Extrapolation of Wideband Speech From the Telephone Band

by

Aryn Alexandra Pyke

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science
Graduate department of Electrical and Computer Engineering
The University of Toronto

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-29415-3

Canada

Extrapolation of Wideband Speech From the Telephone Band

Aryn Alexandra Pyke, M.A.Sc.
Department of Electrical and Computer Engineering
The University of Toronto, 1997

Telephone speech is bandlimited to the frequency range between 300 and 3300 Hz, which compromises its quality. Wideband speech, accommodating frequencies up to 7000 Hz, provides higher quality but at a cost of increased transmission bandwidth. The proposed pseudo-wideband (PWB) speech algorithm regenerates approximations of the bands missing from telephone speech. This is possible because of the strong inter-band correlations which stem from the acoustics of the production apparatus.

For this receiver-based algorithm, the improvement in effective bandwidth requires no extra transmission bandwidth, and involves no codec standardization issues. The spectral envelope and spectral detail are deconvolved via linear predictive analysis, and each is mapped independently to its PWB counterpart. The algorithm is based on parametric analysis using a uniform tube tract model, and has good potential for speaker independence. Performance was encouraging for a preliminary investigation, but a more sophisticated acoustic model is desirable for additional quality improvement.

Acknowledgments

I would like to thank my supervisor, Professor Frank Kschischang, for his invaluable advice and encouragement throughout the research and preparation of this thesis. I would also like to acknowledge my family for their unerring support. Finally, I would like to thank my friends, especially Joel Alo and Lucy Pegoraro, for their continual support and tolerance.

Contents

List of Figures

Chapter 1

Background

1.1 Introduction

Perceptually, telephone speech is less natural and sometimes less intelligible than face-to-face speech. The quality of telephone speech is primarily compromised by the band limiting done in the Public Switched Telephone Network (PSTN) to reduce the sampling rate and save on transmission bandwidth. The PSTN uses a Pulse Code Modulation (PCM) coding scheme. The speech signal is band limited to avoid aliasing and then sampled at 8000 samples per second. Each sample is then quantized according to the 8-bit, non-linear μ-law quantizer[1] [16].

Although the frequency content of speech can extend up to 20 kHz [20], telephone speech is band limited to the range of approximately 300 to 3300 Hz. The range of perceptually significant frequencies for speech perception extends to about 10 kHz, which is considerably beyond that of telephone-band (TB) speech. Specifically, the range from 50 to 200 Hz contributes to increased naturalness, presence and loudness. Its exclusion from the telephone band causes the speech to sound 'tinny', but is presumably justified by the fact that this range has little influence on intelligibility [20] [18]. The supra TB range from about 3400 Hz to 7000 Hz is thought to contribute to increased intelligibility, sound differentiation and crispness [7].

In all fairness, TB speech is a well-justified tradeoff between speech quality and the call capacity of the PSTN. However, advances in speech coding and speech processing can now enable the network to support so called wideband (WB) speech (with a bandwidth of approximately 8 kHz) over its voice grade channels. This contra-intuitive capability can

[1] Outside of North America, an A-law quantizer is used.

actually be accomplished in two ways. Wideband speech can be sampled and efficiently digitally encoded in real time using less than 16 kbits/s [10]. This can be easily accommodated over voice grade channels since bit streams as fast as 33.6 kbits/s can be sent using modems complying with the V.34Q standard. Naturally, to recover the speech at the receiving end, a modem and corresponding wideband speech decoder are needed. This thesis presents an alternative way to obtain wideband speech with no cost in transmission bandwidth. An algorithm is presented which regenerates pseudo-wideband speech at the receiver using only the received TB speech.

A more formal definition of the problem and an outline of objectives is presented in the following section. To provide the reader with the necessary speech background, an overview of speech signal characteristics and relevant speech production and perception issues is presented in Section 1.3. This is followed by a summary and discussion of relevant previous work in Section 1.6. The details of the proposed algorithm are provided in Chapter 3. Chapter 4 outlines the experimental methodology, simulation results, and performance evaluation. A discussion of the results and their implications is presented in Chapter 5. Finally, conclusions and recommendations for future work are presented in Chapter 6.

1.2 Problem Definition

The primary objective of this work was to develop a receiver-based digital speech processing algorithm to produce a bandwidth and quality enhancement of TB speech. Essentially, the algorithm will effect a mapping, T, from TB speech to pseudo-wideband (PWB) speech, as shown in Figure 1.1.

In order that the algorithm have the potential to function in conjunction with any existing narrowband speech coders, in particular, the pulse code modulation of the PSTN, the only allowed input to the algorithm is TB speech. For the purpose of this study, TB speech is assumed to incorporate only frequencies in the range 300 Hz to 3300 Hz, and wideband speech is defined as speech incorporating frequencies in the range 0 Hz to 8000 Hz.

The speech produced by the algorithm will be dubbed pseudo-wideband (PWB) speech to distinguish it from true wideband (WB) speech. Research in speech perception indicates a decrease in frequency resolution in the upper frequency band, which affords a certain leeway in generating the wideband speech. The objective is to generate a perceptually viable approximation to the true WB speech, such that the subjective quality is

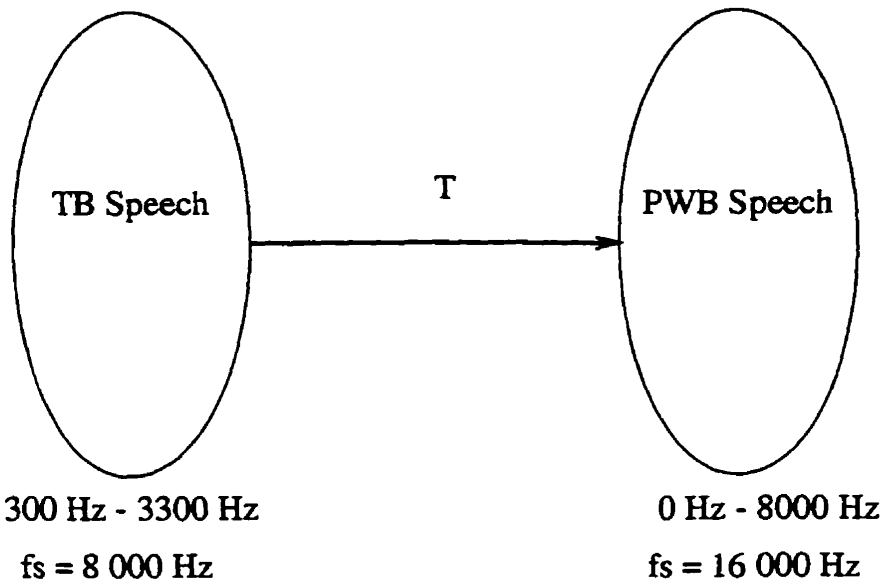


Figure 1.1: Pseudo-wideband speech generation.

notably improved as compared with TB speech. Admittedly, subjective quality is difficult to quantify, therefore objective measures, such as segmental SNR and segmental spectral SNR were used in the design and evaluation of the algorithm. These objective measures are defined and described in Section 1.5.2. Final performance is also evaluated based on informal listening tests.

Unlike the previous attempts described in [1] [2] [6] [8], it was desired that the algorithm be analytic and physically motivated in nature, rather than being based solely on signal statistics and pattern matching. It was believed that such an analytic nature would furnish it with a strong potential for speaker and language independence.

Among the set of secondary goals was the desire to regulate the algorithm's complexity to make it amenable for real-time implementation on a low-cost DSP processor. Furthermore, it was desired to arrive at a better general understanding of the nature and extent of the correlation between narrowband/TB and WB speech.

1.3 Speech Production and Speech Signal Characteristics

Speech has a complex structure embodying a great deal of redundancy. In particular, if only the upper band is isolated (3300-8000 Hz), the speech is still fully intelligible, as it is when only the TB part is present. The fact that the semantic message content is fully

preserved in different bands is a testament to the strength of the correlation between the different frequency bands of the speech signal, and the potential feasibility of the current endeavor.

To understand the structure and characteristics of speech signals it is useful to think in terms of a speech production model. The two main physical structures involved in speech production are the vocal chords and the vocal tract. As air egresses from the lungs, the vocal chords can be relaxed or can vibrate at various frequencies, usually in the range of 55 Hz to 333 Hz, thus applying a periodic pressure signal to the tract [7] [28] [17]. Within the tract, differences in the cross-sectional area, influenced by the position of the tongue, lips, and jaw, cause sound wave reflections which give rise to resonances or *formants*, which appear as peaks in the speech spectrum. In the simplest and most common speech production model, the linear source-filter or terminal-analogue model, the contributions of the chords and the tract are partitioned [24]. As depicted in Figure 1.2, the whole system can be modeled by a source, $\epsilon(t)$, isolated from, and leading into a linear filter, $T(z)$, modeling the tract [19]. The *excitation* signal, $\epsilon(t)$ models the stimulating signal from the chords, and $T(z)$ models the modulation of that signal by the tract.

The dynamics of the tract and chords produce speech signals which are non-stationary, that is, the frequency composition of speech varies with time. In particular, vocal tract structures, or *articulators*, rarely stay fixed for more than 40 ms, so the required $T(z)$ is actually a time varying filter [23, p. 206]. The rate of vibration of the vocal chords can change as quickly as one octave per 100 ms [23, p. 233]. It is such time variation of vocal chord frequency that is the physical basis of intonation, such as the raising of pitch[1] at the end of a question. These movements are sufficiently slow and smooth that the speech signal is generally accepted to be effectively stationary within time segments on the order of 20 ms. Within such a segment, or *frame*, all the signals in the source-filter model can be considered stationary, and $T(z)$ represents a linear time-invariant filter.

In analyzing speech segmentally, speech can be partitioned into two main classes of sounds based on whether the vocal chords are relaxed or vibrating during that frame. In *voiced* speech, exemplified by vowel sounds, the vocal chords are vibrating. The rate at which they vibrate is called the fundamental frequency, $F0$, or *pitch*, which can be assumed to be approximately constant over the duration of the frame. As depicted in the example in Figure 1.3, voiced speech is characterized by a quasi-periodic time-domain signal, $s(t)$. For *unvoiced* speech, the chords are relaxed, but a constriction is present

[1] Strictly speaking, pitch is the *perceived* tone of the sound, while frequency of vibration is a property of the stimulus.

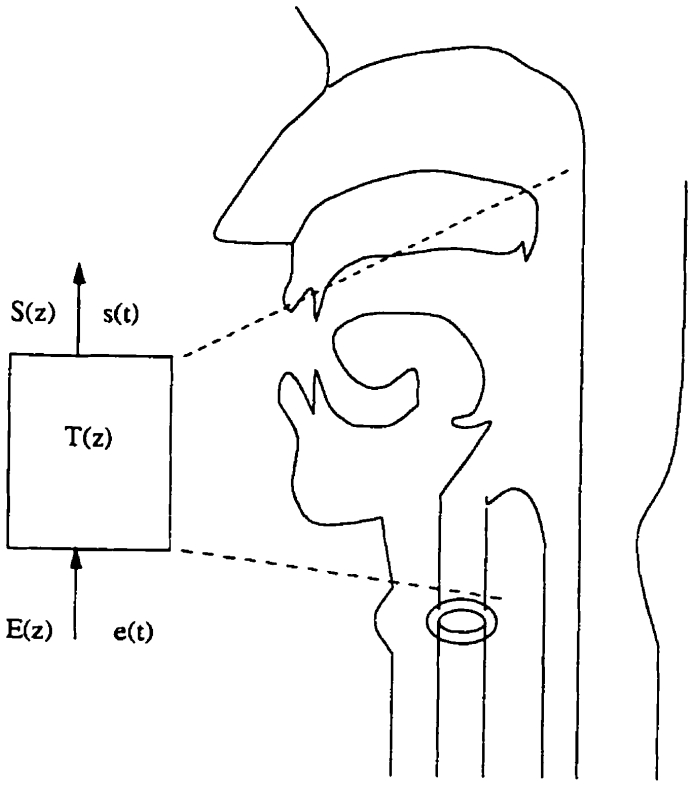


Figure 1.2: Source-filter model of speech production.

somewhere in the tract which results in turbulence as air rushes through. This turbulence serves as an excitation for the remainder of the tract. As shown in Figure 1.4, unvoiced speech is characterized by a lower amplitude, non-periodic, noise-like signal. Examples of unvoiced sounds are *fricatives* such as /s/ and /f/. Typically, unvoiced sounds have lower amplitudes and energies than voiced sounds, and have a greater proportion of their energy concentrated in higher frequency bands. A few sounds, such as /z/ in 'zip', have a mixture of unvoiced and voiced characteristics, and are referred to as mixed-mode sounds. They occur if there is a constriction causing turbulence somewhere in the tract, but the vocal chords are also vibrating.

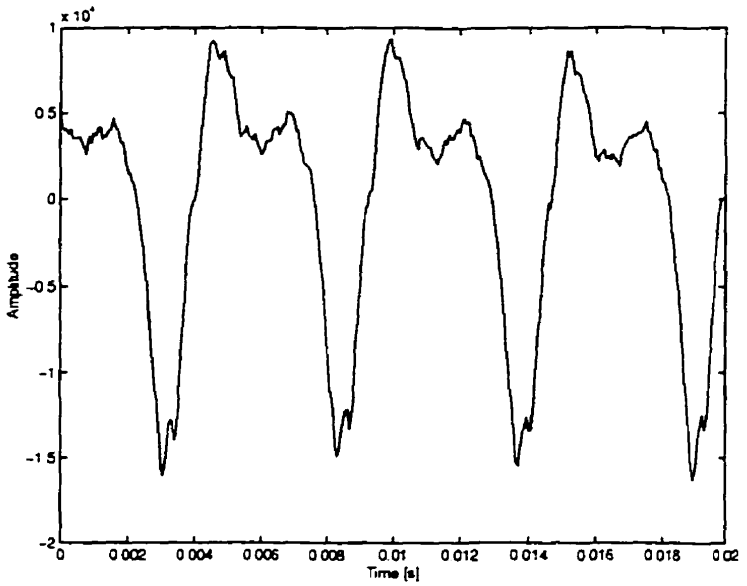


Figure 1.3: 20 ms frame of voiced speech.

The voiced/unvoiced classification leads to the extended source-filter model of speech and speech production depicted in Figure 1.5. Idealized time and frequency domain representative signals are included. For voiced speech, the glottal excitation signal is periodic. The spectrum obtained by taking the Short-Term Fourier Transform (STFT), exhibits the harmonic structure expected for a periodic signal, with the harmonics separated by $F0$, and the spectrum has a typical roll-off of -12 dB/octave [24]. The noise-like excitation for unvoiced speech is spectrally flat. For the model to handle mixed mode sounds, the switch could be replaced with a summation block allowing both sources to be active simultaneously.

Although most individuals possess the same basic apparatus and most languages

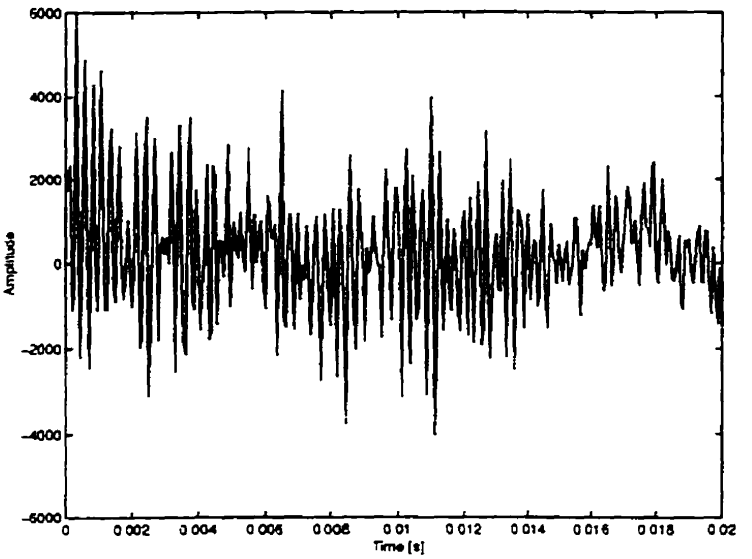


Figure 1.4: 20 ms frame of unvoiced speech.

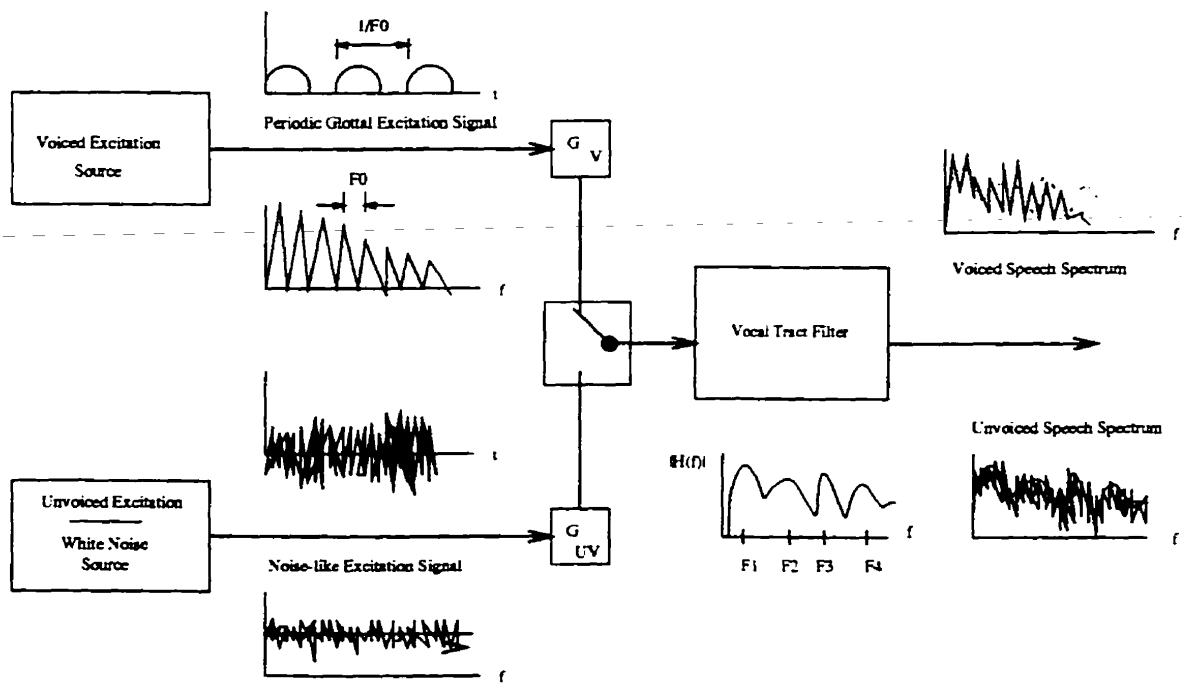


Figure 1.5: Extended source-filter model of speech production.

employ the vocal tract in similar ways, speech signal characteristics are often highly speaker and context dependent [23]. As shown in Figure 1.5, the main speech parameters are: the voicing mode; the fundamental frequency, $F0$, for voiced speech; the gain, G; the formant locations, or *F-pattern* $\{F1, F2, F3, \ldots\}$; and the respective bandwidths of the formants. There is a one-to-many mapping between a semantic unit of sound, or *phoneme*, and the corresponding acoustic signal as described by the parameters. The average fundamental frequency, $F0$, for women is approximately 210 Hz, and that for men is approximately 125 Hz [27] [24][1]. Within a given speaker the pitch usually ranges over an octave of values during speech [24]. In terms of tract variation, the average length for the pharengyl-oral tract is 17 cm for men and 13 cm for women. Tract dimensions affect the resonant frequencies and therefore the positions of the formants in the speech spectrum. The general principle can be understood in terms of a simple resonating tube of uniform cross-section and length, L, for which the resonances occur at odd multiples of $F1 = \frac{c}{4L}$, where c is the speed of sound (approximately 340 m/s for air at sea level) [24]. For men, the formant frequency spacing is approximately 1000 Hz, while for women, it is about 1301 Hz. The telephone band typically contains about four formants worth of male speech, and about three formants worth of female speech. The fundamental frequency, $F0$, and up to three harmonics often fall below the telephone band.

1.4 Linear Predictive Analysis

The deconvolution of excitation and envelope is often accomplished by Linear Predictive Analysis (LPA). LPA identifies the spectral envelope by finding the best all-pole fit of a specified order for the spectrum of that frame. The deviations from this all-pole spectral approximation constitute the excitation, or *residual* as it is called in the LPA literature.

The tenet of linear predictive speech analysis is that a linear predictor can be used to estimate the value of the next sample of the digital speech signal, based upon a linear combination of a set of preceding speech samples. Mathematically, the estimate of the next sample, $\hat{s}_n$, is expressed as

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}, \tag{1.1}$$

[1] Men have more massive cords than women which is why they tend to vibrate at lower frequencies. Within a given speaker, the cricothyroid muscles can increase the tension on the cords, and thus raise their frequency of vibration [5].

where s_n is the speech sequence, p is the order of the prediction, and the a_k's are the prediction coefficients. The residual, r_n, can therefore be expressed as

$$r_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}. \tag{1.2}$$

LPA is the process of selecting the predictor coefficients, a_k, to minimize the residual in the mean squared sense. Since speech in non-stationary, the analysis must be conducted on a segment-by-segment basis. Within a frame of N samples, the mean squared error, E is given by

$$\begin{aligned} E &= \frac{1}{N}\sum_{n=1}^{N} r_n^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(s_n - \sum_{k=1}^{p} a_k s_{n-k}\right)^2. \end{aligned} \tag{1.3}$$

To minimize E, the partial derivatives of equation 1.3 are taken with respect to the prediction coefficients, a_k, and set to zero

$$\frac{\partial E}{\partial a_i} = 0, \ for\ 1 \leq i \leq p. \tag{1.4}$$

After some algebra, this yields p linear equations in the p unknown predictor coefficients of the form

$$\sum_{k=1}^{p} a_k \sum_{n=1}^{N} s_{n-i} s_{n-k} = \sum_{n=1}^{N} s_{n-i} s_n, \tag{1.5}$$

with $1 \leq i \leq p$. The solution of this system can be computed with matrix inversion, however in practice $8 \leq p \leq 26$, and this method becomes computationally expensive [27]. Two alternative practical techniques to approximate the solution are the covariance method and the auto-correlation method whose details are well presented in [21] and [27]. In the LPA simulations conducted for this thesis, the auto-correlation method was employed, and the details of this method will be presented, along with the other implementation details, in Chapter 3. Aside from finite precision errors, analysis with this method is guaranteed to produce a stable filter [23].

Although LPA is most often described in the time domain, the frequency domain interpretation yields more insight for the current application. In particular, p^{th} order linear prediction in the time domain corresponds to modeling the spectrum of that frame with a p^{th} order all-pole model. To see this, note that the z-transform of equation 1.2 reveals that

$$\frac{S(z)}{R(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^k} = H(z), \tag{1.6}$$

where H(z) is a p^{th} order IIR filter. LPA is, in fact, equivalent to maximum entropy spectral estimation [21]. For an appropriate value of p, the macroscopic spectral shaping is fully captured in the *envelope*, $H(z)$, while the spectral detail is captured in the residual, $r(n)$, which has a generally flat spectral trend.

The formulation in (1.6) reveals the consistency between LPA and the source-filter speech production model in Figure 1.2. Thus, $H(z)$ is often referred to as the tract or synthesis filter. The $H(z)$ computed in LPA actually encompasses not only the tract effects, but also glottal flow and radiation effects which are distinct from the tract filter, $T(z)$, in the source-filter production model [32]. Thus, in LPA, all major spectral shaping effects are encompassed in $H(z)$, and the residuals, both voiced and unvoiced, exhibit an overall flat spectral trend.

Viewing the residual signal as an output of the LPA process, it can be seen that

$$\frac{R(z)}{S(z)} = \frac{1}{H(z)} = A(z). \tag{1.7}$$

where $A(z)$ is referred to as the analysis or prediction filter.

Figure 1.6 displays the speech spectrum, $|S(f)|$, and corresponding envelope, $|H(f)|$, found by performing 16^{th} order LPA on the 20 ms WB speech frame depicted in Figure 1.3. The peaks in $|H(f)|$ can be interpreted as revealing the location of formants. Since speech is a real signal, a p^{th} order LPA spectral model effectively deduces the location of $\frac{p}{2}$ positive frequency formant locations[1]. Mathematically, the formant frequencies can be found by finding the poles, z_i, from the denominator of $H(z)$. Then the positive frequency formant, F_i, associated with pole,

$$z_i = |z_i| \exp j\varphi_i, \tag{1.8}$$

[1] In the example shown, there are formants at 0 Hz and 8000 Hz which don't stand out on the graph.

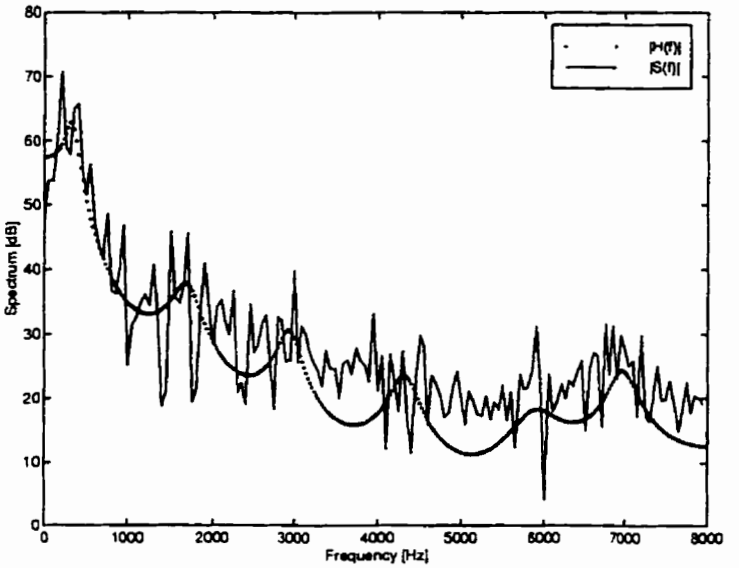


Figure 1.6: Speech spectrum and formant-scape from 16^{th} order LPA.

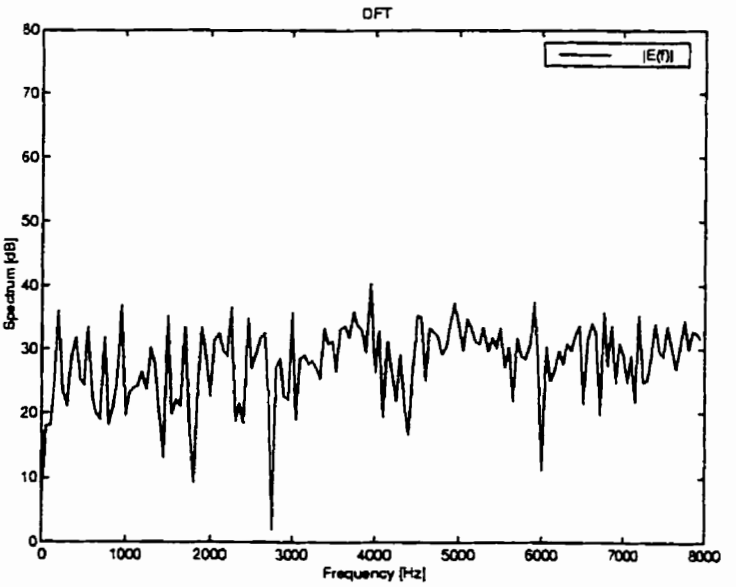


Figure 1.7: Example of a residual spectrum.

is given by

$$F_i = \frac{\omega_i}{2\pi T_s}. \tag{1.9}$$

where F_i is in Hertz, and T_s is the speech sampling period.

Figure 1.8 illustrates the decomposition and synthesis method for a speech segment according to the source-filter model via LPA.

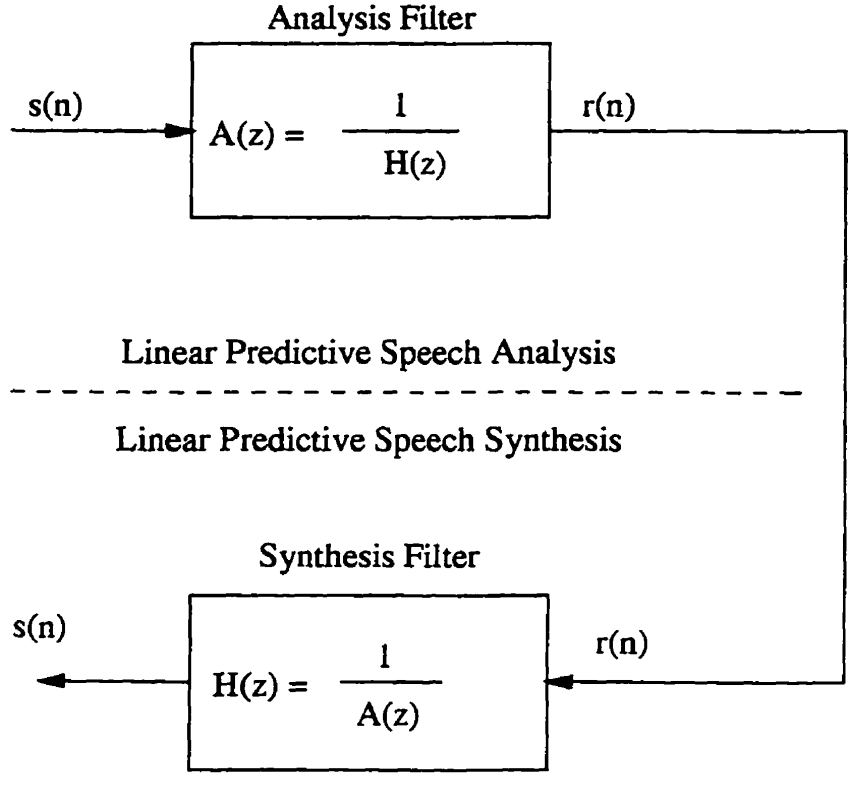


Figure 1.8: Linear predictive analysis and synthesis.

Aside from limitations in numeric precision, the deconvolution of a speech segment into spectral envelope and excitation via LPA is a lossless process. The original speech frame can be reconstructed by filtering the residual, $r(n)$, through the synthesis filter, $H(z)$. Speech synthesis by this technique is known as Residual Excited Linear Prediction (RELP). In applying LPA to speech compression, the prediction coefficients, a_k, can be transformed and quantized by various means to obtain a very compact representation of the spectral envelope. Even more coding gain is achieved when advantage is taken of the fact that relatively less information is contained in the spectrally flat residual than

in the spectral envelope. Speech of acceptable quality can be synthesized according to Figure 1.8 even when only a fairly rough approximation to the residual is used. Such residual compression and approximation techniques are discussed in Section 1.6.2.

1.5 Speech Quality: Factors and Measures

Although the primary goal of this project was to produce a significant enhancement of speech quality, it is a very difficult property to quantify. Speech quality is inherently subjective, and is dependent on incompletely understood aspects of speech perception. Nonetheless, some formal subjective and objective measures measures of speech quality were selected for the design and evaluation of the PWB speech algorithm. In this section, some basic speech quality perception factors are presented along with the selected measures of speech quality.

1.5.1 Speech Perception

A distortion is only perceptually significant if the magnitude of the distortion exceeds the resolution of the human auditory system [26]. However, there is no consensus on the best speech perception model [23]. In this section, a few general speech perception facts are provided which have provided some guidance for speech coding and enhancement techniques in the literature.

Speech perception is particularly tuned to frequency domain aspects of the signal. Spectral amplitude is much more important than spectral phase in speech perception [23, p 79]. Speech is a complex tone, in that in contains components of many different frequencies/harmonics simultaneously. Ohm's law of hearing states that the sound quality of a complex tone, often called its *timbre*, depends only on the amplitudes of its harmonics and not on their relative phases [3].

The ear is especially sensitive to frequencies in the range 200 to 5 600 Hz, in that we can discriminate between small differences in time and frequency in this range [23, p128]. Furthermore, sounds outside this band require significantly more energy to be heard than those inside it. The minimum change detectable for formant frequency is 5% while for formant bandwidth it is about 40% [24]. For regular speech, intensity is such that all harmonics should be audible up through $F4$ [23]. $F1$ is the most intense formant and perceptually the most prominent [23].

It is interesting to note that most of the perception system characteristics were described in the frequency domain. Waveform reconstruction is a sufficient but not a

necessary condition for high quality speech.

1.5.2 Objective Quality Measures

Informal subjective listening tests are ill equipped for finely quantifying distortions over multiple simulations or for evaluating the 'quality' of isolated speech components such as subbands, envelopes or excitation. In the design and evaluation of the algorithm, several, so called objective speech quality measures, were employed to assess signal quality, and aid in system parameter optimization. Objective quality and comparison methods attempt to measure those physical characteristics of the speech signal that are correlated with factors that determine speech quality. These measures allow an analytic calculation to be performed on a signal to quantify its quality, or to assess its similarity to another signal.

Unfortunately, since there is no consensus on the best speech perception model, there is also no consensus on the best objective measures of speech quality. For this project, the choice of objective measures was dictated by simplicity, flexibility and popularity. Specifically, distortion measures were chosen to enable a comparison with the results presented in [1] and [8] whose PWB algorithms are outlined in Section 1.6.

Rather than using separate excitation only, envelope only, or subband only objective measures, to isolate the effects of a given distortion, in the simulations, a distortion of interest is isolated and all other speech components are controlled to be ideal. This permits an overall subjective evaluation of each distortion to complement the set of objective measures, and allows measures for each distortion to be compared to a single set of 'ideal' control reference measures.

Since different measures tend to focus on different types of distortions, several different objectives were employed. Both time domain and frequency domain methods are desirable to get a complete picture. These measures selected were: (i) signal-to-noise ratio (SNR); (ii) segmental signal-to-noise ratio (SEG-SNR); (iii) segmental spectral signal-to-noise ratio (SS-SNR); (iv) segmental log-spectral signal-to-noise ratio (SLS-SNR); (v) misaligned SLS-SNR (vi) segmental spectral root mean square error (S-RMSE) (vii) segmental log-spectral root mean square error (LS-RMSE); (viii) segmental root mean square error (TIME-RMS). The measure are defined mathematically in Table 1.1(N is the frame length, and the distorted version of the signal is indicated with a hat).

SNR and SEG-SNR focus on the time domain waveform. SNR, evaluates the signal as a whole, and can captures the inter-frame boundary distortions which the segmentation produces. Because it deals with energy which is related to the magnitude spectrum, SNR

does not focus on slight phase errors. A synthesis frame misaligned spectral measure might also be useful for revealing any inter-frame continuity distortions.

SNR	$10log_{10}\frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}(x(i)-\hat{x}(i))^2}$
$SEG-SNR$	$\frac{10}{M}\sum_{m=1}^{M} log_{10}\sum_{n=Nm}^{Nm+n+1}\frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}(x(i)-\hat{x}(i))^2}$
$SS-SNR$	$\frac{10}{M}\sum_{m=1}^{M} log_{10}\frac{\sum_{k=1}^{N}[X^2(k)}{\sum_{k=1}^{N}\left(X(k)-\hat{X}(k)\right)^2}$
$SLS-SNR1$	$\frac{10}{M}\sum_{m=1}^{M} log_{10}\frac{\sum_{k=1}^{N} X_{dB}^2(k)}{\sum_{i=1}^{N}\left(X_{dB}(k)-\hat{X}_{dB}(k)\right)^2}$
$SLS-SNR2$	$y(n)=x(n-D)$ $\frac{10}{M}\sum_{m=1}^{M} log_{10}\frac{\sum_{k=1}^{N} Y_{dB}^2(k)}{\sum_{i=1}^{N}\left(Y_{dB}(k)-\hat{Y}_{dB}(k)\right)^2}$
$SL-RMSE$	$\frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(X_{dB}(k)-\hat{X}_{dB}(k)\right)^2}$

Table 1.1: Definitions of objective speech quality measures.

1.6 Previous Work

A few attempts to solve this problem or parts thereof appear in the literature. There has been, however, no complete or commercial solution. All related articles are particularly vague and informal in terms of the subjective and objective performance of the techniques.

The British Broadcasting Corporation considered this problem in the 1970's [9], but did not undertake a focused study. Patrick in the 1980's [25], developed some methods which required side information about the WB signal rather than just the TB component, so in essence they were wideband speech coding techniques. In 1984, Dietrich succeeded in further enhancing the bandwidth of already wideband signals, however the method could not be generalized to work using NB or TB speech as a starting point [11].

Four recent attempts also were surveyed [1] [2] [6] [8]. Figure 1.9 illustrates the basic structure common to these PWB speech generation systems, including the one proposed in this thesis. The source-filter speech production model proved valuable because the independent consideration of envelope and excitation significantly reduces the dimensionality of the mapping from TB to WB speech. In each case, the deconvolution of excitation and envelope was accomplished by Linear Predictive Analysis (LPA). The speech is partitioned into effectively stationary frames on the order of 20 ms in duration, and LPA is conducted on each frame.

It is noteworthy that within the telephone band region, the PWB speech is simply composed of an interpolated version of the input TB speech. As indicated in the envelope mapping and excitation mapping blocks, reconstruction of the upper band (UB), extending from approximately 3.3 to 7 or 8 kHz, is often considered separately from the reconstruction of the lower band (LB), extending below 300 Hz. The next sections discuss the previously attempted techniques for implementing the key blocks in Figure 1.9. Section 1.6.1 focuses solely on previously proposed techniques to reconstruct the WB envelope from the NB or TB envelope. Then, known techniques to reconstruct a WB excitation from the TB or NB excitation are discussed in Section 1.6.2. The excitation techniques come from the field of speech compression as well as PWB speech generation systems.

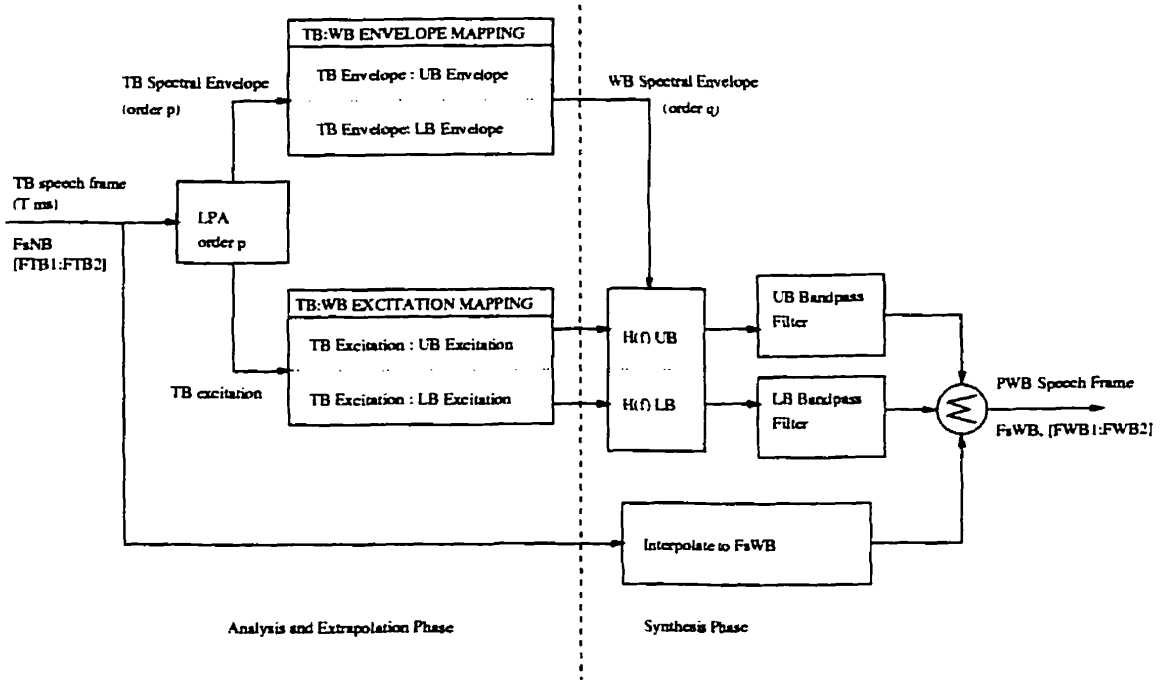


Figure 1.9: High-level block diagram of a PWB speech generation system.

1.6.1 Spectral Envelope Mapping

Having utilized LPA to extract the spectral envelopes of TB and WB speech, the task remains to discover and effect a mapping between them. Three methods will be discussed in the subsections below. Although not so named by their proposers, the methods will be identified, for descriptive clarity in this document, as: (i) the spectral envelope codebook

mapping technique: (ii) the linear prediction of WB cepstral coefficients: and (iii)the hidden Markov model (HMM) statistical recovery function [1] [6] [2] [8].

Spectral Envelope Codebook Mapping Technique

Two separate papers describe essentially the same technique which will be identified here as the Spectral Envelope Codebook Mapping Technique [1] [6]. During the training phase for this technique, a large database or corpus of wideband (WB) speech is obtained and filtered to yield a corresponding corpus of narrowband (NB) speech or telephone band (TB) speech.

To adequately describe the WB speech, a higher order LPA is usually used than that for NB or TB speech. Specifically, a NB-LPA order of p=10 and and WB-LPA order of q=16 were used in [6], although p=q=14 was used by [1]. Each spectral envelope is parametrically described by the coefficients of the denominator of the all pole model.

To effect the mapping from an NB or TB spectral envelope to a WB spectral envelope, parallel codebooks were used. Such codebooks are simply tables of representative envelopes (as described by a set of LPA coefficients or some transform thereof) which can be used to categorize/vector quantize a new envelope. In [1] the LPA descriptions of the WB speech were vector quantized according to the Linde, Buzo and Gray algorithm. The TB speech was then coerced to cluster according to how its WB counterparts clustered. Thus two codebooks were formed, a TB one and a NB one, with a one-to-one mapping between them. In [6], the same rationale is used, but the TB LPA coefficients were allowed to cluster naturally, and the WB LPA coefficients were forced to cluster according to their TB counterparts. In actual fact, for the method described in [1], two pairs of codebooks were used: one two describe the mapping between the missing lowerband spectrum and the TB; and one to describe the mapping between the missing upperband spectrum and the TB.

In operation, the TB speech is partitioned into frames. LPA is then conducted on the TB frame, and the TB-LPC codebook is searched for the closest match. This gives the position of the appropriate entry in the WB-LPC codebook. During analysis, excitation parameters are also extracted from the TB frame, however the full details of the excitation generation procedures are not provided in [1] [6]. As LPA can separate the excitation and the spectral envelope, they can also be recombined by filtering the excitation with the envelope according to the terminal-analogue model. The sources of error in this WB signal stem from errors in the synthetic excitation and vector quantization

error in the LPC codebooks. To minimize error in constructing the final WB frame, the synthetic signal is only used in the frequency bands outside the TB. The TB frame is interpolated, in order to make its sampling frequency appropriate for WB speech, and the missing frequency ranges are filled in with an appropriately filtered version of synthetic WB signal.

The algorithm is based on a fairly primitive pattern matching structure. There is no attempt to employ physical speech production modeling. Different speakers and languages could presumably be accommodated with larger, and larger codebooks.

Linear Prediction of WB Cepstral Coefficients

A method of envelope extrapolation involving the linear prediction of WB LPC cepstral coefficients was proposed in [2]. This method differs substantially from the codebook technique in two respects. First, it is analytical, and second, unlike the codebook technique which is memoryless from frame to frame, this technique employs inter-frame temporal correlations.

As with the codebook technique the speech is partitioned into overlapping frames[1], and LPA is conducted on each frame. The LPA order for the WB speech was 16, while that for NB speech was 8. The LPA *predictor coefficients*, a_k, are then transformed into LPA *cepstral coefficients*, c_k[2], according to [21]:

$$c_k = a_k - \sum_{m=1}^{k-1} \frac{m}{k} c_m a_{k-m}, 1 \leq k \leq p. \tag{1.10}$$

where p is the order of the LPA. This follows from the definition [21]

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log A(\exp jk\omega) d\omega \tag{1.11}$$

and the fact that A(z) is minimum phase. These cepstral coefficients, c_k, constitute a full description of the envelope.

The analytical method proposed essentially entails a linear prediction of the cepstral coefficients of the current WB frame in terms of the those of the current NB frame, the $M = 5$ preceding NB frames, and the $M = 5$ subsequent NB frames according to [2]

[1] In this case the frames are 25 ms in duration with a 10 ms frame rate.

[2] c_k are the cepstral coefficients for $A(z)$, and $-c_k$ are the cepstral coefficients for $H(z)$.

$$\hat{c}_j^{WB}(m) = \sum_{k=1}^{p} \sum_{l=-M}^{M} W_{k,j}(l) c_k^{NB}(m-l), 1 \leq j \leq q. \quad (1.12)$$

where $\hat{c}_j^{WB}(m)$ is the estimated j^{th} WB cepstral coefficient for frame m. m and l are frame indices. and q and p are the respective orders of the WB and NB LPA. In filter parlance. as the technique is described in [2]. the time trajectories of the NB cepstral coefficients are passed through a bank of q. p-input. single-output FIR weighing filters. W. The output of each filter in the bank is the estimated time trajectory of one of the q WB cepstral coefficients. It is perhaps more intuitive to view the system as a simple linear prediction where each wideband coefficient estimate. $\hat{c}_j^{WB}$ is calculated as linear combination of 88 NB cepstral coefficients (2M+1 neighboring frames and p NB coefficients per frame). $W_{k,j}(l)$ is then the prediction weight for the k^{th} NB coefficient in the l^{th} frame from the 'current' frame. Thus. each WB cepstral coefficient is effectively computed via an 88^{th} order linear prediction. The actual values for the predictor weights (FIR filter coefficients). $W_{k,j}(l)$. were selected/optimized during training using a least squares error criterion. The values obtained and the speech corpus used were not provided in [2]. These filters were presumably fixed after training. so that the algorithm is not adaptive during operation.

The technique described above was used to map the NB envelopes to UB envelopes. The same technique was used to map from TB envelopes to LB envelopes. In this case the sampling rate was 8 kHz for both signals and LPA order was 8 for both envelopes [2]. The technique was not applied to the full TB-to-WB problem. in particular to the mapping from 300-3300 Hz to 3300-8000 Hz.

It should be noted from the filter interpretation that the system is non-causal. With $M = 5$. a frame size of 25ms. and a frame rate of 10ms. the 'prediction' of WB cepstral coefficients depends on a total time segment of 120 ms of speech. 50 ms of which comes after the 'current' frame. Aside from any computational delay. this non-causality yields a huge operational delay which would be unacceptable for real-time applications.

The article included no objective or clear subjective measures of the performance of the technique. However it was mentioned that the mappings appeared to be highly speaker dependent. so that a system trained with one speaker would be incapable of working for another. This is a significant drawback to this technique.

Statistical Recovery Function (SRF)

This method involves an explicit statistical source model for speech. Speech in the training corpus was partitioned into an NB component (300 Hz to 3750 Hz, in this case) and a UB component (3750 Hz to 8000 Hz). Gaussian mixture generative models (GMGM) were applied separately to the two ranges to model their source characteristics. This model is described below.

The SRF constitutes a sophisticated generalization of the classification functionality of the spectral envelope codebook mapping technique described in section 1.6.1. The codebook method involves a hard classification of a TB (NB) speech frame according to the best TB codebook match, followed by a one-to-one map to the UB (LB) based on that classification. The GMGM regards a speech frame as a probablisticly weighted sum of contributions from several simultaneous sources, λ_i, rather than as a member of a single class. Similarly, the UB frame is synthesized as a probablisticly weighted sum of contributions from the UB sources, θ_j, with the weights being a function of the cross-correlation probabilities, $\alpha_{ij} = P[\theta_j|\lambda_i]$. As with the codebook mapping technique, no correlations across separate time frames are considered.

Simultaneously, each source produces a frameworth contribution of time-domain values. The statistical sources employed, for both the NB and UB, were 16^{th} order *autoregressive* Gaussian sources, which effectively means that each source is characterized by a specific 16^{th} order all-pole spectral envelope (with the excitation possibilities modeled by Gaussian noise). $N = 64$ sources/envelopes are used to model the NB ensemble, and $M = 16$ sources/envelopes are used to model the UB ensemble. Abe used a 4 bit UB spectral envelope codebook, corresponding to 16 mutually exclusive speech classes [1]. Although $M = 16$ in the SRF, these sources are simultaneous, and so can be used to generate a wider variety of UB speech frames.

The training and source parametrization was done using the estimation and maximization (EM) algorithm which is a method for maximum likelihood estimation [8]. In the E step, the respective 'responsibilities', $h_{i,d}$ of each source, s_i for each data frame, $\bar{x^d}$, in the training corpus is computed. In the M step, for each source, s_i, a responsibility weighted centroid, μ_i, of all the data is computed according to

$$\mu_i = \frac{\sum_{alld} h_{i,d}\bar{x^d}}{\sum_{alld} h_{i,d}}, \tag{1.13}$$

and that source is then moved to centroid. There are similar update rules for the covari-

ances and for the priors.

The weights, β_j used to synthesize the UB are given by

$$\beta_j = \sum_{i=1}^{N} \alpha_{ij} P[\bar{x}_{TB}|\lambda_i] P[\lambda_i]. \tag{1.14}$$

Energy normalization was also done in training so that, for a speech frame, $\bar{X}_{NB}$, $P[\bar{x}_{NB}|\lambda_i]$ would be independent of the absolute energy of the speech frame, $\bar{X}_{NB}$, rendering classification would be independent of volume. Source dependent ratios of UB signal energy versus NB signal energy were also calculated in training to ensure that in synthesis, appropriate energies for the UB components could be determined. So actually, the final weight for the UB component, (1.14) would be multiplied by an appropriate gain, G_j.

The synthetic UB tract filter is thus effectively comprised a filter bank of M fixed 16^{th} order all-pole filters (found during training), each with a variable gain factor calculated from the current NB frame as the product of (1.14) with G_j.

1.6.2 Excitation Extrapolation

As mentioned in Section 1.4, the residual, being spectrally flat, contains relatively less information than the envelope. In applying LPA to speech compression, considerable effort has been directed to compressing the residual. Some of these coding techniques and results have significant bearing on the current problem.

There are four basic relevant techniques for excitation generation and/or extrapolation in linear predictive speech processing: parametric excitation synthesis; code excitation, non-linear high frequency regeneration; and spectral duplication. In the first two techniques, the excitation signal is essentially synthesized from scratch. The latter two techniques were developed in the context of Residual Excited Linear Prediction (RELP) speech coders, and their focus is the extrapolation or interpolation of a portion (sub band) of the residual [32]. An additional technique called harmonic modeling is mentioned in the literature for generating a LB excitation from a TB excitation [6]. Some of the methods deal separately with the two classes of excitations: voiced and unvoiced. Each of the methods is briefly outlined below.

Parametric Excitation Synthesis

In many coding schemes, the excitation is parameterized in terms of voiced/unvoiced status, gain and pitch. An approximation to that excitation can then be synthesized at

the receiver according to the parameter values. For unvoiced speech, the spectrally flat excitation can be modeled with Gaussian noise of appropriate mean and variance [22] [8]. For voiced speech, the pitch, $F0$, is the crucial parameter, and there are methods of varying complexity for synthesizing an appropriate periodic excitation signal. The simplest technique, known as mono-pulse coding, involves approximating the voiced excitation with an impulse train with impulses separated by the pitch period, and scaled according to the gain [27]. In multi-pulse coding, a more complex excitation signal can be synthesized. In this case the parameters are the positions and magnitudes of the pulses, which are often selected via an analysis-by-synthesis method.

The parametric scheme is amenable to the current problem since the TB contains enough signal information to extract parameters such as gain and pitch. It is probable that parametric excitation synthesis was used in conjunction with the envelope extrapolation codebook method described in [1] since power and pitch are extracted in the algorithm. There are, however, two main disadvantages of the parametric scheme. The voiced/unvoiced classification is not always straight forward, and the binary choice is particularly unsuited to handle mixed mode sounds. Second, accurate pitch determination is often complex and error prone [27].

Code Excited Linear Prediction (CELP)

Code-excitation is essentially parametric excitation synthesis taken to the limit. A codebook of excitation signals is compiled and used to vector quantize the excitation at the transmitter. However, instead of analyzing the excitation (residual) directly to extract parameters, or to compare it to members in the codebook, an analysis-by-synthesis technique is used to select the appropriate excitation. After LPA is performed, each codebook excitation is tried with the LPA envelope, to see which gives the closest output to the true speech frame, according to a perceptually weighed mean squared error (MSE) distance criterion. Many modern coders employ this technique, in particular, the 16 kb/s International Telecommunication Union (ITU) standard G.728 [7].

In terms of the current problem, this excitation codebook method could be extended in a similar manner to the Spectral Envelope Codebook technique described in Section 1.6.1 for envelope codebooks. Thus two excitation codebooks could be used to effect a mapping from the NB to WB excitation. No mention of such a scheme was found in the literature. However, it was noted in [2] that an NB CELP coder, when applied to TB speech, produced an NB excitation which included the sub-TB frequency components

(50 to 300 Hz). Thus, a CELP coder could be used to find a suitable LB excitation from the TB speech.

In particular, baseband coders separate the envelope and excitation at the transmitter, and the compact envelope parameters are sent along with a subband of the excitation signal which has been filtered and decimated to make it more compact. At the receiver, techniques are employed to regenerate the missing portion of the excitation signal.

Spectral Duplication

The techniques described in this and the following section were developed in the context of base-band RELP coders. In these coders, the residual is bandlimited and decimated to make it more compact, and the envelope parameters along with this residual subband are forwarded to the receiver. At the receiver, techniques such as spectral duplication or non-linear frequency regeneration are employed to regenerate the missing band(s) of the residual signal.

There are two types of spectral duplication: spectral folding and spectral translation. Figure 1.10 illustrates the spectral effects of the processes. These techniques are dubbed spectral duplication, since the spectrum in the missing band is generated to be a duplicate of that in the known band. For spectral folding, the duplicate is the mirror image with respect to the folding frequency. In spectral translation, the frequency spectrum is duplicated by shifting along the frequency axis. Spectral folding has the advantage that it preserves spectral continuity at the folding frequency and it is intuitive to assume greater correlation between the lower and upper bands in the vicinity of the fold than if they were separated by 4000 Hz.

Spectral folding is the most popular method in the literature for generating a WB residual from a NB residual. It is a technique used in base-band RELP coders [32], and the method used to generate the UB excitation in the SCM-PWB speech generation algorithm of [6] and the LCC-PWB algorithm of [2]. In spectral folding, the UB excitation is simply generated by upsampling the NB excitation to the WB frequency. For mapping 8 kHz sampled NB speech to 16 kHz sampled WB speech, this is accomplished by simply inserting a zero between every sample of the NB residual.

The main advantages of this technique are its computational simplicity and the fact that it can be applied in exactly the same manner to all classes of excitations, avoiding the need for any classification block. For an unvoiced excitation, which is noise-like and spectrally flat, folding appropriately yields a noise-like, spectrally flat signal throughout

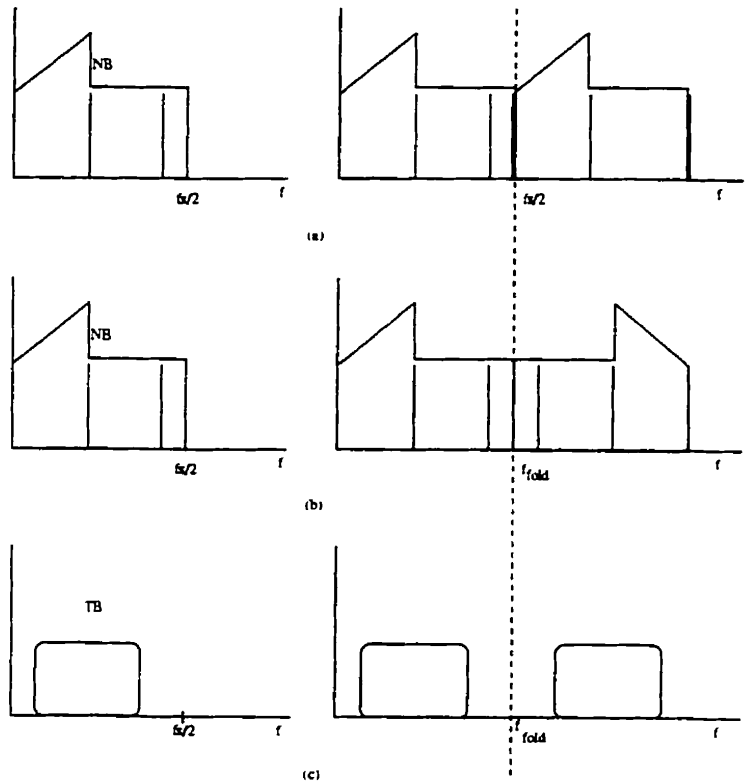


Figure 1.10: Spectral duplication: (a) NB spectral translation; (b) NB spectral folding; and (c) TB spectral folding.

the WB frequency range. For voiced speech, the fold will yield an UB excitation with a harmonic structure. For mixed mode speech, spectral folding has an advantage over parametric excitation generation since it preserves the mixed characteristics of mixed mode speech.

Despite the popularity and simplicity of spectral duplication there are some noteworthy drawbacks. For voiced speech, unless the folding frequency coincides with a harmonic or falls midway between two harmonics, folding causes an interruption in the harmonic pattern at the fold frequency. This is emphasized in Figure 1.10, in which the two vertical lines representing harmonics on either side of the fold frequency are improperly spaced. In addition to the problem with harmonic interruption, spectral folding of the TB excitation yields a significant frequency gap in the interval 3300 Hz to 4700 Hz. Similarly, spectral translation of TB speech yields a gap from 3300 Hz to 4300 Hz.

Non-Linear Frequency Regeneration

A signal with additional frequency components not present in the original can be obtained by passing the original through a non-linear device. This principle has been applied to base-band residuals, and is known in the literature as high-frequency regeneration (HFR) [22].

The non-linear distortion commonly used is waveform rectification as described by [22]

$$e^{WB}(t) = \frac{[(1+\alpha)|e^{NB}(t)| + (1-\alpha)e^{NB}(t)]}{2}, \tag{1.15}$$

where α is a constant such that $0 \leq \alpha \leq 1$, e^{NB} is the baseband residual, and e^{WB} is the WB excitation produced. This successfully produces UB frequency components, but yields a non-flat spectral trend, which would conflict with the envelope shaping. The spectrum can be flattened by performing LPA on the HFR signal and retaining the resultant residual [22]. This technique was considered for in [6], but it was reported that the resulting speech had a harsh and synthetic sound.

This method has also been applied to generate lowerband frequencies from telephone band speech. The TB speech was full-wave rectified, i.e., $\alpha = 1$, and bandpass filtered to obtain the frequency band from 80 Hz to 300 Hz, but no quality improvement measures were reported [9].

1.6.3 System Evaluation

Having discussed the basic excitation and envelope extrapolation techniques, it is possible to appreciate and compare the entire systems in [1], [2], [6], and [8]. The two main criteria for evaluating the algorithms are output speech quality and complexity.

Performance and complexity for the codebook methods hinges upon the sizes of the codebooks and the search mechanism. Abe's method was evaluated using an 8 bit low-band codebook and a 4 bit high-band codebook [1]. The inter-envelope distance measure employed to classify a new envelope according to the NB codebook was the Euclidean distance of the LP cepstrum. The corpus consisted of a total of 216 phonetically balanced words uttered by 10 male and 10 female speakers. Paired comparison listening tests were conducted using 70 TB-PWB word pairs spoken by two male and two female speakers[1], and evaluated by six listeners. It was reported that 88% found the PWB speech to be 'wider' [1]. In terms of objective measures, Abe reported a 'spectrum distortion' of 6.5 dB for the upperband reconstruction and one of 3.5 dB for the lowerband reconstruction. In [6], a 12-bit tree-structured codebook was employed. The speech corpus consisted of approximately 30 minutes of WB speech from the TIMIT database. The performance was evaluated in a speaker independent fashion, using more than 10 listeners, and 90% preferred the PWB speech. The article claimed that the method improved the discriminability of /s/ and /f/ sounds.

Avendano used a CELP coder to obtain the LB excitation, and spectral folding to approximate the UB excitation, and regenerated the WB envelope via a linear prediction of the WB LPA cepstral coefficients employing inter-frame correlations [2]. The primary drawbacks of his system are the huge delay inherent in the prediction, and the speaker dependent nature of the mapping. Furthermore, the UB reconstruction was only accomplished using NB speech as input rather than TB speech. No quantitative performance measures are provided.

[1] These speakers were different individuals than those which produced the training corpus.

Chapter 2

Speech Extrapolation Model

As initially indicated in Figure 1.1, it was desired to devise a mapping from TB speech to PWB speech. The feasibility of an acoustic modeling approach stems from the fact that there are elements of spectral periodicity and predictability intrinsic to resonating systems. This applies specifically to the behaviour of the tract. Spectral correlations between the TB and non-TB bands of the excitation also exist, although these correlations can be gleaned and exploited readily from the residual signal, making it unnecessary to model the underlying acoustics. Excitation modelling issues are presented in section 2.1. The selected tract resonance model is discussed in Section 2.2. The exploitation and application of the models for the purposes of speech extrapolation is discussed in Section 2.2.2. A summary is included at the end of the chapter.

2.1 The Excitation Source

The excitation spectrum contains relatively less information (more obvious redundancy) than the tract filter. It embodies such speech characteristics as volume and pitch. As previously discussed, there are two main classes of excitation: voiced and unvoiced. In both categories, there is a strong spectral inter-band correlation structure which is well documented in the literature. Thus, for this application, it is not warranted to delve into complex acoustic models of the excitation generation process. Instead, the discussion in this section will provide details on the basic nature of the two types of excitation magnitude spectra and the simple models which capture their inter-band correlation structure.

Voiced Excitation

The action of the vocal chords during voiced speech causes a modulation of the respiratory air stream which presents an approximately sawtooth periodic volume velocity signal to the tract [13]. This waveform is more of less independent of the acoustic impedance looking into the vocal tract from the glottis, and consequently the internal impedance of the volume velocity source can be considered to be high [29]. This supports the assumption of independence of excitation and envelope.

An idealized voiced excitation signal is depicted in Figure 2.1, with a period of 5 ms ($F0 = 200$ Hz) which is within the range of typical values for a female speaker. The magnitude spectrum of such a signal exhibits a harmonic structure with a -12 dB/octave roll-off, as shown in Figure 2.2, with harmonics separated by $F0$, the frequency of vocal fold vibration. Nuances in the exact shape, periodicity, and duty cycle of the signal can vary the spectral slope somewhat. However, there are indications that this simple voiced source model provides fairly natural speech [13].

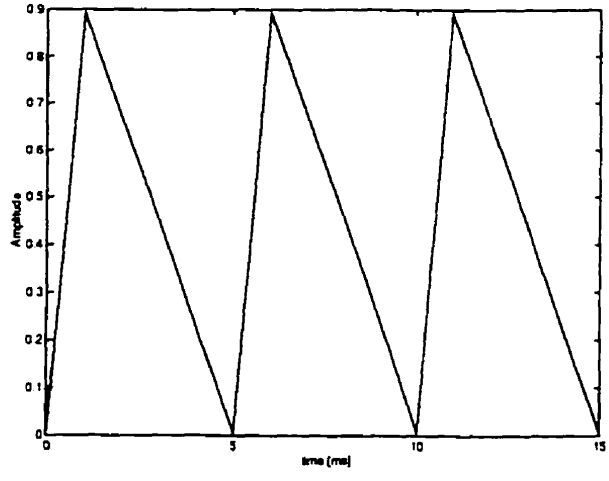


Figure 2.1: Idealized voiced excitation signal.

This ideal structure implies a theoretically easy means of excitation extrapolation. The voiced excitation can be parameterized by two quantities, signal strength (as described by amplitude, variance, or energy, E), and fundamental frequency, $F0$. Unfortunately, analysis of TB speech to determine $F0$ is an error prone and complex problem. It is also one for which incentive is limited in this case. For UB extrapolation, depending on the strength of the signal, the roll-off may reduce contributions in the upper frequency ranges

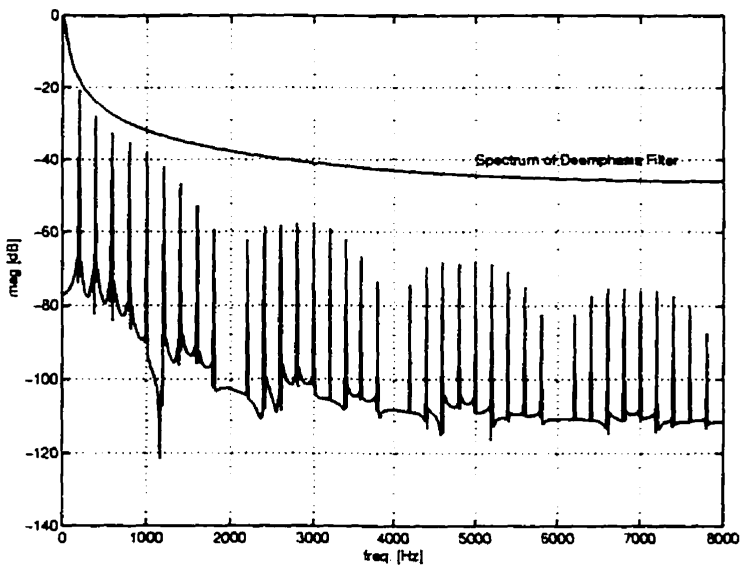


Figure 2.2: Magnitude Spectrum of the idealized voiced excitation signal.

to below the threshold of audibility. Even in cases where the UB is audible, the harmonic structure contributed by the excitation in the UB region is less relevant than that in the TB, due to the poorer frequency resolution in the UB. Furthermore, in the literature, it has been reported that spectral folding, which is significantly less complex than pitch detection, produces a subjectively satisfactory approximation to the UB residual [32] [22].

Consequently, the simplified voiced excitation spectral model will incorporate the -12 dB/octave roll-off characteristic, however the harmonics and fine spectral structure will be generated primarily by spectral duplication. This method is discussed in Section 3.5.2.

Unvoiced Excitation

Unvoiced speech is usually associated with turbulent sources. For fricatives, the excitation is produced by turbulent air-flow caused by a constriction in the vocal tract. This excitation acts predominantly on the section of the tract following the constriction. Naturally, only the resonance behaviour of the section of tract acting on the excitation is of interest for the extrapolation process, and these resonances are the ones which will be encoded in the spectrum.

Strictly speaking, the spectrum of a turbulent excitation will depend on the degree and area of the constriction, and the shape of the orifice [31]. However, such turbulence

typically generates a noise-like signal, and the excitation can be assumed to have an essentially flat spectral trend over the WB frequency range of interest. Turbulent excitation can therefore be modeled as WB Gaussian noise with zero mean and variance corresponding to that of the true residual. This involves only a single parameter, variance, which reflects the energy/volume of the frame. The noise generation model is justified on the basis that the exact spectral detail of the excitation is not crucial perceptually in synthesis applications. This is presumably due to the fact that sounds generated with turbulent excitation have tract filters which emphasize predominantly the high frequencies, because the effective length of the tract is short between the supra-glottal excitation source and the mouth. Thus most of the signal content is concentrated at high frequencies, where frequency resolution is not particularly acute.

2.2 The Tract Filter

Air cavities within the vocal tract act as a multi-resonant filter on the vocal cord excitation [14]. Physically, the tract is an acoustic resonator which is a body of air which will resonate in response to sound stimuli containing frequencies which match the natural resonant frequencies of the volume of air. The term *formant*, encountered in Chapter 1, is defined in [29] to mean a normal mode of vibration of the vocal system[1]. The spectral properties of a resonance are summarized in Section 2.2.1.

From a complete physical perspective, the vocal system requires a daunting number of modeling parameters in terms of air flow, tract size and tract shape. A three dimensional mapping of the vocal cavities would be necessary for a complete analytical prediction of the corresponding speech wave [13, p. 85]. However, the shape of the air cavities is not as important as the volume [5]. Also, while variations in air-flow have a major effect upon the sound intensity they only have a negligible effect upon the spectrum of a sound produced. For these and other reasons, it was believed that a greatly simplified model might suffice to account for and extrapolate spectral speech characteristics.

From the PWB speech extrapolation perspective, there are three basic requirements of the tract model: (i) it must represent, albeit in a simplified form, the acoustic behaviour of the tract; (ii) in doing so, it must readily direct the extrapolation of resonances into the missing bands; and (iii) it must be parameterizable based strictly on an analysis of TB speech. Based on these criteria, the uniform tube model was selected. This model

[1] In this report, the word formant has been used more generally to denote spectral peaks such as those identified by LPA, which do not always coincide with tract resonances.

was selected over more complicated models because it was readily parameterizable from the TB speech, and it was not of interest to determine any exact geometrical knowledge of the tract for it's own sake, only to deduce, if possible any non-TB formant resonances which would be expected from the ensemble detected in the TB speech.

2.2.1 Transfer Function of a Resonance

As revealed in the example formant-scapes in Figure 1.6, the tract is a multi-resonant filter. Each resonance in isolation behaves essentially like a bandpass filter with an amplitude spectral contour like that depicted in Figure 2.3. The corresponding transfer function is given by [29].

$$T_1(j\omega) = \frac{s_1 s_1^*}{(j\omega - s_1)(jw - s_1^*)}. \tag{2.1}$$

where $s_1 = \sigma_1 + 2\pi F_1$, s_1^* is the complex conjugate pole pair, and σ_1 controls the roll-off which is related to physical dissipation. The entire formant-scape is produced by a cascade of such resonant filters, plus a filter describing radiative effects at the mouth. The transfer function of the tract is a product of the transfer functions in the cascade, so its spectrum results from the addition of their dB spectra. This accounts for why shifts in the frequencies of resonance, as caused by strategicly located constrictions in the tract, result in amplitude changes of the higher spectral peaks, depending on where on the tail of the lower resonant they occur.

As parametrized by the transfer function in (2.2), each resonance has two defining characteristics: resonant frequency, $F = \omega/2\pi$, and bandwidth, $B = F/2\pi$. In terms of f, the transfer function, $H(f)$, is given by [13]

$$H(f) = \frac{F^2 + (B/2)^2}{\sqrt{(f-F)^2 + (B/2)^2}\sqrt{(f+F)^2 + (B/2)^2}}. \tag{2.2}$$

The peak value of $|H(f = F)|$ is $Q = F/B$ [13, p. 54].

This frequency curve, displayed in Figure 2.3, is identical to that for the voltage transfer function of a series RLC circuit [13]. In considering such electrical models, sound pressure, ρ, is analogous to voltage and volume velocity, U, is analogous to current. This so called lumped model is of limited utility since it describes only a single resonance of the system.

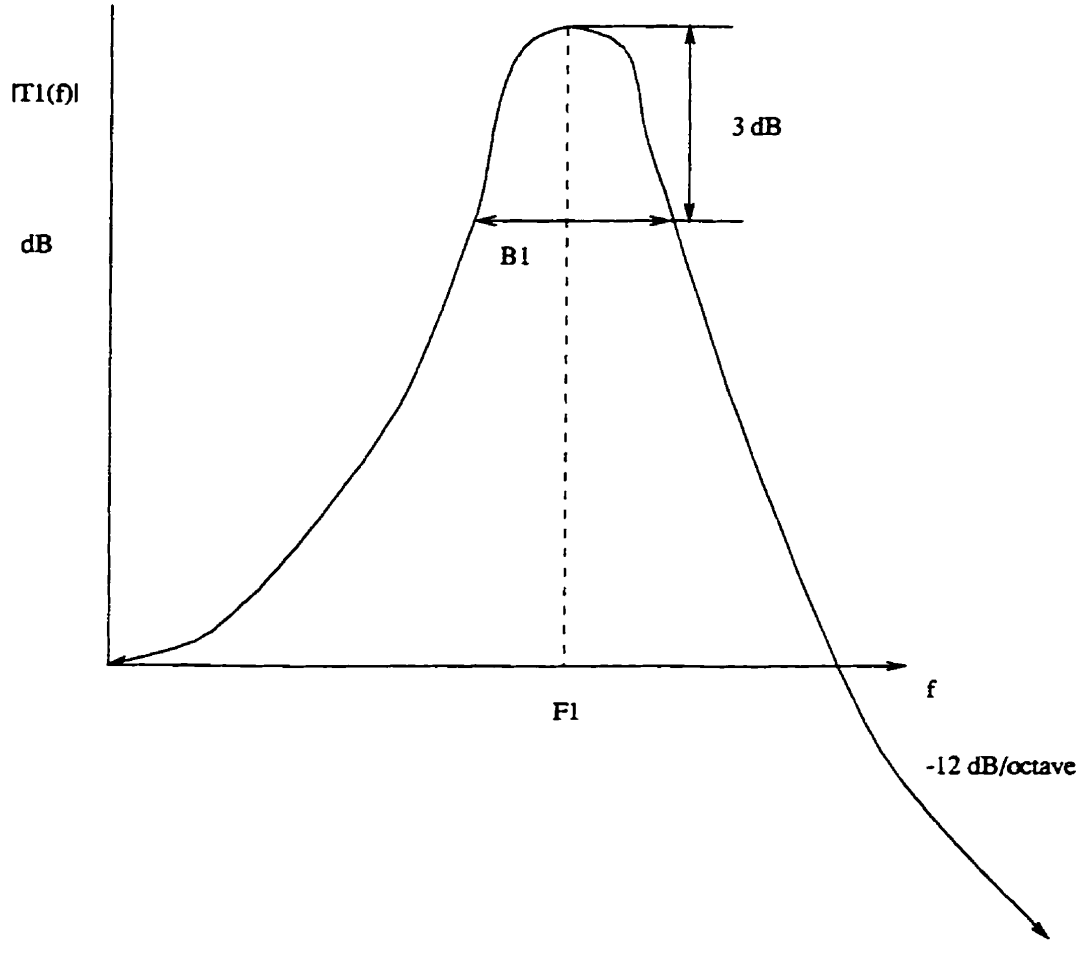


Figure 2.3: Spectrum of a single resonance.

2.2.2 The Uniform Tube Model

A popular physical model of the vocal tract is a cylindrical tube of uniform cross-section which is closed at one end, corresponding to the glottis/excitation, and open at the other, corresponding to the mouth. This model is especially applicable to vowel production, since the tract is relatively unconstricted and the glottal excitation is subject to the entire tract [5]. This model is characterized by a single parameter, length, L. Incidentally, to correspond to human tract, the diameter of the tube would be roughly 2 cm [24], however, this dimension does not affect the resonance behaviour of the tube, provided the diameter is appreciably less than the sound wavelengths[1]. Such a tube resonates at odd harmonics of a fundamental resonance frequency, $F1 = \frac{c}{4L}$, where $c \approx 340$ m/s is the speed of sound [24]. Thus, this tube would exhibit a static F-pattern $\{F1, F2 = 3F1, F3 = 5F1, F4 = 7F1 \ldots\}$, with a uniform spacing of $2F1$ between formants. The transfer function for such a tube would be of the form

$$T(s) = \frac{1}{\prod_{n\ odd}^{\infty}(1 - \frac{s}{s_n})(1 - \frac{s}{s_n^*})}. \tag{2.3}$$

which is effectively the cascade of an infinite number of the resonances described in (2.1). Thus, the uniform tube model yields an all-pole spectral envelope, and is fully compatible with the LPA and LPS techniques.

Figure 2.4 depicts the first four odd-harmonic resonant modes for a tube closed at one end[2]. The modes are maintained by sound wave reflections at the ends of the pipe[3]. The letters 'P' and 'V' denote the respective regions of maximum sound pressure and volume velocity for the different modes.

Until this point, the discussion has not considered resonance bandwidths. The formant-scape of the uniform tube would have infinite peaks at resonant frequencies for the ideal lossless case, but exhibits peaks of finite amplitudes for the actual dissipative case [29]. According to [23, p. 225], formant bandwidths can range from 30 Hz to 500 Hz, but this data, like most in the literature, probably applies to narrowband speech. In [29] and [13], $B = 100$ Hz for all formants was used in most of the examples, and was justified

[1] This condition guarantees planar propagation of the sound waves in the tube, and certainly applies for frequencies below 4000 Hz, and hopefully also to those below 8000 Hz which have wavelengths over twice the tract diameter.

[2] Note that speech is a complex tone, so that these resonances occur simultaneously.

[3] At the closed end, the reflection entails no change of phase, while at the open end there is a change of phase. Furthermore, the factor of the amount dissipated is $\frac{8\pi^2 R^2}{\lambda^2}$, if the ratio of tube radius, R, to wavelength, λ, is small [30].

by the fact that this is the approximate bandwidth of $F1$, $F2$, and $F3$ in spoken vowels. For the UB, for both voiced and unvoiced speech, formant bandwidths in the simulation corpus exhibited an average bandwidth of 763 Hz (which corresponds to a z-pole magnitude of 0.8614).

This single tube model provides insight on certain properties of speech and their relation to the physical production process. The average length of a male vocal tract is 17 cm [23, p.181] [15, p. 24], which corresponds to $F1_M = 500$ Hz, and a formant separation of $\Delta F_M = 2F1_M = 1000$ Hz. This consistent with the order of the formant separation experimentally observed for male speech [14], and similar results apply for female speech. The average tract lengths of 17 cm for men and 13 cm for women lead to an expectation of 7 or 8 formants in the wideband speech range, which justifies the fact that 16^{th} order LPA (which identifies 8 spectral peaks) is the most popular choice in the literature for wideband speech analysis. The observation in [14] that an increase in the length of the lip passage causes a lowering of all frequencies is also consistent with this model, since a lengthening of the vocal tract would decrease $F1$, resulting in a decrease in the harmonic formant frequencies, and additionally, a decrease in the separation between formants. From the latter example, it is clear that, not only does the effective length of the tract vary from speaker to speaker, but also within a speaker over time due to articulatory dynamics.

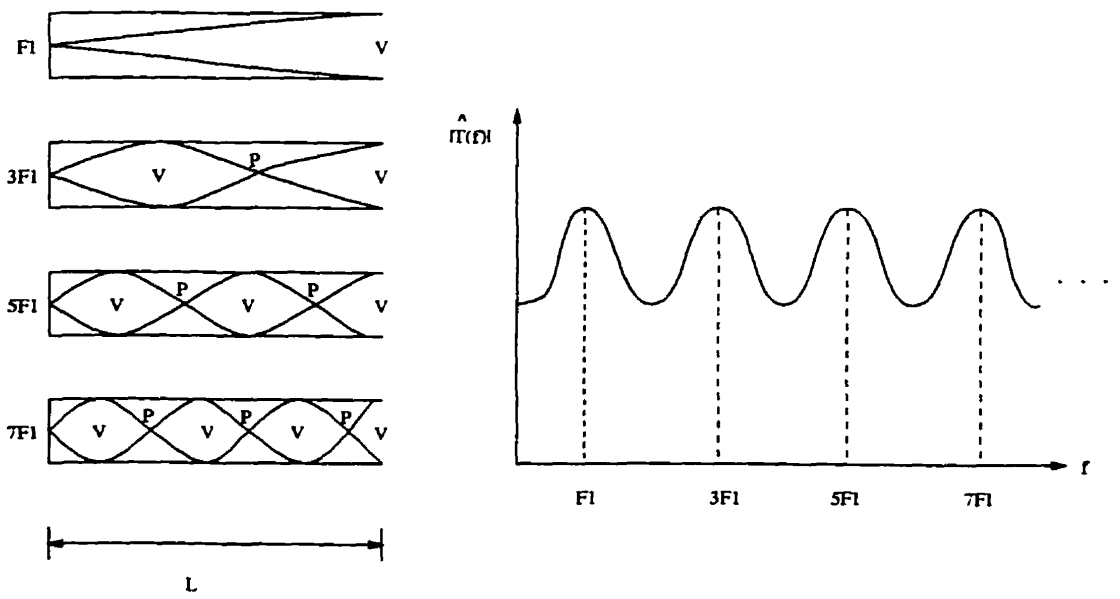


Figure 2.4: Odd-harmonic resonances produced in a tube closed at one end.

For vowels, in a real static tract, elastic properties, constrictions and contortions perturb the odd-harmonic progression of the F-pattern suggested by the uniform tube model. These perturbations can have marked effects of the relative amplitudes of the spectral resonance peaks, and this phenomena is described in Section 2.3. Each of the resonances depicted in Figure 2.4 are particularly sensitive to constrictions in the tract which coincide with their P and V regions [5]. In particular, constrictions at a point of maximum velocity, tend to lower the associated resonant frequency, while constrictions at a point of maximum pressure tend to raise it [14]. It is noteworthy that the mouth (open end of tube) is a point of maximum velocity for all odd-harmonic resonances. Since mouth constrictions affect all harmonics, such an effect could be modeled as an effective increase in tract length.

Despite these F-pattern perturbations, statistically, their average spacings will not change provided the total vocal tract length does not change [29] [13, p. 61]. This suggests that total effective tract length, L_{eff}, is a dominant parameter of the system. Thus, $F1_{eff}$ can be found from an infinite series of actual tract resonances according to

$$F1_{eff} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \frac{F_{i+1} - F_i}{2}. \tag{2.4}$$

This suggests the means by which to parametrize the model and apply it to extrapolation.

In the preceding discussion of the speech production mechanisms, the analysis was presented looking from the inside out. That is, from an acoustic cause to a spectral effect. In applying the model to PWB extrapolation, the first stage in the process is the analysis of the TB speech to parameterize the tract model. Then, the model can be used to extrapolate out-of-band spectral effects. This process is outlined in the following sections.

2.2.3 Estimation of Tract Length from TB speech

To deduce the 'instantaneous' tract length for a current frame, it is necessary to identify the tract resonances within the TB and calculate the average separation between them. Identification of tract resonances from a segment of speech is an inherently difficult problem. However, when the resonances are reasonably well separated in frequency and when their bandwidths are not abnormally wide, the frequencies of the spectral maxima, as identified by all-pole modelling such as LPA, are good measures of the resonant frequencies [23].

Thus, formants found via TB-LPA can serve as estimates of the tract resonance

frequencies. Appropriate order of LPA is important to ensure that all true resonances in the TB are identified, but no extra 'resonances' are identified which are actually pitch harmonics or other excitation artifacts. The choice of $p = 6$, identifying three resonance frequencies, is a logical one because three formants can be expected in the telephone band for most sounds and speakers. This is verified by the fact that the expected frequency ranges for the first three formants are $F1\epsilon[100, 1000]$, $F2\epsilon[700, 2500]$, and $F3\epsilon[1500, 3500]$ [24, p. 93].

2.2.4 Limitations of the Uniform Tube Model

The uniform tube model will automatically yield uniformly spaced formants, however, actual speech does not exhibit such regular formants-scapes. At best, the model can only approximate the true F-pattern with an optimal uniformly spaced approximation. Similarly, the bandwidths of actual UB formants are not all equal. Although these limitations are inherent to the model, it is possible that these inaccuracies are perceptually tolerable. Perceptual considerations are outlined in Section 2.2.5.

For some sounds the uniform tube is a better approximation to tract configuration than others. For example, for nasal sounds, there is a coupling between the tract and nasal passages, resulting in a different acoustic behaviour. Also, for unvoiced sounds, the excitation acts primarily on the front cavity section of tract (including the narrow passage at the articulatory constriction). There does, however, exist some coupling between this front section and the back section between the glottis and the constriction. Unfortunately, the degree of coupling to the back cavities increases with increasing frequency [13]. Nonetheless, the main shape of the vocal tract filter function (ie. the TB observable spectral envelope) may still be essentially conditioned by the front cavity resonances. Thus, as for parameterizing the tract length for vowels, the average frequency spacing between formants can be used to deduce the length of the front part of the tract.

To more rigorously quantify the effects of tract constrictions on the resonance behaviour, in some studies, the tract has been modeled physically as a set of concatenated tubes of different lengths and diameters [12]. Independently, these tubes have resonance properties, but the extent to which a given *system* resonance is affiliated with a particular cavity depends on the amount of coupling between the cavities, which in turn is dependent on the dimensions of the constrictions [29]. Mapping from such a geometric model to a spectrum can be fairly complex. Parameterizing such a complex model solely from a TB signal would be prohibitively complex if not impossible. All parts of the vocal cavities

have some influence on all formants and each formant is dependent on the entire shape of the complete system [14]. It is for these reasons that such models were excluded from preliminary consideration for this study.

2.2.5 Perceptual Considerations for the Tract Model

Naturally, a distortion is only perceptually significant if the magnitude of the distortion exceeds the resolution of the human auditory system [26]. Conveniently, certain perceptual sensitivities and resolutions decrease with increasing frequency. Some results concerning the effects of predictor coefficient quantization errors in LP speech coding are of particular interest. In [24] it was reported that the minimum change detectable for formant frequency is 5% while for formant bandwidth it is about 40%. This implies an encouraging amount of leeway for UB speech generation. Figure 2.5 provides a tolerance guideline for formant frequency errors as a function of frequency, although admittedly, compound errors could have exaggerated effects.

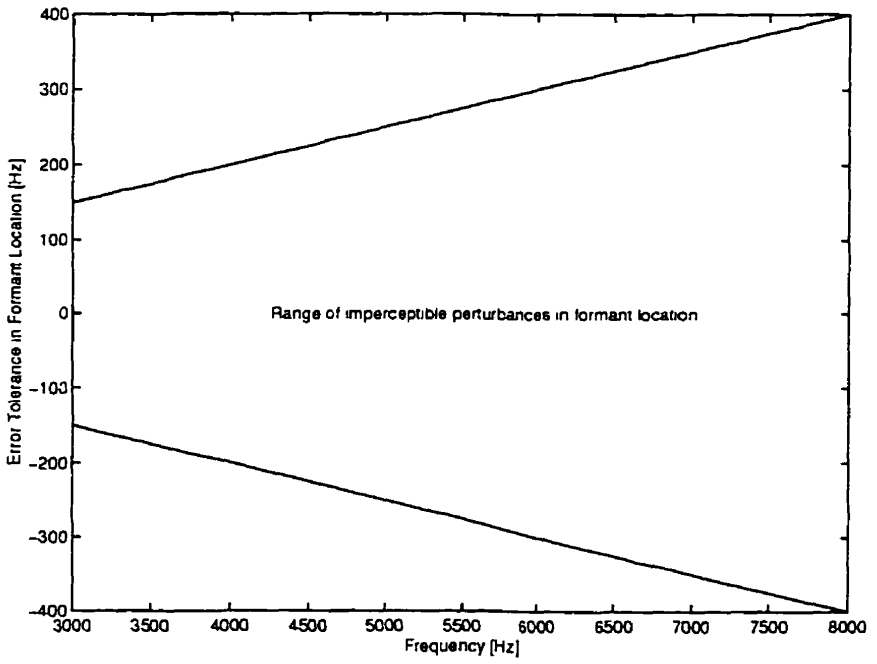


Figure 2.5: Tolerance guideline of just noticeable differences in formant location and bandwidth as a function of frequency.

2.3 The Entire Speech Spectrum

From the point of view of the perceived speech, the tract influences don't occur in a vacuum. The final speech spectrum, as perceived or recorded by a microphone external to the mouth is the result of several influences. In particular, the speech spectrum, $S(f)$ can be expressed

$$S(f) = U(f)T(f)M(f), \tag{2.5}$$

where $U(f)$ is the spectrum of the excitation source, $T(f)$ defines the resonance properties of the tract which have already been discussed above, and $M(f)$ is the spectrum defining radiative effects from the mouth. For voiced speech, $U(f)$, represents the spectrum of the volume velocity delivered by the vocal chords, which has been seen to have a spectral trend of -12 dB/octave (this roll-off actually begins after 100 Hz [13, p. 51]).

At the mouth opening, the effects of acoustic radiation impedance come into play. The spectral effects of radiation from the mouth can be modelled as a first order high-pass filter, increasing at a rate of 6 dB/octave in the range up to 3000 Hz [24].

The tract response itself can be viewed as a cascade of its individual resonance transfer functions, $T_i(f)$, of the type described by (2.2).

$$T(f) = T_1(f)T_2(f)T_3(f)\ldots \tag{2.6}$$

This corresponds to summing the individual contours of the dB spectra.

When the resonances of $T(f)$ deviate from the odd-harmonic pattern for the uniform tube, the final relative amplitudes of the spectral peaks can vary markedly depending on the frequencies of the resonances. As the position along the frequency axis of a resonance is shifted with respect to a lower one, it will be superimposed on a different section of the lower resonance's tail, raising or boosting the higher resonance's spectral peak. The amplitude of the spectral peak of the lower resonance is also effected, but to a lesser extent, since the resonance roll-off approaches 0 dB quickly on the side of the resonance closer to 0 Hz.

This raises an important issue for the synthesis of bandlimited speech. It might not be sufficient merely to model the tract by including resonances that fall within the band of interest. An additional, compensating contour may be required to account for

'fore shadowing' effects of resonances above the band of interest. To correct for this effect a function, $K(f)$, is desired. One such correction factor suggested to correct for formants above the fifth is [13]

$$|K(f)|_{dB} = 0.54(\frac{F1}{f})^2 + 0.00143(\frac{F1}{f})^4. \tag{2.7}$$

2.4 Summary

A speech spectrum is the result of the combined influences of the excitation, the resonance properties of the tract, and radiative effects which occur when the volume velocity signal exits the mouth.

The magnitude spectra of voiced frames generally tend to slope downward over the the WB region and those of unvoiced frames tend to slope upwards. It is important to note, however, that the former phenomena stems from a roll-off in the spectrum of the voiced *excitation source*, while the latter is due, not to the excitation spectrum, but to the shorter effective portion of the *tract* to which it is applied by virtue of it's location. A shorter tract will yield higher resonance frequencies, as noted by blowing in a bottle with and without water in it. So, although the unvoiced excitation is flat, it's low frequency components are often damped by the tract.

The tract and excitation features discussed illuminated an acoustic basis for the relative independence of the speech excitation and envelope. For a given frame, the F-pattern frequencies can be viewed as natural frequencies of free vibration for that tract configuration, and they are independent of how the system is excited. Similarly, due to high 'impedance', the excitation signal is relatively independent of the tract configuration.

The resonance properties of the tract can be quite complex[1]. For vowels, it is reasonable to consider the vocal tract as a single tube, whereas for some constants, a more complex transmission model might be warranted [5]. Nonetheless, the uniform tube model was shown to possess significant merit, both in terms of it's ability to account for observed spectral characteristics and the ease with which it allows the extrapolation of out-of-band components. It can be parametrized by conducting LPA on the TB speech. Based on considerations of complexity, perceptual factors, and speaker and sound independence, the uniform tube model was selected for UB PWB extrapolation.

[1] Even when the cross-sectional area of the vocal tract is known at all points along its length, and exact calculations of resonance frequencies can be made but they are complex [29]

The PWB system has two potential Achilles heels. The TB may be insufficient to properly parameterize the otherwise adequate (from the point of view of extrapolation capability) model. Or, the model itself might be too simplistic, yielding evenly spaced formants, to produce perceptually satisfactory approximations.

Chapter 3

Proposed PWB Speech Extrapolation Algorithm

In this chapter the proposed PWB speech generation algorithm is described. The purpose of the algorithm is to employ the production models described in Chapter 2 to permit extrapolation of the frequency content from the TB into non-TB ranges. The algorithm is primarily focussed to extrapolate into the UB rather than the sub-TB range. Based on results outlined in Chapter 4, the UB yields a better gain in terms of objective measures than the sub-TB range. As mentioned in Chapter 1, the supra-TB range is thought to contribute to increased intelligibility, sound differentiation and crispness [7].

Section 3.1 outlines the key assumptions inherent in the design of the algorithm. An overview of the whole system is presented in Section 3.2. The subsequent sections discuss the implementation details of the various operational blocks of the system.

3.1 Design Assumptions

Several assumptions are inherent in the modeling and design philosophy used for this algorithm. Foremost is the assumption that the telephone band contains sufficient information that a perceptually viable wideband version can be extrapolated. This assumption appears validated by the promising preliminary results reported in [1], [2], [6] and [8].

More to the point, it is assumed that the correlation structure can be adequately captured according to the extrapolation model described in Chapter 2. As detailed in Section 2.2.3, it is assumed that the first three resonances are available in the TB. This indicates that the primary application of the model will be to extrapolate resonances into

the UB.

Frames are regarded independently, so no inter-frame time correlations were investigated or employed in the algorithm. This decision was based on the desire to avoid error propagation between frames and to minimize the operational delay of the algorithm to make it amenable for real time implementation. This decision was also incorporated into the previous spectral codebook algorithms [1] [6] and the Statistical Recovery Function [8] discussed in Section 1.6. Thus it is assumed that inter-frame time correlations can safely be neglected, and that all necessary information for the wideband extrapolation is contained within the current telephone band.

It is also assumed that the excitation and envelope components of speech produced by LPA can be treated independently, and that any correlation between these components is insignificant for the current application. This tactic was used in the previous PWB methods discussed in Section 1.6 since it reduces the dimensionality of the mapping, and allows the exploitation of the powerful techniques of linear predictive analysis and synthesis. Furthermore, as discussed in Chapter 2, this assumption is physically justified.

Finally, it is assumed that the input TB speech is ideal within the TB range of frequencies. Consequently, the PWB speech contains a faithful reproduction of the input TB signal in that range.

3.2 System Overview

A high-level block diagram of the proposed PWB speech extrapolation system is presented in Figure 3.1. Due to the dynamic nature of speech, time segmentation or *framing* is the first order of business to prepare the signal for LPA. All processing in the system is conducted on a frame-by-frame basis. Once the speech is framed for block processing, there are essentially three operational phases: analysis; extrapolation; and synthesis.

Details of the framing procedure are presented in Section 3.3. During the analysis phase a TB frame is classified as voiced or unvoiced, and partitioned into envelope and excitation via LPA. These procedures are discussed in Section 3.4. In the extrapolation phase, tract resonance extrapolation and residual extrapolation are conducted in accordance with the models discussed in Chapter 2. This extrapolation is geared specifically to reproduce the UB. The extrapolation mechanisms are discussed in Section 3.5. Finally, in the synthesis phase, LPS is conducted using the extrapolated residual and tract filter. Any corrective spectral shaping contour is then applied. Finally, since it is assumed ideal, the TB frame is interpolated and spliced into the synthetic WB signal. These synthesis

steps are described in Section 3.6.

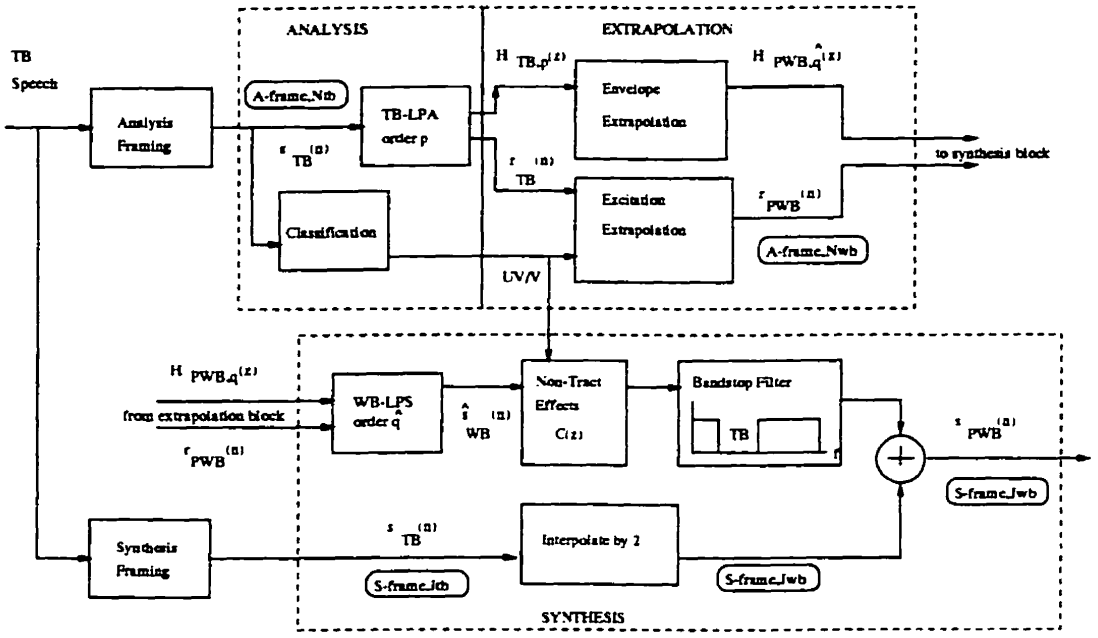


Figure 3.1: High-level block diagram of the proposed PWB speech extrapolation system.

3.3 Framing for Block Processing

Figure 3.2 depicts the framing used for linear predictive analysis and synthesis. The input TB speech is partitioned into overlapping analysis frames of 20 ms duration ($N_{tb} = 160$ samples at 8000 samples per second). As discussed previously, this is the duration of time over which speech can be assumed to be approximately stationary, and thus have a meaningful spectrum whose envelope contours can be deduced with LPA[1]. A frame advance rate of 10.5 ms was used, which is consequently the length of each synthesis frame, and corresponds to $J_{wb} = 168$ samples at 16 kHz. This duration was chosen based on precedents in the literature, and an awareness of the tradeoff between the additional computational complexity incurred with numerous short synthesis (and analysis) frames and the better time resolution and implicit inter frame smoothing they yield.

The framing mechanism described touches upon the issue of inter frame correlation. Analysis frames are overlapping while synthesis frames are not. Each analysis frame is

[1]For speech applications in the literature, LPA frame durations were typically in the range of 20 ms to 25 ms.

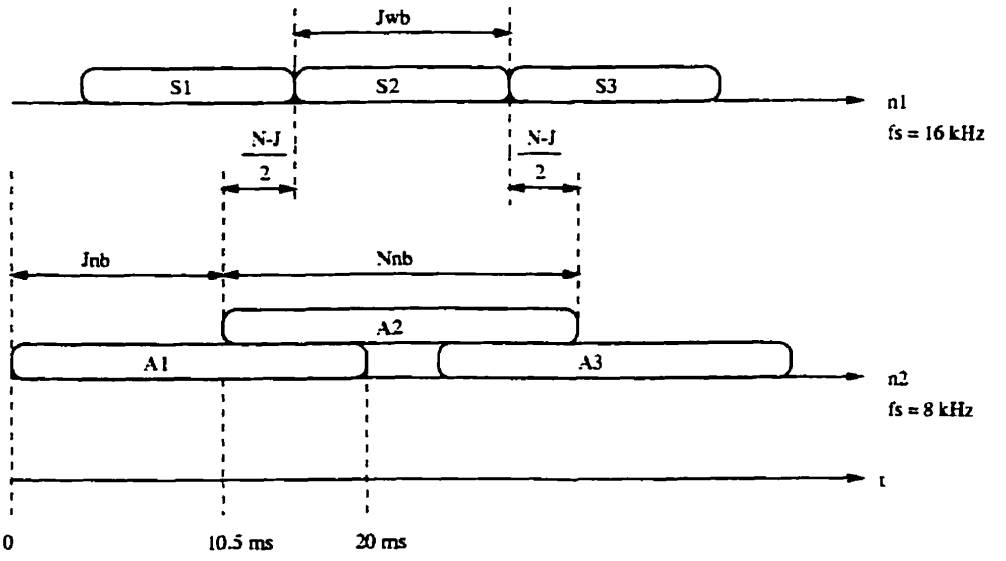


Figure 3.2: Relationships between the analysis and synthesis frames used for LPA and LPS.

windowed with a Hamming window prior to LPA so that the characteristics of the synthesis portion of the frame receive the most weight. Such windowing reduces the significance of samples at the frame edges, so large, overlapping analysis frames are employed to ensure all speech events receive their due attention. The larger analysis time window permits a better averaging of the correlation characteristics. Thus, although inter frame correlations are not explicitly considered in the extrapolation process *for the non-TB frequency regions*, the overlapping TB analysis frames smooth the $H_{TB}(z)$ transitions, which in turn yields closer F-patterns between adjacent frames for the TB, and consequently in the extrapolated range as well.

3.4 Analysis

3.4.1 Linear Predictive Analysis (TB-LPA)

The innards of the TB-LPA block are revealed in Figure 3.3. The basics of the LPA of speech were reviewed in Section 1.4. In this section, implementation details of the process are provided, as well as the parameter values used in this application.

For each analysis frame there are two outputs of the TB-LPA block, a spectral envelope, $H_{TB}(z)$, and a residual, $r_{TB}(n)$. The envelope becomes the input for the tract resonance identification block, while the residual is the input for the excitation extrapo-

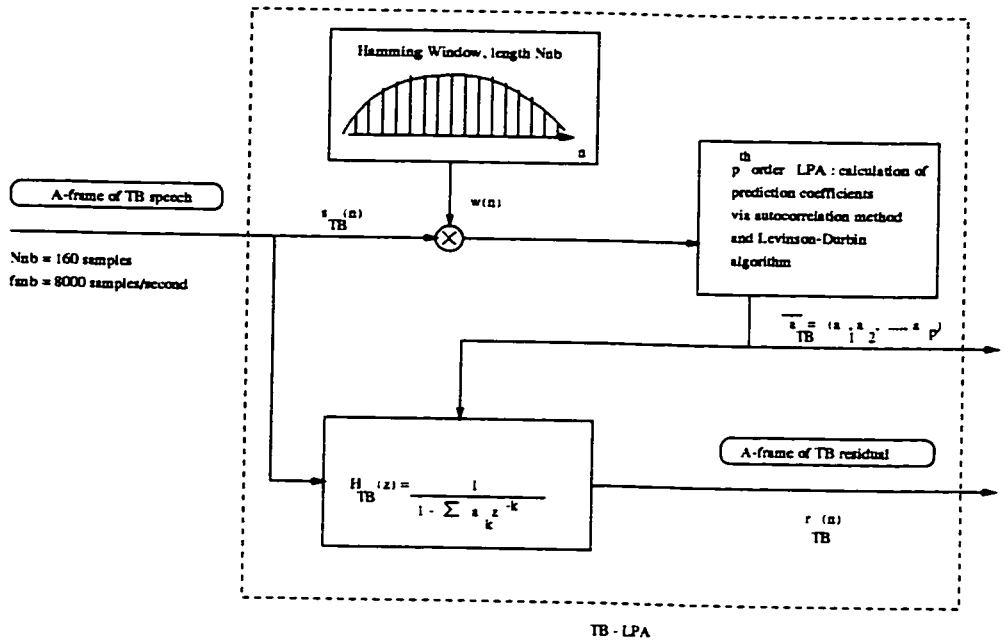


Figure 3.3: Block Diagram of TB-LPA.

lation block. The all-pole envelope is parameterized by the p predictor coefficients and is of the form

$$H_{TB}(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{3.1}$$

The autocorrelation method for calculating these predictor coefficients from the analysis frame is detailed in a subsection below.

The residual, $r_{TB}(n)$, a time sequence of length $N_{tb} = 160$ at the sampling rate of 8000 Hz, is generated according to Figure 3.3, by applying the analysis frame of speech, without the Hamming window weighings, to the filter $H_{TB}(z)$. The un-windowed speech is used in order that the output of the filter better reflects the energy content of the speech frame. This is important for the excitation extrapolation phase discussed in Section 3.5.2. Although in synthesis, only the non-overlapping duration of the residual is necessary, producing a residual of analysis frame length helps smooth frame edge effects, by allowing the synthesis filter of each frame to be primed/initialized with a sequence of previous samples. This is useful, since the synthesis filters are of different shapes and orders, which is not conducive to state initialization in the usual sense.

The order of TB-LPA is $p = 6$, and thus identifies three formants, which, as

discussed in Section 2.2.3 is the typical number to be found in the TB range. High order LPA would result in misidentifying voicing pitch harmonics as tract resonances.

Autocorrelation Method

The computation of the predictor coefficients, a_k, was performed using the autocorrelation method of LPA. Inspection of (3.2), which is a reiteration of (1.5), reveals that the summations over the frame are reminiscent of autocorrelation calculations.

$$\sum_{k=1}^{p} a_k \sum_{n=1}^{N} s_{n-i} s_{n-k} = \sum_{n=1}^{N} s_{n-i} s_n. \tag{3.2}$$

with $1 \leq i \leq p$. It is also noteworthy the calculations depend not only on samples within the current frame (for simplicity numbered s_1 to s_N), but also on p samples preceding the current with frame, to accommodate the ranges of i and k and ensure that the autocorrelation computations are all based on N pairs of samples. In the autocorrelation method, the frame is windowed with a cosine window (Hamming), so samples outside the frame are presumed to be zero, and samples near the edge are less potent. This has the effect that

$$\sum_{n=1}^{N} s_{n-i} s_{n-k} \approx \sum_{n=1}^{N} s_{n-(i+1)} s_{n-(k+1)}. \tag{3.3}$$

The computation on the right will be based on one less non-zeroed product pair than that on the left. The maximum discrepancy in number of product terms is p fewer pairs for the ranges of i and k, but since $N \gg p$, and the additional pairs include frame edge points which are severely attenuated by the Hamming window, (3.3) can be assumed true and equal to an autocorrelation function $R(i-k)$ as defined by[1]

$$R(i-k) = \sum_{n=1}^{N} s_n s_{i-k}. \tag{3.4}$$

The the system described by (3.2) becomes, in matrix form

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(2) & R(3) & \dots & R(p-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(0) \\ R(1) \\ \vdots \\ R(p-1) \end{bmatrix} \tag{3.5}$$

[1] The autocorrelation function $R(i-k)$ could also be normalized by N as in [27].

Levinson-Durbin Algorithm

The leftmost matrix in (3.5) is a Toeplitz matrix, in which the elements on each diagonal are equal, and for which an efficient method for inversion exists. In particular, the a_k's can be solved by using the Levinson-Durbin recursive algorithm which is described in pseudo code below [27].

$E = R(0)$
$k_1 = R(1)/E$
$c_1(1) = k_1$
loop j from 2 to p
    $k_j = \frac{R(j)-\sum_{m=1}^{j-1} c_{j-1}(m)R(j-m)}{E}$
    $c_j(j) = k_j$
    loop i from 1 to $j-1$
        $c_j(i) = c_{j-1}(i) - k_j c_{j-1}(j-i)$
        $E = E(1-k_{j-1}^2)$
    end i loop
end j loop
loop k from 1 to p
    $a_k = c_p(k)$
end k loop

E is the error term, k's are known as reflection coefficients, and $c_j(i)$ is the j^{th} predictor coefficient for an i^{th} order LPA. The final predictor coefficients for the p^{th} order LPA are denoted, as usual, a_k for $1 \leq k \leq p$.

3.4.2 Frame Classification

The classification block classifies the speech as voiced or unvoiced. At the synthesis stage, this determines which excitation extrapolation technique to employ and is used as a flag to apply a voiced spectral roll-off trend to the UB spectrum of voiced frames.

Using WB speech, a simple approximate method for V/UV classification was to classify as UV those frames with a greater proportion of their energy above 3000 Hz than below it. To ascertain the appropriate speech class for a frame of TB speech, the slope of the LPA envelope was used. It may be recalled that LPA incorporates the +6 dB/octave radiation effect and the -12 dB/octave voiced excitation roll-off, into the envelope, $H_{TB}(z)$.

Thus it is possible, by examining the slope of the spectrum, to determine if the frame is voiced (negative slope of approximately 6 dB/octave) or unvoiced (positive slope).

3.5 Extrapolation

3.5.1 Odd-Harmonic Tract Resonance Extrapolation

The input to this block is the TB envelope, $H_{TB,p}$ (or equivalently the set of predictor coefficients, $\bar{a}_{TB}$), and the output is an all-pole PWB tract filter, $T_{PWB,\hat{q}}$.

Based on the uniform tube model, it is desired to parameterize the tract by estimating the effective length, L_{eff} (or $F1_{eff} = c/4L_{eff}$), from the TB resonances. To construct such an estimate, it should be recalled that, statistically, the average spacing between formants depends only on tract length. The TB F-pattern is calculated by finding the poles of the envelope. Each complex conjugate pole pair contributes one positive resonance frequency, F_i, according to

$$F_i = \frac{phase(z_i)}{T_s \pi}. \tag{3.6}$$

where T_s is the sampling interval for TB speech, which is 0.125 ms (which corresponds to 8000 samples per second).

Assuming that the TB formants, $F1_{TB}$, $F2_{TB}$, and $F3_{TB}$ correspond to the first three tract resonance frequencies, an estimate of $F1_{eff}$ can be made according to

$$F1_{eff} = \frac{1}{3}(F1_{TB} + \frac{(F2_{TB} - F1_{TB})}{2} + \frac{(F3_{TB} - F2_{TB})}{2}). \tag{3.7}$$

Having estimated $F1_{eff}$, it is a trivial matter to extrapolate odd-harmonic resonances according to

$$\hat{F}_i = iF1_{eff}, \tag{3.8}$$

where i is an odd integer such that $3300 \leq iF1_{eff} \leq 8000$. Depending on $F1_{eff}$, the WB might accommodate a different number of formant harmonics. The implication for synthesis is that the order, $\hat{q}$, of the envelope filter can vary from frame to frame, depending on the number of extrapolated 'formant resonances' which fall within the WB range.

As discussed in Section 2.2.1, the contribution of each resonance to the tract transfer function can be parameterized in terms of two quantities: frequency, F_i, and bandwidth,

B_i. The bandwidth of each extrapolated formant was set to 763 Hz (which corresponds to a z-pole magnitude of 0.8614), since this was found to be the average bandwidth of UB formants in the speech corpus used in the simulations. As mentioned in Section 2.2.5, there is a relatively high tolerance for formant bandwidth errors in the UB. Furthermore there is a precedence in the literature for using fixed bandwidths for high frequency formants [23].

Each of the $\hat{q}/2$ resonances contributes a conjugate pole pair (z_i, z_i^*) of the form

$$z_i = 0.8614e^{j2\pi F_i}. \tag{3.9}$$

The final PWB tract filter, $T_{PWB}(z)$, is thus constructed according to

$$T_{PWB}(z) = \frac{K_{\frac{\hat{q}}{2}}(z)}{\prod_{i=1}^{\frac{\hat{q}}{2}}(z - z_i)(z - z_i^*)}, \tag{3.10}$$

where $K_{\frac{\hat{q}}{2}}(z)$ is a correction factor to account for the non-localized 'fore shadowing' effects of resonances whose centre frequencies lie above the WB. As a first approximation, $K_{\frac{\hat{q}}{2}}(z)$ was assigned to be a resonance at 8 kHz, however, it was ultimately decided that assigning the correction factor to unity yeilded better results.

Since it is assumed above that the first tract resonance is already contained in the TB, this technique does not generate sub-TB resonance features.

3.5.2 Excitation Extrapolation

The innards of the excitation extrapolation block are presented in Figure 3.4. The process generates a wideband excitation frame, $r_{PWB}(n)$, from the input residual, $r_{TB}(n)$. As indicated, the details of the extrapolation technique depend on whether the frame is voiced or unvoiced. The technique for voiced excitation extrapolation was devised by the author.

For unvoiced frames, wideband white noise is used for r_{PWB}. This technique is prevalent in the literature, and was found to yield adequate subjective and objective results as described in Section 4.4.2. The basis for the noise signal was produced with a random number generator implemented with the *randn* function in MATLAB which produces random numbers according to the Gaussian probability distribution function, $N(0, 1)$. This noise signal was then adjusted to have the mean and variance of the actual TB residual[1].

[1] Actually, the mean of the residual is approximately zero anyway [21].

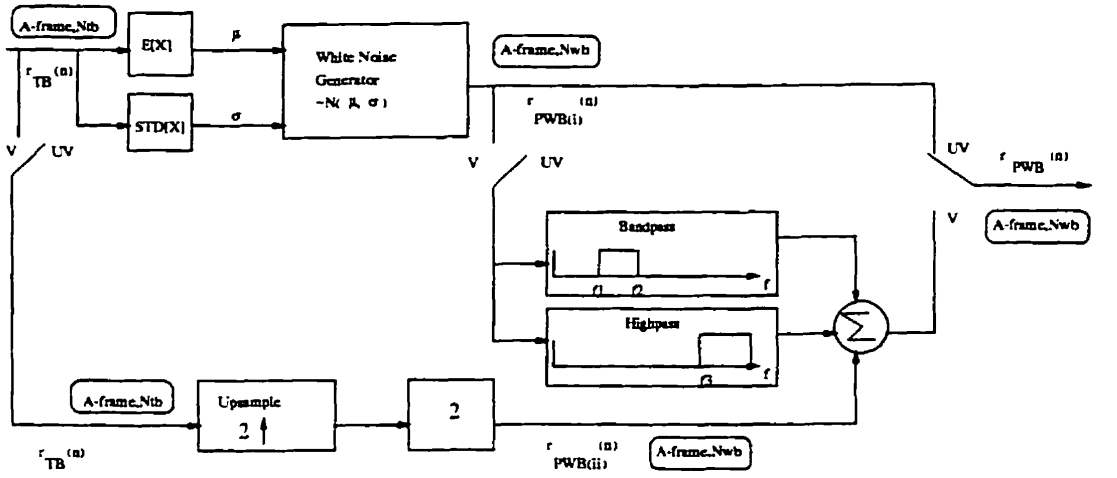


Figure 3.4: Block diagram of the excitation extrapolation technique.

A different technique was necessary for voiced excitation, since it was found that modeling the missing bands of voiced excitation with white noise produced artifacts. The usual alternative of spectral folding results in spectral gaps when applied to the TB bandlimited residual. The author found that filling the gaps with suitably scaled white noise (as for the unvoiced excitation) yielded satisfactory results. Spectral folding is accomplished by upsampling r_{TB} to 16 kHz by inserting a zero-valued sample after each sample. A gain factor of 2 is necessary to preserve the short-term average power of the baseband and extrapolated residuals [32]. Tenth order Chebychev type I filters were used extract the gap noise from the wideband noise using cutoff frequencies $f1 = 3300$ Hz, $f2 = 4700$ Hz, and $f3 = 7700$ Hz.

Figure 3.5 presents examples of the $|r_{PWB}(f)|$ generated for typical voiced and unvoiced frames. It is noteworthy that, even in the voiced case, the excitations exhibit flat spectral trends. The residual, $r_{TB}(n)$, yielded by the TB-LPA block is spectrally flat, because the major contours are extracted as the envelope. For the PWB case, all major contours, including voiced spectral roll-off will also be included in the envelope during the synthesis phase as described in Section 3.6. Therefore, strictly speaking, this block performs *residual* extrapolation rather that *excitation* extrapolation.

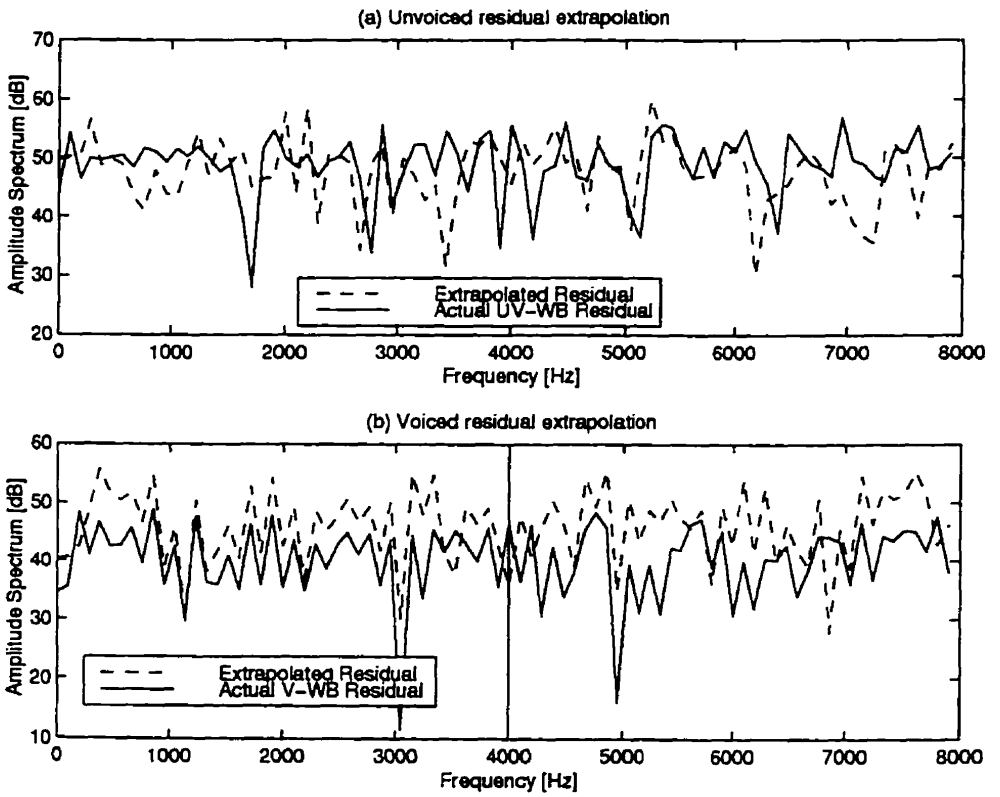


Figure 3.5: Examples of actual and extrapolated wideband residuals for (a) an unvoiced frame; (b) a voiced frame.

3.6 Synthesis

The synthesis block has four inputs: the extrapolated tract transfer function, $T_{PWB}(z)$; the extrapolated residual, r_{PWB}; the frame's classification (V/UV); and the original TB frame. The synthesis proceeds in three phases. The first phase is $\hat{q}_{th}$ order WB-LPS whereby the extrapolated residual is modulated by the extrapolated tract transfer function. The second phase involves a spectral contour correction filter, $C(z)$. This is included to model non-tract spectral effects on the envelope such as the radiative effects from the sound leaving the mouth and the -12 dB/octave harmonic attenuation of the voiced excitation. In the final phase of synthesis, the TB speech is interpolated and spliced into the synthetic signal.

3.6.1 Correction Filter

For voiced speech, a correction filter specified by

$$C_V(z) = \frac{1}{1 - 0.96z^{-1}} \tag{3.11}$$

was used. This contour is depicted in Figure 3.6. This helps generate the roll-off which was stripped from the voiced residuals.

For unvoiced speech, it was found that the UV-UB spectral slope exhibited a great deal of variance. Since no single, acceptable correction filter could be found, nominally, $C_{UV}(z) = 1$.

3.6.2 Splicing the TB into the WB Synthetic Signal

Based on the telephone band speech model described in Section 4.2, it is assumed that the input TB signal is effectively ideal within the telephone band (despite actual non-idealities in the passband filter). This assumption was also made in the PWB algorithms described in [1], [2], [6], and [8], and is the motivation for interpolating and splicing the TB signal into the final PWB synthetic signal. The goal of the PWB algorithm is not to enhance the signal within the TB, but to enhance it by introducing frequency components outside the TB.

The cutoff frequencies for the 10^{th} order Chebychev type I splicing filter depicted in Figure 3.1 are 265 Hz and 3380 Hz. These values were chosen based on a simulation described in Section 4.4.1. The TB signal is interpolated by a factor of two and and added to the synthetic signal.

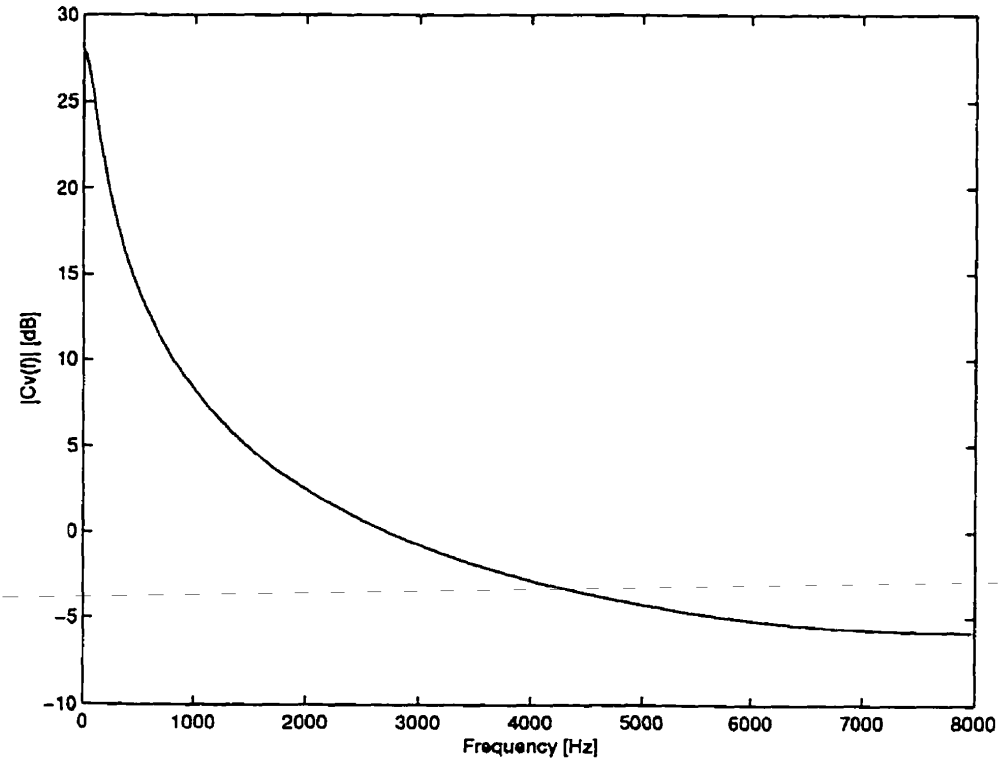


Figure 3.6: Contour of the Voiced Spectral Shaping Filter, $-C_V(f)-$.

3.7 Complexity

A detailed investigation of the complexity of the algorithm was not undertaken, however, some general comments can be made. The LPA and LPS processes are not computationally intensive. They are comparable, in terms of order and frame rate, to those used in modern speech codecs, and implementable in real-time. Similarly, the excitation extrapolation techniques have been employed in RELP coders of reasonable complexity [32]. A potentially expensive computation is the root finding computation to determine TB resonance frequencies.

3.8 Summary

As with the methods discussed in Section 1.6, the techniques of LPA and LPS are cornerstones of the system. TB-LPA does double duty in the front end of the system since it partitions the speech in envelope and excitation, and does so in such a way as to permit the ready idenification of the TB formant locations.

In the envelope extrapolation block, the TB tract resonances identified are then used to parameterize the acoustic tract model. Excitation extrapolation is also conducted on the TB-residual, r_{TB}, to generate a PWB-excitation, r_{PWB}. The manner in which this extrapolation is done depends on whether the frame is classified as voiced or unvoiced by the classification block. Once the envelope and excitation have been extrapolated, speech synthesis is undertaken. Corrective spectral shaping filters are also incorporated into the model at this stage to account for such effects as mouth radiation, supra-WB resonance effects within the WB, and the voiced excitation. Finally the interpolated TB signal is spliced into the LPS output signal, $\hat{s}_{WB}$, to generate the final PWB speech frame.

Performance results are discussed in Chapter 4.

Chapter 4

Experimental Results

4.1 Methodology

4.1.1 Equipment

This section describes the hardware and software used in the simulations. Figure 4.1 illustrates the basic experimental setup for speech I/O and processing. The speech input device was an Andrea Electronics model ANC100 anti-noise microphone. It was selected for its active noise cancelation capabilities and sound card compatibility. This microphone has a frequency response spanning the range from 20 Hz to 10 kHz, and therefore sufficiently accommodates wideband speech. Approximately linear analog-to-digital (A/D) and digital-to-analog (D/A) conversions were performed by a Sound Blaster 16 card (SB16) in 16-bit precision mode. The card is installed in a 486SX-33 personal computer controlled by the Linux operating system version 1.2.13. The software programs used to record and play were, respectively, *srec* and *splay*, which were authored by Hannu Savolainen, and were available at the time of this writing from *ftp://sunsite.unc.edu/pub/Linux/kernel/sound/snd-util-3.0.tar.gz*. To allow the simulations to proceed in a timely manner, the raw sound files were then converted from the little endian binary numeric format used in personal computers to the big endian numeric format used by SPARC machines. All simulations were carried out in MATLAB version 5.004064 running on a UNIX SPARC machine. Within this environment, all computations are done in double precision. A custom toolbox of speech analysis and processing functions for the MATLAB environment was designed by the author to conduct the simulations. For speech output, Labtec CS-550 amplified computer speakers with a frequency response of 40 Hz to 16 000 Hz were used, as well as SONY MDR-W10 dynamic walkman earphones.

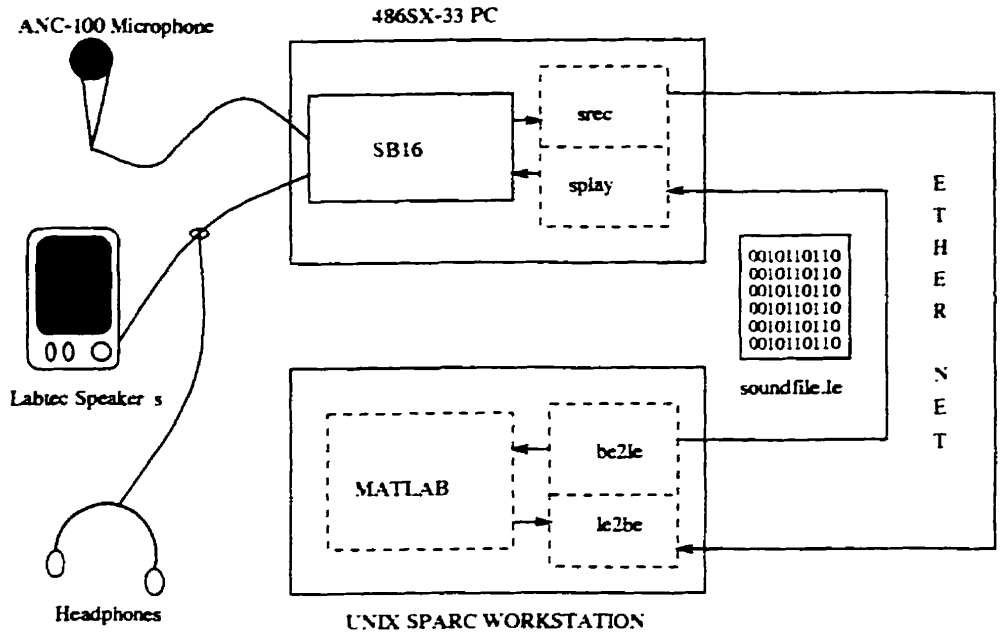


Figure 4.1: Experimental Set-Up for Speech I/O and Processing

4.1.2 Speech Corpus

The sentences for the primary WB speech database or *corpus* were found in [26, p84], and are listed below. The sentences are phonetically balanced, which means that they each contain at least one consonant from each of the three main categories of consonants (*sibilants, stops, fricatives*) and at least one vowel from each of the three main categories of vowels (*front, middle,* and *back*). Examples of these phonemes are provided in Table 4.1.

These sentences were recited by one female speaker in a quiet environment, and recorded using the experimental setup described in the previous section. The SB16 sound card was set for a sampling rate of 16 Hz, and a quantization precision of 16 bits. A corresponding TB corpus was derived from this WB corpus as detailed below. In evaluating the proposed system, the TB corpus was used as input while the WB corpus served as a reference database for the distortion measures.

Despite the speaker dependency issues raised in Chapter 2, it was felt that a single speaker corpus was adequate since inter-frame (phonemic) differences within a single speaker pose similar challenges to the algorithm as inter-speaker differences in that they entail a variation in $F1_{eff}$. In the simulations, the values calculated for $F1_{eff}$ according to (3.7) ranged from 356 Hz to 2150 Hz (corresponding to approximate effective lengths

between 24 cm and 4 cm). For voiced frames, it was found that the average value of $F1_{eff}$ was 566 Hz with a standard deviation of 55 Hz, and the corresponding values for unvoiced frames were 624 Hz and 155 Hz. The higher mean $F1_{eff}$ for unvoiced frames was expected, since the excitation applies to the post constriction segment of the tract, which has a relatively short length.

Phonetically Balanced Speech Corpus

1. Go and sit on the bed.
2. The book is about trash.
3. Moths still turned yellow.
4. Fire consumed the paper.
5. The bowl dropped from his hands.
6. Don't thrash around that way.
7. Those children are dirty.
8. Dress sleeves are much too warm.
9. Tractor plowed the fields.
10. We made some fine brownies.
11. They broke out of prison.
12. He drank our lusty brew.

4.1.3 Objective Measures

The simulations were evaluated on the basis of several objective measures which were defined in Figure 1.1. These objective measures quantify the accuracy with which the synthetic signals mathematically approximate the original WB signal. It was found that a subset of the measures sufficed to quantify the distortions. In the simulations below two measures will be the primary measures quoted: (1) segmental log spectral signal-to-noise ratio (SLS-SNR1); and (2) spectral log root mean squared error (SL-RMSE). These measures are redefined here for convenience.

Consonant Categories		
Sibilants	Stops	Fricatives
/z/ zip	/p/ pat	/v/ vat
/s/ sit	/t/ top	/f/ for
/č/ chat	/b/ bat	/θ/ thin
/š/ shot	/d/ dot	/∂/ that
/j́/ jot	/g/ get	/k/ kit
Vowel Categories		
Front	Middle	Back
/i/ team	/ʌ/ ton	/u/ tool
/I/ tip	/3̂/ bird	/took/
/ɛ/ ten		/o^u/ tone
/æ/ tap		/ɔ/ talk
		/a/ top

Table 4.1: Examples from each of the consonant and vowel categories [26].

$$SLS-SNR1 = \frac{10}{M}\sum_{m=1}^{M} log_{10}\frac{\sum_{k=1}^{N} X_{dB}^2(k)}{\sum_{i=1}^{N}\left(X_{dB}(k)-\hat{X}_{dB}(k)\right)^2}. \tag{4.1}$$

and

$$SL-RMSE = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(X_{dB}(k)-\hat{X}_{dB}(k)\right)^2}. \tag{4.2}$$

where X_{dB} and $\hat{X}_{dB}$ are, respectively the DFT's expressed in decibels, of the original and distorted signals.

The first measure implicitly weights errors less heavily when they coincide with a strong signal. The second measure provides data on the absolute error produced by the distortion. These measures were selected because, in the course of the simulations, they proved to correlate well with subjective quality assessments, and they enable a comparison with results presented in previous work.

Unfortunately, distortion measures were seldom formally defined in the articles, which renders the results open to interpretation. It is believed that SLS-SNR1 and SL-RMSE correspond respectively to the two spectral objective measures mentioned in the article on the HMM-PWB system, there identified as 'segmental spectral SNR' and 'spectral log rms' [8]. It was reported that the HMM-PWB system achieved a spectral SNR gain of 3 dB [8].

Since the algorithm focuses on UB extrapolation, the objective measures are applied only to the UB region. Other bins in both the WB reference and the PWB candidate are spectrally nulled prior to evaluation.

4.2 Telephone Band Speech Model

Figure 4.2 depicts the system used to obtain the TB corpus from the WB corpus. The passband of 300 Hz to 3300 Hz for the bandpass filter corresponds to an intersection of the generally excepted bands for telephone speech [27] [1] [2]. To implement the bandpass filter, a 10^{th} order Chebychev type I filter was employed using zero-phase forward and reverse digital filtering[1]. A Chebychev filter was selected to be consistent with the technique used to generate the TB corpus in the HMM-PWB system [8]. A zero-phase characteristic was desired to yield a perfect time alignment between the WB and TB signals, so that distortion measures such as signal-to-noise ratio (SNR) could be more conveniently applied. Decimation by a factor of two is used to reduce the sampling rate from 16 kHz to 8 kHz.

The final block in the WB-to-TB system truncates the samples to 16 bit (linear) precision. Telephone speech is subject to 8-bit μ-law quantization used in the PSTN, which corresponds to 13 or 14 bits of precision in uniform PCM [16]. It was desired to capture all of this inherent TB precision for PWB processing, and since the SB16 sound card afforded only two precision choices for sound I/O, 8 bit and 16 bit, the latter was chosen.

Thus, telephone speech is modeled as speech bandpassed to the region 300 Hz to 3300 Hz. Since the PWB algorithm is focused on combating the bandwidth truncation distortion, the TB speech model does not incorporate any phase distortions, coding distortions, or transmission errors. In theory, these distortions could be minimized in a preprocessing/equalization phase. To accommodate analog input in operation, it is assumed that the digital PWB system is preceded with an A/D converter to sample the signal at 8 kHz and uniformly quantize it to 16 bits precision. Thus, for all simulations, the input to the PWB system is of the form described for the TB corpus.

[1] The zero-phase filtering was simulated using the *filtfilt* command in the MATLAB environment.

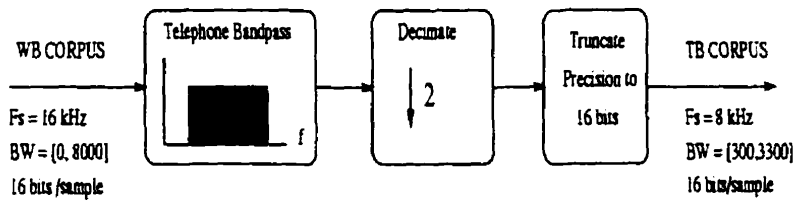


Figure 4.2: Generation of the TB corpus from the WB corpus.

4.3 Performance Baselines

To calibrate and correctly interpret the objective measures used, it is useful to consider performance baselines. Figure 4.3 presents objective measure results to quantify the distortion between TB speech and WB speech and to provide an informal view of the relative information content and importance of the different bands under consideration. In the corpus, the TB speech (with no partitioning V and UV frames) exhibited a SLS-SNR1 of 14.6 dB, and a SL-RMSE of 15.6 dB. These are lower bound performance measures for the system. In terms of both measures, the TB and UB together constitute a closer match to the WB original than the TB and sub-TB together. This verified that it would be worthwhile to concentrate on extrapolation to the UB.

Figure 4.4 quantifies the default distortions present in the UB and sub-TB ranges as a result of the bandpass filtering done to generate the TB. Specifically, it quantifies the distortion in the UB due to the absence of significant UB components in the interpolated TB speech.

4.4 Simulations

The system described in Figure 3.1 has many features which required parameterization. The following sections outline the simulations and results which justify the choice of system architecture and parameter values described in Chapter 3. The simulations were coded for the MATLAB environment employing a speech processing toolbox designed by the author.

Aside from any extrapolation errors, there are various procedural distortions that occur in the PWB system. In particular, errors associated with linear predictive analysis, TB signal interpolation, and sub-band splicing errors, and errors associated with continuity at frame boundaries. These do produce an upper bound on the performance achievable

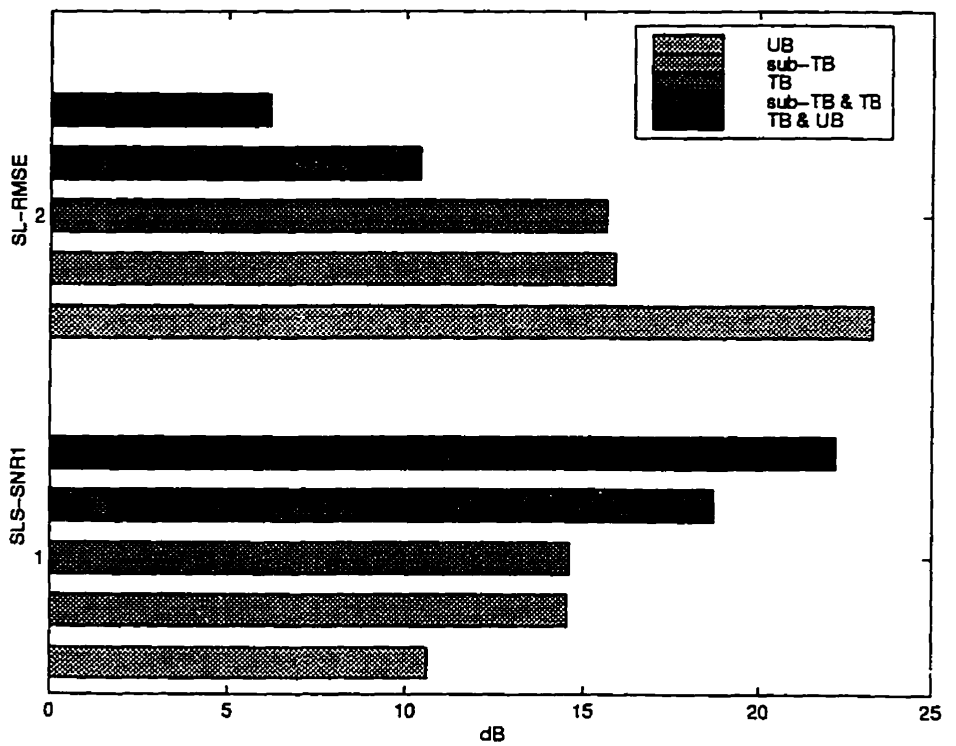


Figure 4.3: Measures of the band-limiting distortion for the TB, UB, sub-TB, TB and UB, and TB and sub-TB bands.

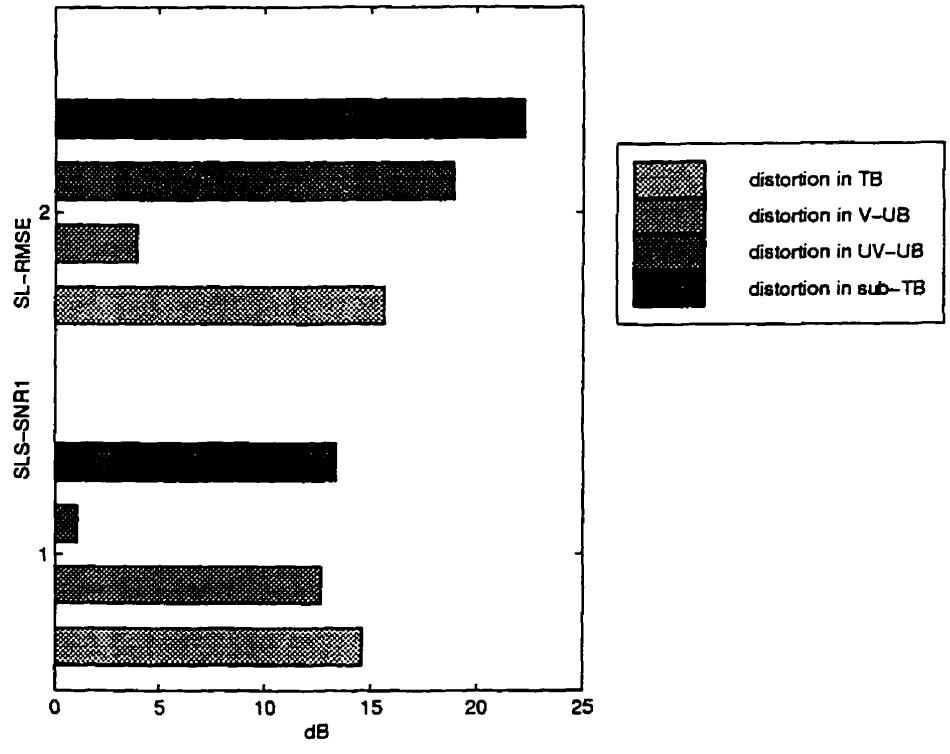


Figure 4.4: Control performance measures for TB speech quality.

by the system, however, after a considerable investment of time, they were found to be effectively negligible (both independently and collectively) in terms of subjective distortion.

4.4.1 Splicing the Bands

For splicing the TB signal into the PWB signal, it was necessary to determine appropriate cutoff frequencies for filters to remove the synthetic TB from the synthetic WB signal. Quadrature mirror filters (QMF) would have been ideal for seamless merging of the subbands, however, the TB speech available as input was not bandlimited with such a filter. The use of quadrature mirror filters would require re-filtering the received TB speech with a QMF, further truncating its bandwidth, which would constitute a self-defeating quality compromise. The splicing itself is justified by the fact that the synthetic PWB signal produced in all the simulations exhibited poorer objective measures within the TB, than the interpolated input TB signal.

Based on SNR, a suitable UB cutoff frequency for the high-pass filter, F_{UB1}, was found to be 3800 Hz. This value was determined using the system in Figure 4.5, and simulation results are presented in Figure 4.6. In the simulations, a 10^{th} order, high-pass, Chebychev type I filter was used to filter the entire signal with *filtfilt* to effect a zero phase characteristic. Similarly, a suitable sub-TB cutoff frequency, $F2_{SUB-TB}$, was found to be 265 Hz, and SNR simulation results are displayed in Figure 4.7. In operation, the ideal filter would operate on a frame-by-frame basis, with appropriate state initializations, and would have a zero phase characteristic to avoid any time misalignment between the bands.

4.4.2 Excitation Simulations

In devising an appropriate excitation extrapolation technique, it was desired to isolate distortions caused solely by excitation extrapolation errors. The simulation system is depicted in Figure 4.8. LPA was conducted on the WB corpus and the TB corpus with orders of $q = 16$ and $p = 6$ respectively. Various excitation extrapolation candidates were conducted on the TB residual, $r_{TB,p}$, to yield r_{PWB}.

For each scheme the synthetic residual, $r_{PWB}(n)$, was compared with the *'ideal' or control residual*, r_{WB} according to the set of objective measures. To construct a signal for subjective evaluation, LPS was conducted using the synthetic WB residual, r_{PWB}, and the *'ideal' envelope*, $H_{WB,q}(z)$(as parameterized by the vector of prediction coefficients $\bar{a}_{WB}$). This PWB speech signal was evaluated both subjectively and objectively in comparison to the to original WB signal. Since, in the final PWB signal, the portion in the TB will be

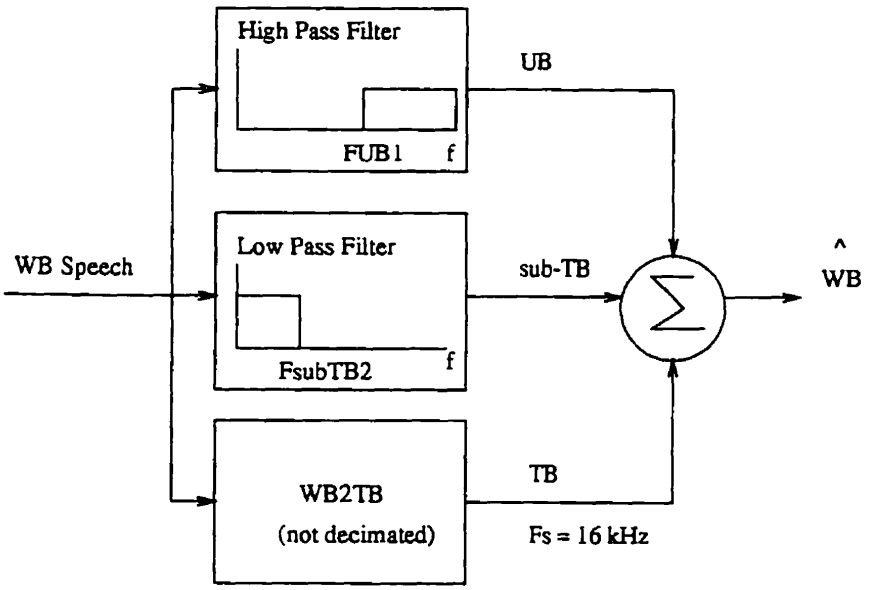


Figure 4.5: Simulation system for determining appropriate cutoff frequency for splicing the TB into PWB speech.

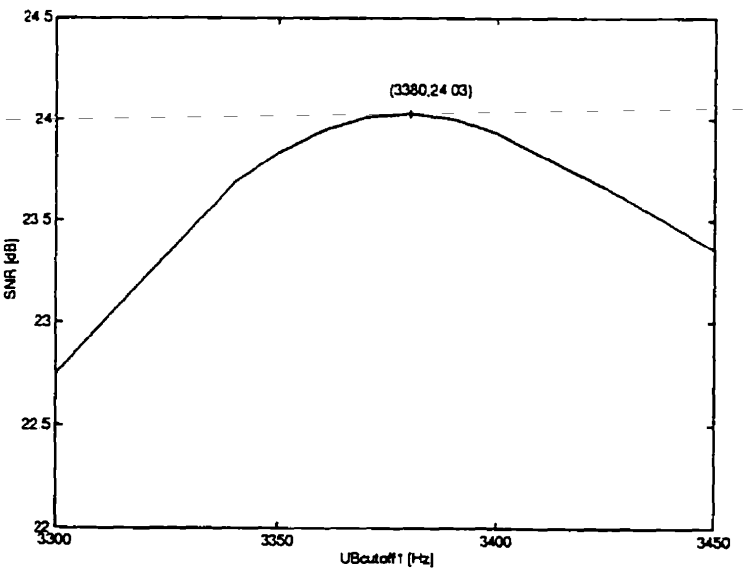


Figure 4.6: Determination of appropriate UB cutoff frequency, F_{UB1} for the highpass filter in the subband splicing phase.

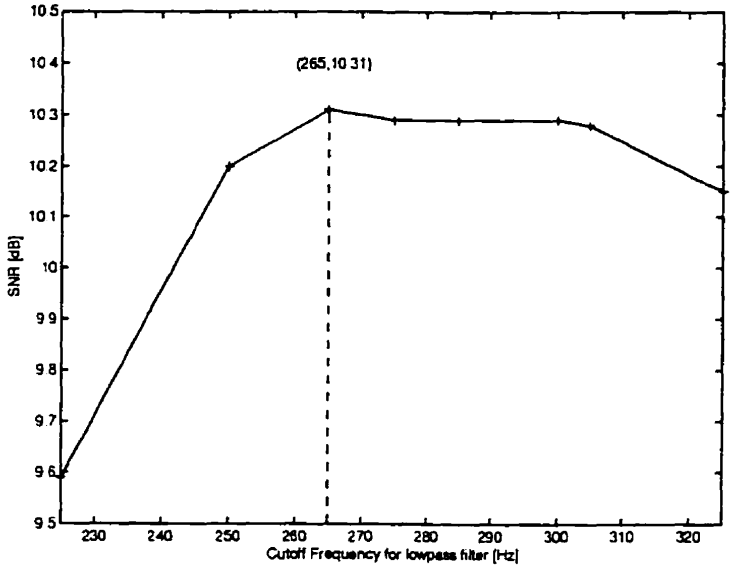


Figure 4.7: Determination of appropriate SUB-TB cutoff frequency, $F_{SUB-TB2}$ for the lowpass filter in the sub-band splicing phase.

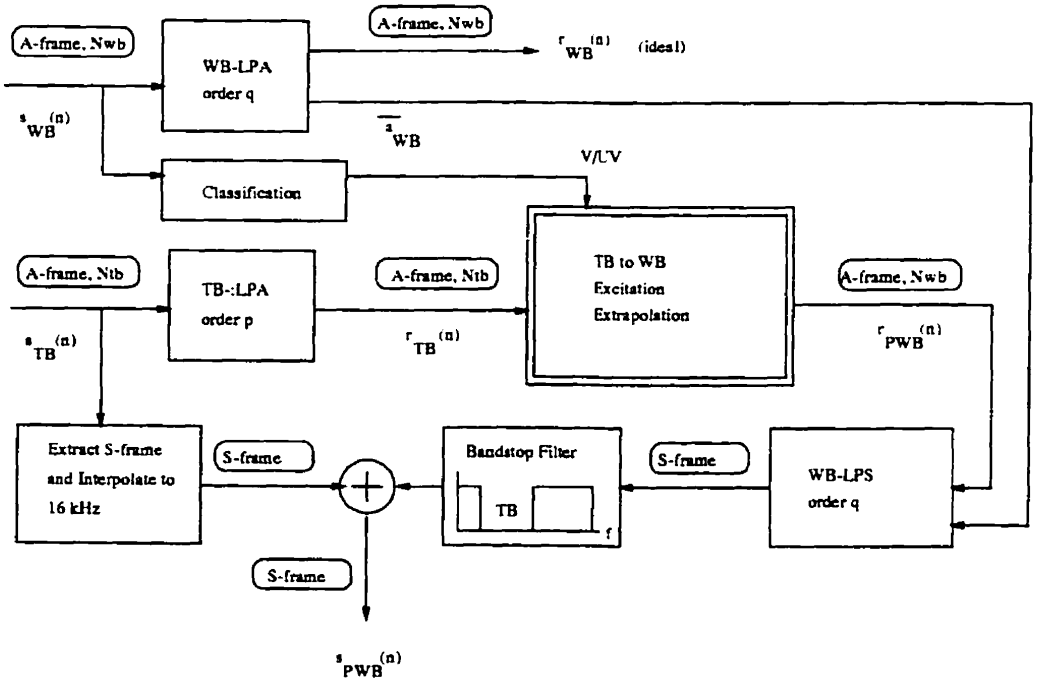


Figure 4.8: Simulation system for evaluating excitation extrapolation techniques.

an interpolated version of the TB signal, distortions in r_{PWB} which fall within the TB are irrelevant. The objective measures are therefore applied in particular to the signal outside the TB range. The actual TB was also spliced into the synthetic signal prior subjective evaluation.

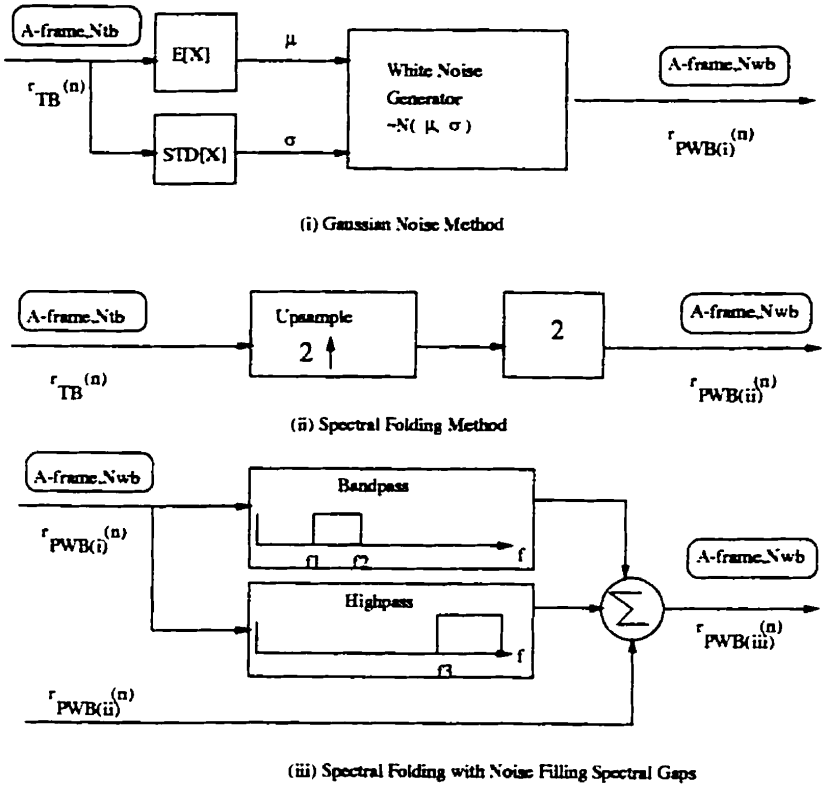
Several techniques were considered as candidates for the UB excitation generation. These included: (i) Gaussian noise: (ii) Spectrally folded TB residual: (iii) Spectrally folded TB residual with Gaussian noise to supplement missing frequency bands. The details for each of these candidates is presented in Figure 4.9, where these systems were inserted in the double outlined block of Figure 4.8. The first two methods have been used historically for excitation extrapolation and were outlined in Section 1.6.2, the third method was adapted by the author. In this simulation, unvoiced speech was defined as speech with the greater proportion of its energy above 3 kHz, and it was found that approximately 12% of the corpus fell into this category.

As predicted in the literature, for unvoiced speech, Gaussian noise excitation produced satisfactory results (in particular, notably crisp /s/'s) with no artifacts. For voiced speech, Gaussian noise produced audible 'chirppy' artifacts. Spectral folding caused no audible artifacts for either voiced or unvoiced speech, however, it did not yield the crisp subjective effect produced by the noise excitation. Spectral folding supplemented with noise in the gap yielded a better signal for voiced frames from a subjective perspective.

According to both subjective and objective measures, Gaussian noise was the best of the three alternatives for generating the UV-UB excitation. Although, noise excitation gave marginally better objective results for voiced speech, this produced audible artifacts, so subjectively it was found that spectral folding with noise was the best choice for V-UB extrapolation. The degradation in quality produced by using extrapolated UB residuals with ideal envelopes was negligible. This suggests that the low frequency ($\leq$ 3300 Hz) region of the residual signal is much more important than the UB residual to the speech quality, and that the UB envelope is more important than the UB residual.

4.4.3 Envelope Extrapolation Simulations

The objective of these simulations was to quantify the performance of the acoustic tract model, in terms of UB extrapolation, and to investigate the use of a correction filter, $C(z)$. As with the excitation simulation system, in evaluating the envelope extrapolation technique, all other components of the signal were controlled by keeping the TB and the residual ideal. LPA was conducted on the WB corpus with an order of $q = 16$ to gener-



(i) Gaussian Noise Method

(ii) Spectral Folding Method

(iii) Spectral Folding with Noise Filling Spectral Gaps

Figure 4.9: Excitation extrapolation candidates.

ate an 'ideal' residual, $r_{WB}(n)$[1]. For both the objective and subjective evaluations, the residual was kept ideal (evaluation of simultaneous excitation and envelope extrapolation distortions is presented in Section 4.6). The simulation system is depicted in Figure 4.10. The objective performance measures quoted were applied strictly to the UB region of the signal.

Odd-harmonic resonance extrapolation was conducted on the TB envelope, $H_{TB,p}$, to yield $T_{PWB,\hat{q}}$ according to the method described in Section 3.5.1. This filter is denoted $T_{PWB}(z)$ to emphasize the fact that it only models *tract* characteristics. The simulations for voiced and unvoiced extrapolation were performed separately. It was anticipated that the model would better apply to voiced sounds (vowels) than unvoiced ones, since the tract is less constricted and more like a uniform tube in the former case. In unvoiced evaluation, the UB for voiced frames was kept ideal and vice versa.

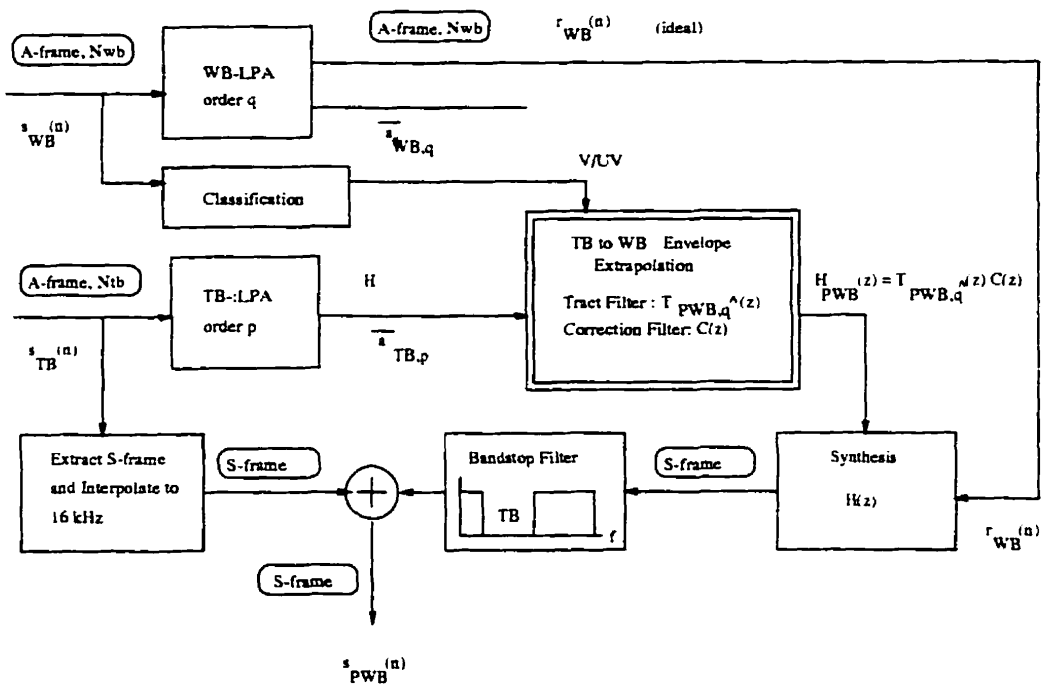


Figure 4.10: Simulation system for evaluating envelope extrapolation techniques.

[1] An order of $q = 16$ was chosen to generate this residual, since the resonance extrapolation technique yielded an average order of 16 for the tube filter, $T_{PWB,\hat{q}}(z)$.

Voiced Speech

For voiced speech, the raw tract UB extrapolation yielded a gain of 4.3 dB in SL-SNR1 over the V-UB performance baseline. The average $F1_{eff}$ found for voiced frames was 565.9 Hz. By inspection of the WB spectra, it was noted that for voiced speech, the voicing roll-off was a dominant factor in the spectral trend. A suitable correction filter, recommended in [23] was

$$C_v(z) = \frac{1}{1 - .96z^{-1}}. \tag{4.3}$$

This correction filter yielded an additional performance gain of 1.3 dB in SLS-SNR1 and a reduction in spectral distortion of 1 dB SL-RMSE with respect to the V-UB performance baseline. Figure 4.11 presents the spectral distortion measures found both with and without the correction factor. Some experimentation was done to simulate the roll-off by other techniques such as excluding poles above 7500 Hz from the extrapolation process, but these attempts were less successful than using $C_v(z)$. Figure 4.12 portrays the actual WB envelope and the PWB extrapolation for a typical voiced frame. As visible in this example, a side effect of the correction filter, $C_V(z)$, is that it boosts the frequency content in the LB, producing a reasonable visual match to the actual LB spectral contour.

Subjectively, some artifacts were perceptible. Avendano confirmed that over or underestimation of the UB envelope is subjectively problematic for voiced frames [2].

Unvoiced Speech

For unvoiced speech, the raw tract UB extrapolation yielded a *gain* of 1.5 dB SLS-SNR1 and a distortion *reduction* of 3 dB in SL-RMSE in the UB range as compared with no extrapolation. For unvoiced speech, the spectral trend is often upwards in the UB. The fore shadowing factor is perhaps a dominant influence[1]. Techniques such as the addition of an extra pole at 8000 Hz were tried but all yielded worse performance than the raw tract extrapolation. The performance of the extrapolation procedure for unvoiced speech is presented in Figure 4.13. Figure 4.14 portrays the actual WB envelope and the PWB extrapolation for a typical unvoiced frame.

Percentage Distortion in Formant Frequencies and Bandwidths

To explain the subjective shortcomings observed for the envelope extrapolation, a simulation was conducted to compare the performance of the uniform tube envelope extrapolation

[1]This spectral contour can be calculated analytically, and is dependent on $F1_{eff}$ [13]. However, the derivation was inaccessible to the author at the time of this writing.

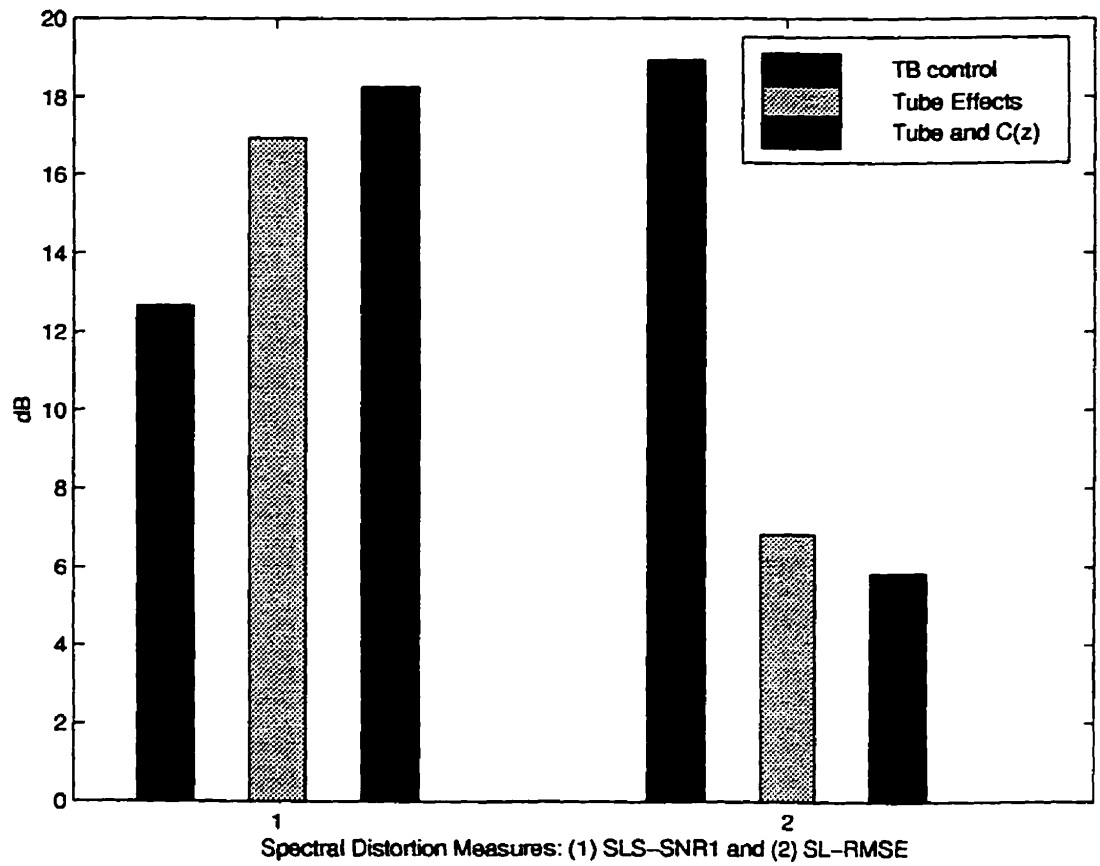


Figure 4.11: Spectral distortion measures for envelope extrapolation of voiced speech.

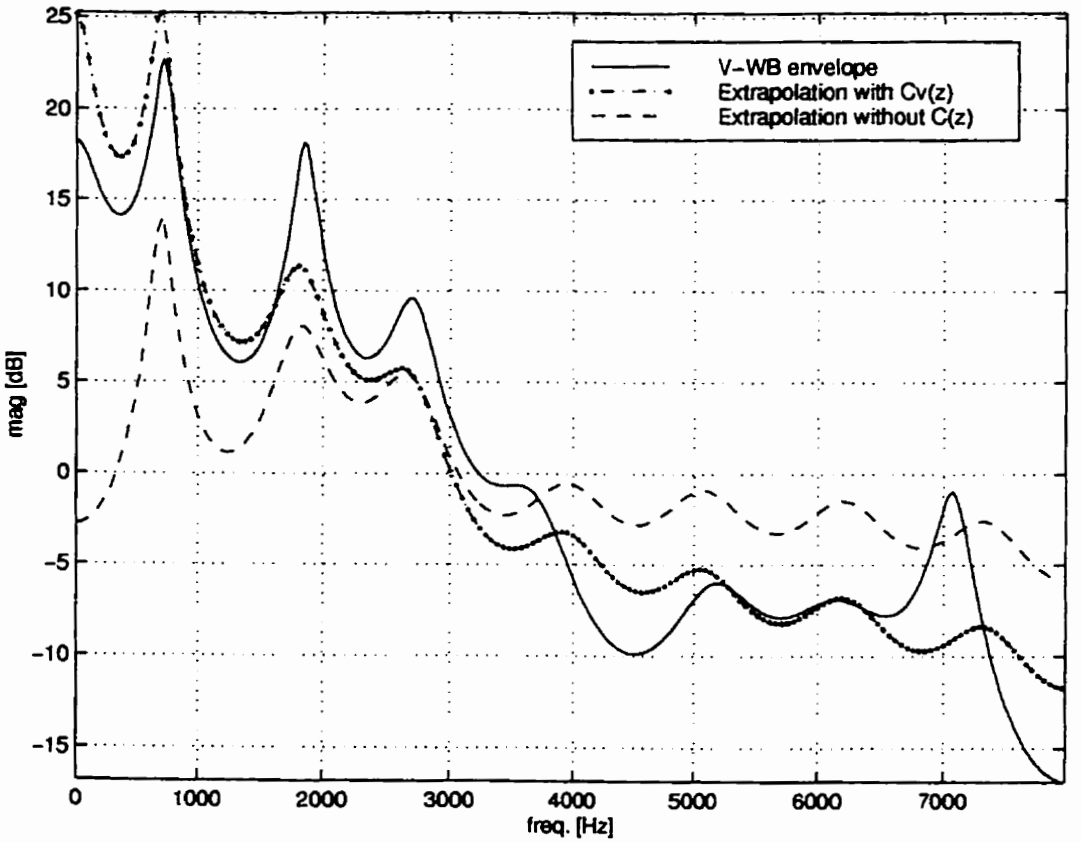


Figure 4.12: Example of envelope extrapolation for a typical voiced frame.

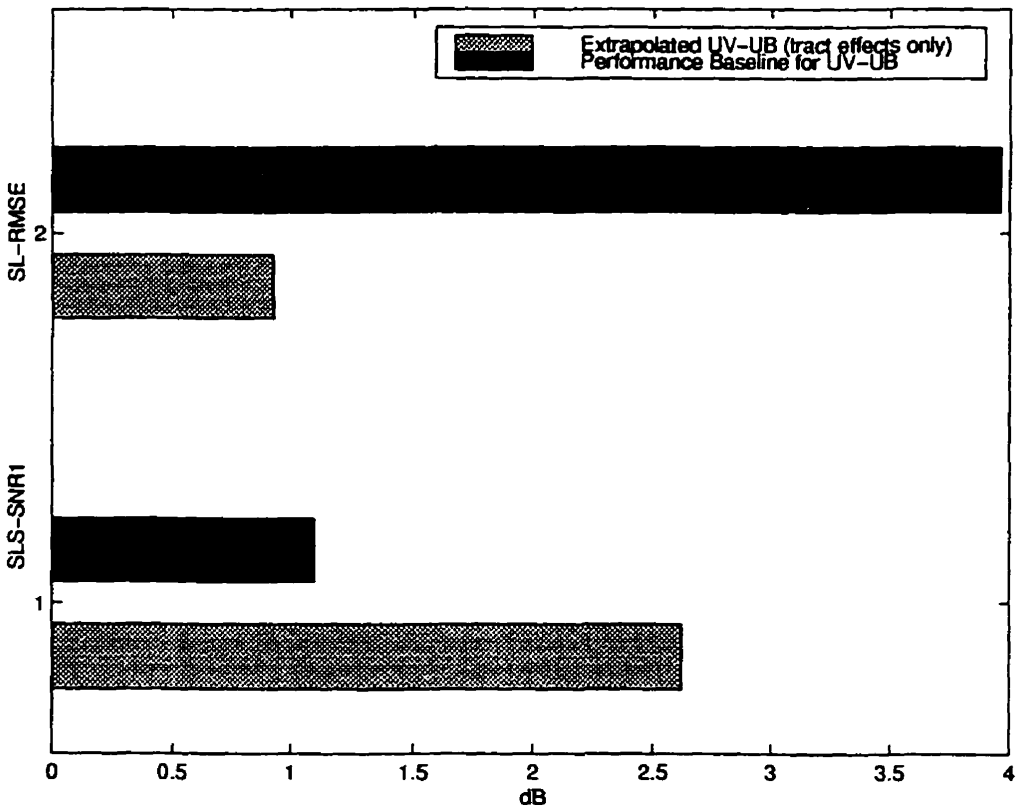


Figure 4.13: Spectral distortion measures for envelope extrapolation of unvoiced speech.

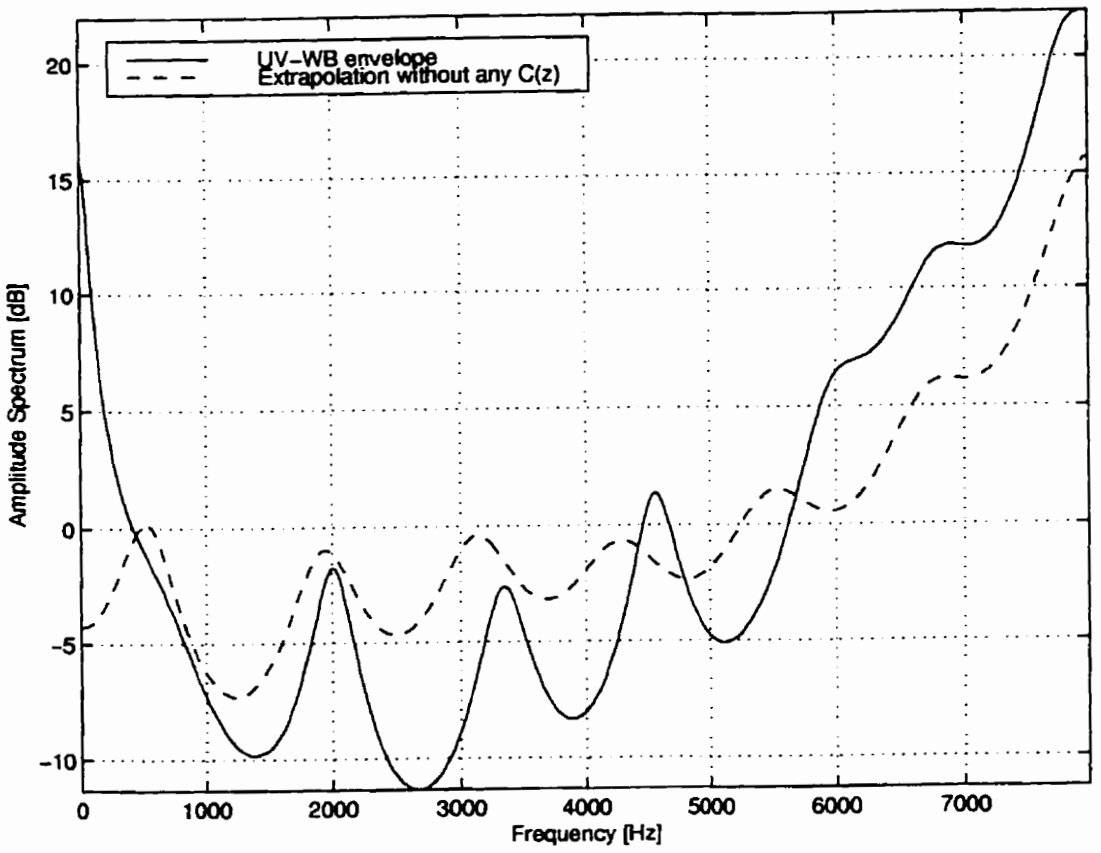


Figure 4.14: Example of envelope extrapolation for a typical unvoiced frame.

technique to the perceptual performance baseline described in Section 2.2.5.

Again, voiced and unvoiced frames were considered separately. In each case, for each extrapolated envelope, $T_{PWB,\hat{q}}(z)$, the order, $\hat{q}$, was noted and used to conduct LPA on the corresponding original WB frame to generate, $H_{WB,\hat{q}}$. Those k formants of the extrapolated envelope above 3300 Hz are then matched up one-to-one with the k highest formants in $H_{WB,\hat{q}}$. Then for each pair a percentage distortion was calculated for both formant frequency and bandwidth according to

$$Ferr = \frac{|F_T - F_H|}{F_H} * 100\%. \tag{4.4}$$

and

$$Berr = \frac{|B_T - B_H|}{B_H} * 100\%. \tag{4.5}$$

where F_T and B_T are the frequency and bandwidth for the extrapolated formant, and F_H and B_H are the corresponding properties for the formant from the WB original.

The results were averaged over the corpus and are summarized in Table 4.2. Several implications are noteworthy. The results for percentage error in formant location validate the intuition that the uniform tube model would perform with better on voiced frames, since it is for these phonemes that the tract most closely approximates the uniform tube configuration. The just noticeable distortion (JND) quoted is for a single formant in the F-pattern [24], while the values given for the voiced and unvoiced scenarios reflect the expected degree of distortion *for each* UB formant. In light of this comparison it is not surprising that the subjective results left something to be desired.

	JND for single formant	VOICED	UNVOICED
Formant Frequency	5 %	7.37 %	11.23 %
Formant Bandwidth	40 %	51.94 %	74.00 %

Table 4.2: Average percentage distortion for extrapolated UB formant frequencies and bandwidths, as compared with perceptual tolerances *for a single errant formant.*

From a subjective perspective, speech which is all ideal except the V-UB sounds just as fraught with artifacts as speech in which all is ideal except the UV-UB. This can perhaps be accounted for by the fact that only about 12% of the frames are unvoiced, so even if those UV frames are fairly distorted, the fact that 88% of the frames are ideal offsets this effect.

4.5 Preliminary Investigations for Alternative Tract Models

Although improvements (over the TB speech) in the objective measures were obtained using the uniform tube model for envelope extrapolation, the subjective assessments were less encouraging. Furthermore, since no universal spectral trend could be found for the UV-UB, no correction filter was suggested by spectral slope analysis of the corpus. In light of the percentage distortion measures presented in Table 4.2 which significantly exceed the perceptual threshold, it was deemed necessary to consider alternative models[1].

4.5.1 Uniform Open Ended Tube for Unvoiced Frames

In applying the tube model to unvoiced speech, it was unclear if the excitation constriction constituted a closed end to the tube. This is an important consideration since the resonant properties of a tube are dictated by application of excitation to its end, in so far as that application renders that end acoustically closed or open. It is generally accepted that during voicing, the vocal folds can be thought to be effectively closed.

The inspiration for modelling the tract as a tube open at both ends for unvoiced, turbulently excited, speech stemmed from the fact that turbulent excitation occurs at the embouchure hole on a flute and a recorder fipple, and these regions are modeled as open ends [4, p. 84]. A tube of length, L, open at both ends, resonates at all the harmonics (multiples) of a fundamental resonance $F1_{open} = \frac{c}{2L}$.

All-harmonic resonance extrapolation of unvoiced speech produced poorer results than the odd-harmonic resonance extrapolation suggested by the uniform tube model with the excitation end closed. In particular, it generated a uniformly spaced F-pattern which was too dense, and consequently resulted in an overestimation of the UB spectrum.

Although the unvoiced speech, like the flute music is produced by an excitation that is turbulent in nature, it appears that their similarities do not extend to the open tube model. The constriction in speech is perpendicular to the tract axis, and may indeed be tight enough to serve as a closed end.

4.5.2 Multiple Independent Resonators

As a loose approximation of the concatenated tube model of the tract, it was assumed that the tube segements were uncoupled, such that each resonated independently of the

[1] Since every effort was made to control all other speech components to be ideal in the simulations, it is justified to find fault with the tract model. Serious investigation was done to ensure that the perceptual artifacts were not caused by some frame or band boundary effects.

others. The net resonance pattern of such a system would be the union of all the sub-tube F-patterns. This is not in strict accordance with the underlying physics on several counts. First, the tube sections are acoustically coupled, and the formants are affected by the interactions between them. Second, in assuming that each tube resonates independently, it was assumed that each acted like a tube *closed at one end* as for the uniform tube tract model. However, this is valid perhaps only for the first tube, which begins at the glottis. An advantage of the model was its ability to yield more random resonance extrapolation patterns than the uniformly spaced UB resonance structure dictated by the uniform tube model. Each TB formant was assumed to be affiliated with it's own segment of tract, such that the TB resonance was the first resonance of its sub tube.

This model yielded poorer results than the uniform tube model, which is appropriate considering it's weaker physical grounding. In particular, the interpretation of each TB formant as a 'seed' for odd harmonic extrapolation yielded an excessive number of predicted UB resonances leading to a gross overestimation of the spectral envelope contour.

4.6 System Performance

Although, as discussed in the previous section, the envelope extrapolation produced by the uniform model was not as subjectively satisfying as desired, this system performance evaluation is included for completeness. The results presented in this section summarize the performance of the proposed PWB system as a whole, as it is described in Chapter 3. The simultaneous effects of excitation extrapolation and envelope extrapolation are considered. Where possible, comparison is made to the performance results reported in the PWB speech systems outlined in Section 1.6.

Figure 4.15 summarizes the performance of the system in terms of the objective measures as applied to the UB. These measures describe how closely the PWB signal approximates the original WB signal. Included on the same plots are the baseline measures for the TB data, to exhibit the improvement or gain achieved via PWB extrapolation.

Overall, the proposed PWB signal exhibits some improvement over the TB in terms of mathematically approximating the original WB magnitude spectrum. For his PWB signal as a whole, Mermelstein reported a 3 dB *gain* in 'spectral segmental SNR' [8], and this can be compared with the 1.41 dB gain in SLS-SNR1 between the TB and PWB results. Abe reported a 6.5 dB 'spectrum distortion' for the UB [1]. Although the term 'spectrum distortion' is not explicitly defined in [1], it is logical, in terms of the units and magnitude, to presume it could be a log-spectral RMS measure, such as LS-RMSE.

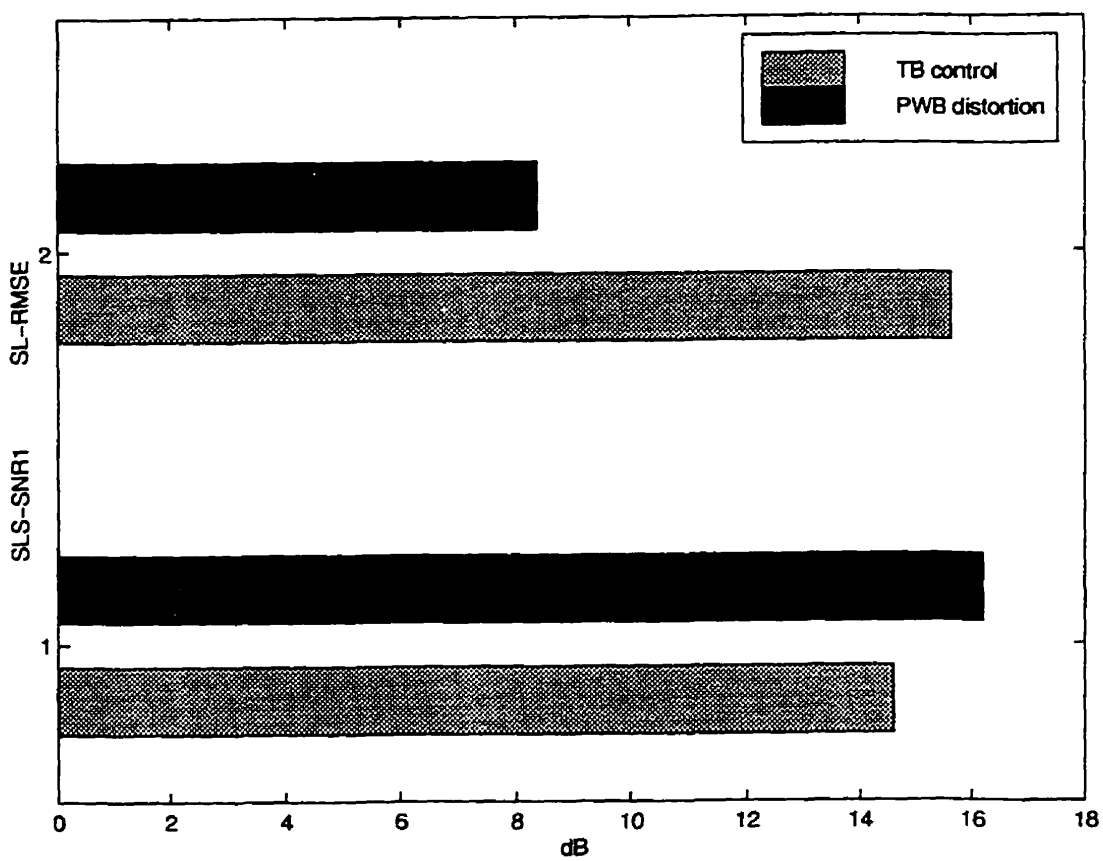


Figure 4.15: Objective results comparing PWB speech with the WB original in the excitation simulations.

If that is the case, the proposed PWB system exhibits a 7.36 dB distortion in the UB.

While the objective results are a positive reflection on the algorithm, the ultimate objective is to produce PWB speech that is of *perceptually* higher quality than TB speech. Unfortunately, the noteworthy 'chirppy' artifacts associated with the envelope alone preclude the need for further subjective testing at this stage.

Chapter 5

Discussion

Based on the promising preliminary results for the PWB speech generation systems described in [1], [8], [2], and [6], the assumption of a sufficiently strong correlation between the TB and the WB for PWB speech extrapolation appears justified. Furthermore, this correlation can only stem from the nature of the underlying physics and acoustics of the speech production process. From this perspective, an acoustic modeling approach to the PWB speech generation problem appears justified and promising. The task becomes one of devising an acoustic model which captures explicitly the nature of the inter-band correlations which are implicitly captured in the SRF ([8]) and spectral envelope codebook techniques.

An adequately parameterized model would, in theory, enable the acoustic parameterization of the tract in a speaker and sound independent fashion. However, it appears that it was overly optimistic to assume that the uniform tube model could yield perceptually superior results.

5.1 Strengths of the Uniform Tube, Fixed Bandwidth Model

For vowels, the uniform tube model is a fairly accepted approximation of the actual tract geometry. Since most frames in the corpus were found to be voiced frames (88%), it was reasonable to hypothesize that a model which catered to the majority could yield satisfactory results.

The uniform tube model yields a spectrum with uniformly spaced formants. Although actual voiced speech frames aren't characterized by so structured an F-pattern, statistically, if properly parameterized, the uniform tube extrapolation should constitute the best approximation of the actual spectrum that a uniformly spaced spectrum can

provide. The simplicity and elegance of the model are also key benefits. Assigning all bandwidths equal to the expected value of UB bandwidths, results in a mean bandwidth error of zero (although the absolute percentage error is much higher as indicated in Table 4.2).

Despite it's advantages, the results yielded by the uniform tube model fell substantially short of the subjective objectives. The two main sources for error in the uniform tube model approach are: errors in parameterizing the model, and inadequacies of the model itself (in terms of its extrapolation capabilities).

5.2 Tract Length Parameterization Errors

The limitations in terms of parameterization stem from the use of LPA to extract formant frequencies. There are some potential drawbacks to this technique. Resonances closely spaced in frequency might yield a single spectral peak, and so be identified as a single formant by LPA. Furthermore, although, on average over an infinitely wide spectrum, the resonant spacings encode the tract length, the three resonances available in the TB might be insufficient to yield an accurate estimate of the global average described in (2.4). Furthermore, the telephone band might contain a number of resonances different than three, but 6^{th} order LPA will forcibly extract three inaccurate formants. A more sophisticated analysis might use an adaptive order for the TB-LPA, by ascertaining the spectral flatness of the segment, and predicting and detecting the number of underlying resonances accordingly.

The scenario in which the first resonance of the tract falls below the TB is discussed in Section 6.2.1, in the context of extending the model to predict sub-TB resonances.

For certain unvoiced sounds, it is not the entire tract which is subject to the excitation, only the post constriction portion, so L_{eff} could be very small. If the constriction is within 2.5 cm of the mouth opening, even the fundamental resonant frequency, $F1_{eff} \approx 3400Hz$, falls above the telephone band. This would render the telephone band effectively barren of tract resonance information, and consequently would preclude the possibility of extrapolation. Conveniently, for $L_{eff} < 1$ cm, even the first resonance falls above the wideband frequency range so it would not be necessary to take it into account in PWB generation.

5.3 Limitations of the Uniform Tube, Fixed Bandwidth Model

Although it was expected to serve as an adequate approximation for voiced speech, the average of 7.37% error for every formant location is substantially above the threshold of perception. A portion of this error is attributable to errors in the estimate of effective length, as discussed above. The performance deteriorates still further when applied to unvoiced speech.

Although the bandwidth error becomes zero in the mean, the percentage error is also well over the perception threshold.

The uniform tube model produces a spectrum in which formants are uniformly spaced. For actual speech spectra, this is rarely the case. Tract constrictions will perturb them, and the locations and existence of these non-uniformities cannot be accounted for with the uniform tube model.

5.4 Potential for Other Acoustic Approaches

In light of the results discussed in Section 4.4.3, the first order approximation of a uniform tube seems to have some applicability to most frames in the corpus. This model does not take into account independent resonant effects of different tract sections. It would appear that such simple acoustic principles are insufficient to achieve the desired quality gain over TB speech. A more rigorous classification of speech and a more sophisticated acoustic model are needed.

The failure of the alternative attempts of multiple independent resonators and the open ended tube for unvoiced frames does not preclude the possibility of a successful acoustic approach to the PWB speech problem. In particular an acoustic/phonetic/articulatory method is proposed in the Future Work section.

A concatenated tube model, such as is described in [12] might yield better results. Concatenated tube models produce, in general, F-patterns which are not limited to the uniform spacing structure. The difficulty is in finding a model which can be parameterized from the TB.

It would appear that there is an inherent tradeoff/performance limitation associated with the acoustic modeling approach. The more fully the model describes the tract configuration, the more accuracy can be expected from the associated analytical prediction of the spectral composition of speech. However, the more complex the model, the more difficult it is to accurately parameterize it based solely on the properties of a TB

speech frame. This is due in part to discrepancies in tract length, which result in spectral shifts and dilations of scale for the same phoneme spoken by different speakers. This issue is discussed more thoroughly in Section 5.5. Furthermore, speech production modeling is still an active field, so there are aspects of the process which are as yet poorly understood, or not well documented. In general, the study of the acoustics of speech has focused on the narrowband range of speech, due to it's significance for intelligibility.

5.5 Speaker Dependence

The speaker dependent nature of speech might become more of an issue, in terms of parameterization concerns, for more complicated models, such as the phonemic one.

In principle, an acoustic aware algorithm is robust against speaker differences, since it attempts to extrapolate resonances based on simple harmonic principles from the resonances present in each TB frame. The difficulty arises in the limitations of the TB spectral window. For the same phoneme and basic tract configuration, two speakers with tracts of different lengths would exhibit different *TB visible* F-patterns, as exemplified in the open uniform tube example of Figure 5.1. For the shorter tube, only two formants are captured within the TB.

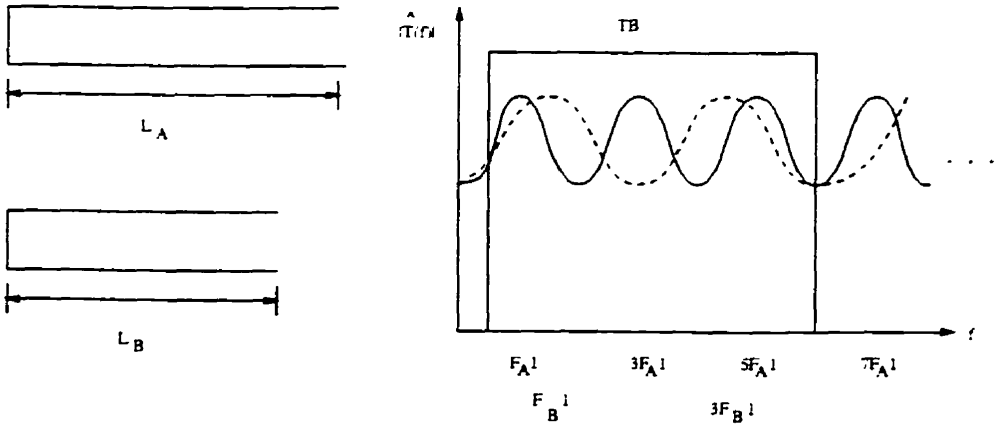


Figure 5.1: Effect of tract length on the F-pattern observable in the TB spectral window.

For the PWB algorithm presented in this thesis, the potentially speaker sensitive parts are the classification block and the LPA formant identifier block. The envelope shape and number of resonances expected in the TB depends on the tract length. For women, approximately three formants are expected, while for men, three or four could be present.

The order of the LPA or some other feature of the TB formant detection block should ideally be adapted to accommodate the more or less information revealed for a given speaker in the TB spectral window. Identification of extraneous resonances or failure to identify resonances would lead to misparameterization of the model and erroneous extrapolation into the missing bands. This problem pertains not only for different speakers, but also for changes in effective tract length within a given speaker.

To determine how much effective information is present a spectral flatness measure could be employed. This is a means to gauge the number of effective formants in the TB. If the same order of LPA is employed on frames from different speakers, interpretation errors are bound to occur.

To normalize the fundamental tract resonance frequency, $F1_B$, of the shorter tract with respect to the longer tract in Figure 5.1.

$$\hat{F1}_B = F1_B \frac{L_A}{L_B} = F1_A. \tag{5.1}$$

where $\hat{F1}_B$ is the 20 cm normalized fundamental frequency for tube B. This is based on the relation for tubes closed at one end that $F1 = \frac{c}{4L}$, where c is the speed of sound.

Cumulative data across multiple frames of the signal might be particularly useful for parameterizing macroscopic tract geometries for better on-line adaptation to a particular speaker.

Chapter 6

Conclusions

A PWB speech generation algorithm has been presented which extrapolates missing spectral components from TB speech. Unlike previous attempts which are based predominantly on the statistical clustering of a training corpus, this method employs an acoustic model of speech production, and thereby delves more explicitly into the physical basis and nature of the correlation between the different frequency ranges of speech.

Although the performance of the algorithm is not up to a commercial level, the results were encouraging considering the simplicity and elegance of the uniform tube model.

The challenges in acoustic modeling and extrapolation are manifold. Speech production modeling is an ongoing field. Furthermore, even where adequate models for various phonemic categories exist, identification of said categories from the TB is difficult, and the process is complicated by speaker normalization issues. Furthermore, such analysis could be prohibitively complex from an implementation standpoint.

6.1 Contributions

6.1.1 Explicit Acoustic Approach to PWB Speech Generation

The use of an explicit acoustic tract model constituted a novel approach to the PWB speech generation problem, and a novel application of acoustic speech production concepts. Most previous work on speech acoustics focussed on the prediction of narrowband speech spectra from an empirically measured tract configuration.

The specific uniform tube model selected had several attractive features. The uniform tube is a simple yet reasonable analog for the human tract, especially during vowel production which embodies the majority of the frames. The resonance behaviour of such a

tube reveals the physical basis for the correlation between TB and WB speech. A physical explanation and formulation for this correlation was notably absent from previous work in the area. The potential for speaker independence is inherent in such a model. Unlike the case with statistical pattern matching methods, performance is not dependent on a specific training corpus or testing corpus, since the acoustic resonance principles are fundamental and universally applicable. When accurately parametrized according to effective tract length, the uniform tube model produces the statistically optimal approximation of the broad spectrum achievable with a uniformly spaced F-pattern. This is achieved with an absolute minimum cost in terms of complexity. The fact that the model is fully described with a single parameter renders it reasonable to parametrize solely from the TB, and precludes compound parametrization errors. In terms of appropriateness, simplicity, complexity, and extrapolation capability, the uniform tube model scores very well.

However, the model, like all approximations, also exhibited some limitations. By nature it produces a uniformly spaced F-pattern, which isn't strictly evident, even for vowels, in natural speech frames. For some unvoiced sounds, the uniform tube is a less accurate anolog of the tract than for vowels, and it's applicability to such sounds is questionable. A consideration in interpreting the subjective results, however, is that the performance of the algorithm is naturally dependent upon how accurately effective tract length is exacted from the TB speech. It is difficult to partition the responsibility of the approximation errors between the model itself, which is limited to produce a uniformly spaced F-pattern, or the parametrization. The parametrization issue is again addressed in Section 6.2.1.

This study can be interpreted as a preliminary investigation into the application of acoustic modeling to the PWB speech generation problem. Although the UB extrapolation produced by the algorithm was not subjectively sufficient, the objective results reveal the underlying merit to the acoustic approach, even as exemplified by the simplistic uniform tube model.

6.1.2 Voiced Excitation Extrapolation from TB to WB

A drawback of the traditional spectral folding technique, as applied to TB residuals, is the fact that it leaves spectral gaps in the ranges from 3300 Hz to 4700 Hz and from 7700 Hz to 8000 Hz. This spectral folding technique was extended by the author to incorporate suitably scaled noise in the spectral gaps, thereby producing a spectrally complete and continuous WB residual[1]. It was found that no significant artifacts are produced when

[1] In [32], noise was used in conjunction with spectral folding, but it was a superimposed arrangement, in which the noise essentially whitened the already spectrally complete folded residual.

such a synthetic WB residual is used in conjunction the true envelope in LPS.

6.1.3 Speech Processing Toolbox

This research necessitated the design and creation of a custom toolbox of speech analysis and processing functions for the MATLAB simulation environment. This toolbox will be posted to the 'shareware' MATLAB web archive, where it will have the potential for productive reuse, since such specialized functionality is not currently supported by the commercial toolboxes.

6.2 Future Work

There are two possible avenues for future work. The first involves a revamping of the acoustic model to allow it to better accommodate the myriad of idiosyncrasies in tract behaviour. This would involve issues such as frame categorization based on tract acoustics, and LB extrapolation. The second route would involve a departure from acoustic modelling, due to the inherent limitations discussed in Section 5. The author is of the opinion that a neural network approach would be promising.

6.2.1 Extension of Acoustic Model

Sub-TB Extrapolation

The algorithm described in Chapter 3 was geared for UB extrapolation. It was assumed in analysis that the first tract resonance fell within the TB range. In frames in which the first resonance occurred below the TB, this assumption resulted in inaccurate extrapolation on two counts. The estimate of $F1_{eff}$ in (3.7) is perverted by the misidentification of $F2_{TB}$ as $F1_{TB}$, and consequently, the UB extrapolation is erroneous. Also, the missing sub-TB resonance remains missing in the PWB speech, since no LB resonance extrapolation is conducted.

For completeness, it would be desirable to accommodate LB extrapolation into the current system. In TB analysis, it would become necessary to detect the situation in which the first formant falls below the TB. In the case where the first three tract resonances fall within the TB, (3.7), applies and

$$F1 \approx F1_{eff} \approx \frac{1}{2}\left(\frac{F2 - F1}{2} + \frac{F3 - F2}{2}\right). \tag{6.1}$$

If the the first TB formant greatly exceeds that predicted by this equation, it is possible there is a tract resonance which occurs below the TB.

In terms of excitation, a more sophisticated scheme involving pitch detection might be needed, since Avendano asserts that for low frequency reconstruction the fine spectral structure of the lower harmonics is more important than envelope shape [2].

6.2.2 Acoustic-Phonetic/Articulatory-Phonetic Model

The voiced/unvoiced partition revealed that the uniform tube model yielded better results for voiced than unvoiced speech. A more accurate and sophisticated model would likely involve a multi-class partitioning of the speech to handle a variety of tract configurations and couplings. A substantial amount of data in the literature relates the tract configuration to an associated phoneme (articulatory-phonetics), and acoustic signal characteristics to an associated phoneme. This information could be exploited to construct a two stage approach to the problem. This method would essentially involve speech recognition, and would pay a penalty in terms of complexity.

Perhaps the biggest challenge for the application of this acoustic model to the generation of PWB speech is the need to classify the frame, *based solely on the TB*, into an appropriate acoustic category for which appropriate tract parameter values can then be determined. A method hinted at, although not employed by Abe in [1] was the observation that the spectral magnitude trend (in dB) is somewhat consistent throughout the wideband range. That is, if the TB spectrum slopes downwards, as for voiced speech, it is expected that the UB will also.

6.2.3 Non-Acoustic Approaches to PWB Speech Generation

Although more intuitively satisfying than statistical techniques, an acoustic approach poses no special advantage from the perspective of a commercial algorithm. A well parameterized acoustic tract model would doubtless be *sufficient* to permit accurate spectral extrapolation, but the preliminary results of the non-acoustic techniques indicate that it is by no means *necessary*. In the case of TB and WB speech, the shortest path between the two points is possibly a straight line. Delving beneath the signal layer into the underlying physical layer might be an unwarranted sojourn.

Speaker dependence, and frame-to-frame variation within a given speaker yield a huge domain and range for the mapping from TB speech to WB speech. It appears as though a complex and involved acoustic model would be required to guarantee perceptually

accurate extrapolation in the missing bands. The extraction of parameters from a single frame of telephone band speech to sufficiently parameterize such an involved model would be challenging indeed.

Many of these issues are handled implicitly in methods which employ a large multi-speaker training corpus, such as the codebook spectral envelope methods. An alternative method, which to the author's knowledge has not been investigated, is the use of neural networks to determine the correlation between the TB and WB speech. At issue in such an approach is how to efficiently represent the TB and WB speech frame data for presentation to the neural network. The obvious candidates are vectors of linear prediction coefficients (or transforms thereof), both for compactness and applicability to speech synthesis. If possible, a preprocessing tract length normalization could be conducted on the training data prior to training the network. The performance of this method would depend on the architecture of the network, the training corpus, and the choice of feature vector description used. This method shares the advantage of the spectral codebook method that many speaker discrepancies and classification issues are handled implicitly.

The spectral shifts and dilations associated with speaker normalization suggest that wavelet analysis also might be a feasible alternative approach to the problem.

References

[1] M. Abe and Y. Yoshida. More natural sounding voice quality over the telephone! An algorithm that expands the bandwidth of telephone speech. *NTT Review*, 7(3):104–109, May 1995.

[2] C. Avendano, H. Hermansky, and E. Wan. Beyond Nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech. In *EUROSPEECH*. 1995.

[3] Richard E. Berg and David G. Stork. *The Physics of Sound.* In Stork [4], 2nd edition, 1995.

[4] Richard E. Berg and David G. Stork. *The Physics of Sound.* Prentice Hall, Englewood Cliffs, NJ, 2nd edition, 1995.

[5] G. J. Borden and K. S. Harris. *Speech Science Primer: Physiology, Acoustics, and Perception of Speech.* Williams & Wilkins, London, 2nd edition, 1984.

[6] H. Carl and U. Heute. Bandwidth enhancement of narrowband speech signals. In *Signal Processing VII, Theories and Applications. Proceedings of EUSIPCO-94. Seventh European Signal Processing Conference.*, volume 2, pages 1178–81, Sept. 1994.

[7] H. H. Chen, R. Cox, Y. Lin, and N. S. Jayant. A low-delay CELP coder for the CCITT 16 kb/s speech coding standard. *IEEE J. Sel. Areas. Commun.*, 10(5):830–849, June 1992.

[8] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein. Statistical recovery of wideband speech from narrowband speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):544–548, Oct. 1994.

[9] M. G. Croll. Sound quality improvement of broadcast telephone calls, BBC Research Report RD1972/26. Technical report, British Broadcasting Corporation, 1972.

[10] R. Drogo de Iacovo, R. Montagna, and D. Sereno. A low-delay wideband speech codec at 16 kbit/s. In *Signal Processing VI - Theories and Applications, Proceedings of EUSIPCO-92, Sixth European Signal Processing Conference*, volume 1, pages 483–485, 1992.

[11] M. Dietrich. Performance and implementation of a robust ADPCM algorithm for wideband speech coding with 64 kbit/s. In *Proc. Int. Zurich Seminar Digital Communications*, 1984.

[12] H. K. Dunn. The calculation of vowel resonances, and an electrical vocal tract. In R. W. Schafer and J. D. Markel, editors, *Speech Analysis*. IEEE Press, 1979.

[13] G. Fant. *Acoustic Theory of Speech Production with Calculations based on X-Ray Studies of Russian Articulations*. Mouton, Paris, 2nd edition, 1970.

[14] G. Fant. The acoustics of speech. In R. W. Schafer and J. D. Markel, editors, *Speech Analysis*. IEEE Press, 1979.

[15] J. L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, New York, 2nd edition, 1972.

[16] CCITT Recommendation G.711. Pulse code modulation (PCM) of voice frequencies. Technical report, IXth Plenary Assembly, Blue Book, Melbourne, Australia, Nov. 1988.

[17] A. Gersho. *Digital Speech processing: Speech coding, synthesis and recognition*, chapter Speech Coding, pages 73–99. Kluwer Academic Publishers, Boston, 1992.

[18] M. Guglielmo, G. Modena, and R. Montagna. Speech and image coding for digital communications. *European Trans. on Telecomm. and Related Technologies*, 2(1):21–44, 1991.

[19] M. J. Hunt. *Digital speech processing: Speech coding, synthesis and recognition*, chapter The Speech Signal, pages 43–72. Kluwer Academic Publishers, Boston, 1992.

[20] N. S. Jayant. High quality coding of telephone speech and wideband audio. *IEEE Communication Mag.*, 28(1):10–20, 1990.

[21] J. Makhoul. Linear prediction : A tutorial review. *Proceedings of the IEEE*, 63:561–580, April 1975.

[22] J. Makhoul and M. Berouti. High-frequency regeneration in speech coding systems. In *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 428–431, 1979.

[23] D. O'Shaughnessy. *Speech Communication: Human and Machine.* Addison-Wesley Pub. Co., Reading, Mass., 1987.

[24] F. J. Owens. *Signal Processing of Speech.* Macmillan P., Basingstoke, 1993.

[25] P. J. Patrick and C. S. Xydeas. Speech quality enhancement by high frequency band generation. In *Proc. Int. Conf. Digital Proc. of Signals in Communications*, pages 365–373, 1981.

[26] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements. *Objective Measures of Speech Quality.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[27] C. Rowden, editor. *Speech Processing.* McGraw-Hill, New York, 1992.

[28] M. Serizawa and K. Ozawa. 4 kbps improved pitch prediction CELP speech coding with 20 msec frame. *IEICE Trans. Inf. and Syst.*, E78-D(6):758–763, 1995.

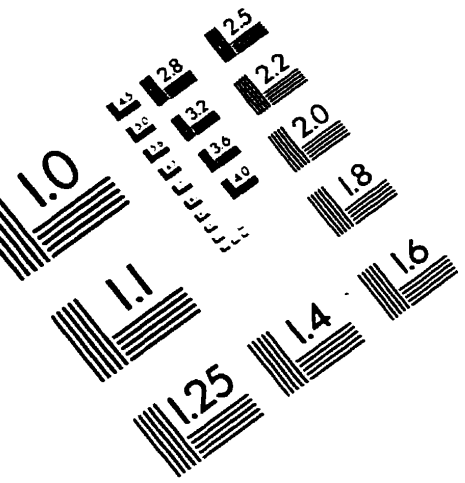
[29] K. N. Stevens and A. S. House. An acoustical theory of vowel production and some of its implications. In D. B. Fry, editor, *Acoustic Phonetics: A Course of Basic Readings.* Cambridge University Press, 1976.

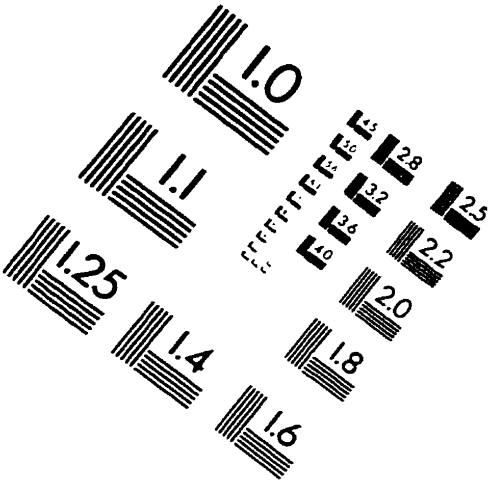
[30] G. W. Stewart. *Introductory Acoustics.* D. Van Nostrand Company, Inc., New York, 1933.

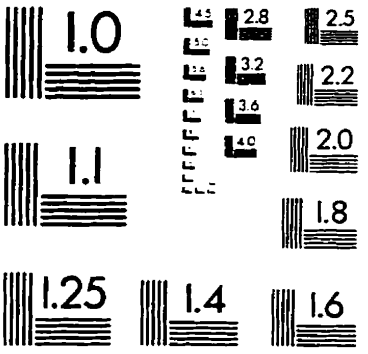
[31] P. Strevens. *Spectra of Fricitive Noise in Human Speech.* Cambridge University Press, 1976.

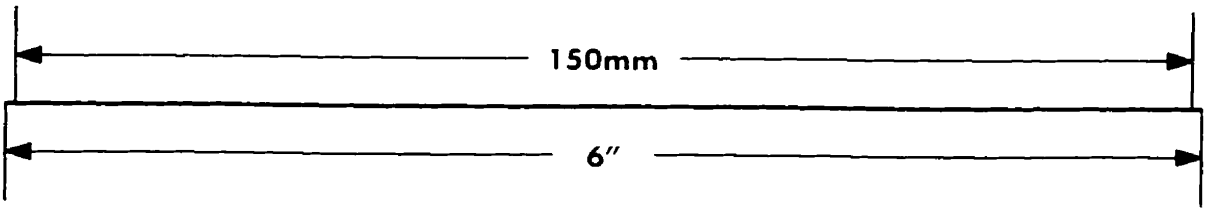
[32] Chong Kwan Un and D. Thomas Magill. The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s. *IEEE Trans. on Communications*, COM-23(12):1466–1474, Dec. 1975.

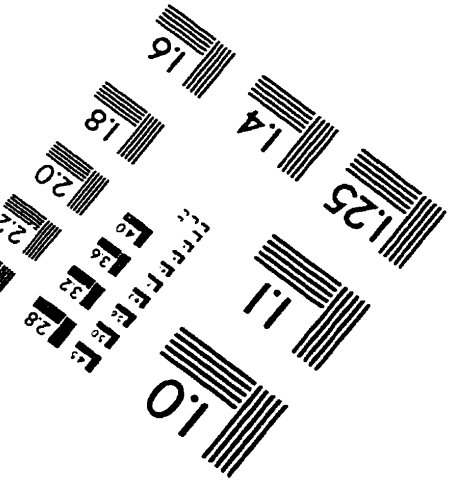
IMAGE EVALUATION TEST TARGET (QA-3)

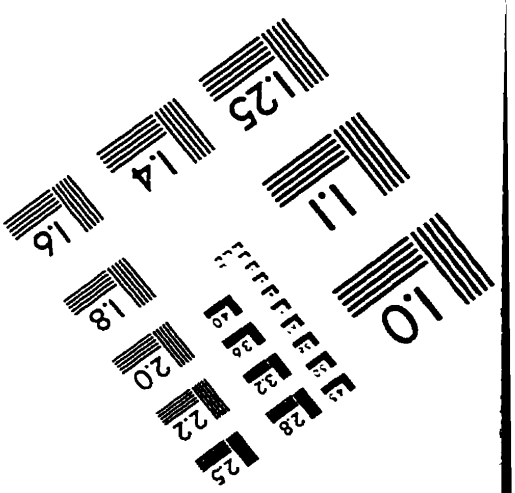












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989