

47 Jayant N, Johnston, Safranek, "Signal compression based on models of human perception", Proc. IEEE, vol. 81, no.10, pp1383-1421, October 1993

48 Kabal P, Ramachandran R, "The computation of line spectral frequencies using Chebyshev polynomials", IEEE Trans. ASSP, vol. 34(6), pp. 1419-1425, 1986.

49 Keele CA, Neil E, *Samson Wright's Applied Physiology*, section 40, 12th edition, Oxford University Press, 1973.

50 Kleijn BW, Krasinski DJ, Ketchum RH, "Fast methods for the CELP speech coding algorithm", IEEE Trans. on A.S.S.P, vol. 38, 1990.

51 Knagenhjelm H, Kleijn W, "Spectral dynamics is more important than spectral distortion", ICASSP, pp. 732-735, 1995.

52 Kroon P, Swaminathan K, "A high quality multirate real time CELP coder", IEEE J. Selec. Areas Comms., vol. 10, no.5, pp. 850-857, 1992.

53 Kryter K, *The effects of noise on man*, Second Edition, Academic Press, 1985.

54 Kryter K, *The Handbook of Hearing and the Effects of Noise*, chapter 4, Academic

Press, 1994.

55 chapter 4 of [54]

56 Kuo CC, Jean FR, Wang HC, "Speech classification embedded in adaptive codebook search for CELP coding", ICASSP, vol. II, pp. 147-151, 1993.

57 Laflamme C, Adoul JP, Salami R, Morissette S, Mabilleau P, "16kbps wideband speech coding technique based on algebraic CELP", ICASSP, pp. 13-16, 1991.

58 Lee JI, Un CK, "On reducing computational complexity of codebook search in CELP coding", IEEE Trans. Comms., vol. 38, no.11, 1990.

59 Licklider J, "Effects of amplitue distortion upon the intelligibility of speech", J. Acoustical Soc. America, vol. 18(2), pp. 429-434, 1946.

60 Lim J (Ed), *Speech Enhancement*, Prentice Hall, 1983.

61 Makhoul J, "Linear prediction: a tutorial review", Proc. IEEE, vol. 63, no.4, pp. 561-580, 1975.

62 Markel J, Gray A, *Linear Prediction of Speech*, Springer-Verlag, 1976.

63 Martens JP, "A new theory for multitone masking", J. Acoustical Soc. America, vol. 72(2), pp. 397-, 1982.

64 McCandles SS, "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. ASSP, vol. 22, no. 2, pp. 135-141, 1974.

65 McLoughlin IV, Chance RJ (both of the University of Birmingham) Simoco International, Cambridge UK, "Method and apparatus for speech enhancement in a speech communications system", patent no. 9714001.6 lodged with the UK patent office, 2nd July 1997.

66 McLoughlin IV, Chance RJ, "LSP-based speech enhancement", 13th International Conference on DSP, Santorini, Greece, July 1997.

67 McLoughlin IV, Chance RJ, "LSP analysis and processing for speech coders", IEE Electronics Letters, pp. 743, vol. 33, no.99, April 1997.

68 McLoughlin IV, "CELP and Speech Enhancement", M.Phil(Q) thesis, The University of Birmingham, April 1996.

69 Moore B, *An introduction to the Psychology of Hearing*, chapter 1, 3rd edition, Academic Press, 1992.

70 chapter 3 of [69]

71 chapter 7 of [69]

72 chapter 8 of [69]

73 Moore B.C.J., *Hearing*, 2nd edition, Academic Press, 1995.

74 Niederjohn R, Grotelueschen J, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression", IEEE Trans. ASSP, vol. 24(4), pp. 277-282, 1976.

75 Nikias CL, Raghuveer MR, "Bispectrum estimation: A digital signal processing framework", Proc. IEEE, vol.75, no.7, July 1987, pp869-891.

76 Olive J, "Automatic formant tracking by a Newton-Raphson technique", J. Acoustical Soc, America, vol. 50, no.2, part2, pp. 661-670, 1971.

77 Ovens M, "Speech intelligibility testing: summary of methods", an internal Simoco International report, published 16th August 1996

78 Oyama G, "A stochastic model of excitation source for linear prediction speech analysis-synthesis", ICASSP, pp 941-944, 1985

79 Paksoy E, Srinivasan K, Gersho A, "variable rate speech coding with phonetic segmentation", ICASSP, pp. 155-158, 1993.

80 Paliwal KK, "On the use of line spectral frequency parameters for speech recognition", Digital Signal Processing, vol.2, pp80-87, 1992.

81 Pickett J, *The Sounds of Speech Communication*, Allyn and Bacon, 1980.

82 chapter 1 of [81]

83 chapter 2 of [81]

84 chapter 4 of [81]

85 Pollack I, "Speech communications at high noise levels: the roles of a noise-operated automatic gain control system and hearing protection", J. Acoustical Soc, America, vol. 29(12), pp. 1324-1327, 1957.

86 Priede T, "The effect of operating parameters on sources of vehicle noise", J. Sound Vib., vol. 43, no.2, pp. 239-252, 1975.

87 Rabiner LR, "Applications of voice processing to telecommunications", Proc. IEEE, vol. 82, no.2, pp. 199-228, 1994.

88 Rabiner LR, Schafcr RW (Eds), *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

89 Raghuveer MR, Nikias CL, "Bispectum estimation: A parametric approach", IEEE Trans. ASSP, Vol. ASSP-33, no.4, OCt 1985, pp1213-1230.

90 Saito, Nakata , *Fundamentals of speech signal processing*, Academic Press, 1985.

91 chapter 9 of [90]

92 appendix 21 of [90]

93 appendix 22 of [90]

94 Saoudi S, Boucher J, Guyader A, "A new efficient algorithm to compute the LSP parameters for speech coding", Signal Processing, vol. 28(2), pp. 201-212, 1992.

95 Schaub A, Straub P, "Spectral sharpening for speech enhancement/noise reduction", ICASSP, pp. 993-996, 1991.

96 Schaefer RW, Rabiner LR, "Digital representations of speech signals", Proc. IEEE, vol. 63, no.4, pp. 662-677, 1975.

97 Schaefer RW, Rabiner LR, "System for automatic formant analysis of voiced speech", J. Acoustical Soc. America, vol.47, no.2 (part 2), pp634-648, 1970.

98 Schroeder MR, Atal BS, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates", ICASSP, pp. 937-940, 1985.

99 Schroeder MR, Atal BS, Hall JL, "Optimizing digital speech coders by exploiting masking properties of the human ear", J. Acoustical Soc. America, vol. 66(6), pp. 1647- , 1979.

100 Sears WG, *Anatomy and Physiology for Nurses and Students of Human Biology*, chapter 10, 4th edition, Arnold, 1967.

101 Sen D, Holmes WH, "Perceptual enhancement of CELP speech coders", ICASSP, pp. 105-108, 1993.

102 Sen D, Irving DH, Holmes WH, "Use of an auditory model to improve speech coders", ICASSP, vol. II, pp. 411-415, 1993.

103 Sharp DWN, White RL, "Determining the pitch period of speech using no multiplications", ICASSP, vol II, pp. 527-531, 1993.

104 Sharpley AD, "Summary of speech intelligibility testing methods", from Dynastat Inc. (internet page http://www.realtime.net/dynastat/), May 1996.

105 Slifka J, Anderson TR,"Speaker modification with LPC pole analysis", ICASSP, pp644-647, 1995.

106 Soong FK, Juang BW, "Line spectrum pair (LSP) and speech data compression", ICASSP, part 1.10, pp.1-4, 1984

107 Sugamura N, Farvardin N, "Quantizer design in LSP speech analysis-synthesis", IEEE Journal Selec. Areas Comms., vol. 6, no.2, pp432-440, February 1988.

108 Sugamura N, Itakura F, "Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP-", Speech Communication, vol. 5, pp213-229, 1986

109 Summerfield Q, Foster J, Tyler R, Bailey PJ, "Influences of formant bandwidth can auditory frequency selectivity on identification of place of articulation in stop consonants", Speech Communications, vol. 5, pp199-215, 1986.

110 Tempest W (Ed), *The Noise Handbook*, Academic Press, 1985.

111 chapter 1 of [110]

112 chapter 3 of [110]

113 chapter 4 of [110]

114 chapter 9 of [110]

115 Thomas IB, "The influence of first and second formants on the intelligibility of clipped speech", J. Acoustical Soc, America, vol. 16, no. 182, 1968.

116 Thomas I, Niederjohn R, "The intelligibility of filtered-clipped speech in noise", J. AES, vol.18(3), pp. 299-303, 1970.

117 Thomas I, Ohley W, "Intelligibility enhancement through spectral weighting", Proc. IEEE Conf. Speec, Comms and Processing, pp. 360-363, 1972.

118 TIMIT database, a CD-ROM database of phonetically classified recordings of sentences spoken by a number of different male and female speakers. Speech disc 1-1.1 of the National Institute of Standards and Technology of the U.S. Department of Commerce.

119 Trancoso IM, Atal BS, "Efficient search procedures for selecting the optimum innovation in stochastic coders", IEEE Trans. ASSP, vol. 38, no.3, 1990.

120 Trans-European Trunked Radio System (TETRA) standard, A European Telecommunications Standards Institute (ETSI) standard.

121 Tsoukalas D, Paraskevas M, Mourjopoulos J, "Speech enhancement using psychoacoustic criteria", ICASSP, pp. 359-362, 1991.

122 U. S. Environmental Protection Agency, *Transportation, Noise and Noise from Equipment Powered by Internal Combustion Engines*, Washington D.C. pp.109, 1971.

123 van Velden JG, Smoorenburg GF, "Vowel recognition in noise for male, femals and child voices", ICASSP, pp. 961-963, 1991.

124 Virag N, "Speech enhancment based on masking properties of the auditory system", ICASSP, pp. 796-799, 1995.

125 Voiers WD, "Evaluating processed speech using the diagnostic rhyme test", Speech Technology, pp30-39, Jan/Feb 1983.

126 Vredestein (company), advertisement, Motoring & Leisure, the magazine of the Civil Service Motoring Assn., back cover, June 1995.

127 White F, *Our Acoustic Environment*, John Wiley & Sons, 1976.

128 chapter 6 of [127]

129 chapter 9 of [127]

130 Williams D, Tempest W,"Noise in heavy goods vehicles", J. Sound Vib., vol. 43, no.1, pp. 97-107, 1975.

131 Witten IH, *Principles of computer speech*, chapter 2, Academic Press, 1982.

132 Yannakoudakis EJ, Hutton PJ, *Speech synthesis and recognition systems,* John Wiler & Sons.

133 Yasheng Q, Kabal P, "Pseudo three-tap pitch prediction filters", ICASSP, pp. 523-527, 1993.

134 Yegnanarayana B, "Formant extraction from linear-prediction phase spectra", J. Acoustical Soc, America, vol. 63(5), pp. 1638-1640, 1978.

135 Zong-Liang-Wu , Schwartz JL, Escudier D, "Modelling spectral processing in the central auditory system", ICASSP, pp. 373-377, 1990.