National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

Study of the Correlation of LSP and LPC Frequencies for Vowel Phonemes

by

Rubina Hussain

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering

Department of Electrical Engineering

McGill University

Montreal, Canada

April, 1994

ISBN 0-315-99966-7

Acknowledgments

I would first like to thank my advisor, Dr. Douglas O'Shaughnessy, for his invaluable advice and insights on the content of this thesis.

Secondly, I would like to express my gratitude to Bell Canada and Stentor Resource Center Inc. for providing me with the opportunity and the financial support to pursue graduate studies at McGill University and at INRS-Telecom. I would especially like to acknowledge Luc Morin and Michel Plouffe, who were my supervisors at Bell Canada when I was awarded the scholarship.

I would like to thank my parents, sisters and brother for their kind support and encouragement throughout my Masters program. I would like to acknowledge my friends and colleagues at INRS-Telecom and McGill, and especially Rachida and Lila for their support.

I would like to extend a special thanks to my dear friend and mentor Susan-Joan, without whom my accomplishments would not have been possible.

Finally, to my loving husband Aleem, who has always been there with relentless courage and an enormous clarity of vision. His profound belief in my abilities and his continuous support and advice have been instrumental in the realization of my achievements.

Abstract

The objective of this thesis is to provide an analysis of Linear Predictive Coding (LPC) frequencies and Line Spectrum Pair (LSP) frequencies. While it is generally known that LSP frequencies congregate about LPC formants, this study will deal with *how* LSP and LPC frequencies correlate. The work done in this study is intended to enhance the recognition of vowel phonemes by providing observations on the similarities between LPC and LSP frequency distribution. New information regarding exact LSP and formant relationships for different vowel phonemes is provided. Specifically, this study divides vowel and diphthong phonemes into categories, or types. These types pertain to specific patterns in the way LSP frequencies relate to the closest corresponding formant, F1, F2, or F3. In drawing relationships between formants and their corresponding LSPs, another indicator is made possible such that, when incorporated within other statistical or knowledge based techniques, it may serve to enhance the effectiveness of recognition systems.

The results of this thesis indicate that it is possible to divide the sixteen vowel phonemes into nine types. The experiment produced a 75% success rate when the test set was applied.

The first part of this thesis provides the theoretical background for the basic understanding of line spectrum pairs and the acoustic-phonetics related to the formation of vowel sounds. The second part of this thesis provides details on how the experiment was conducted and the results of the analysis. Explanations are also provided for these results.

Résumé

Le but de cette thèse est de fournir une analyse comparative des fréquences LPC (Linear Predictive Coding) et des fréquences LSP (Line Spectrum Pairs). Il est généralement reconnu que les fréquences LSP s'amalgament au voisinage des formants LPC, aussi cette étude traitera de la manière dont les fréquences LSP et LPC sont corrélées. Le travail réalisé dans cette étude devrait servir à améliorer la reconnaissance des voyelles en fournissant des observations sur les similarités entre les distributions des fréquences LPC et des fréquences LSP. En traçant des liens entre les formants et les LSP, on obtient un autre indicateur exploitable, lorsqu'il est incorporé soit à d'autres techniques statistiques, soit à des techniques basées sur la connaissance. Il peut alors servir à augmenter l'efficacité du système de reconnaisance. Plus précisément, cette étude répartit les voyelles et les diphtongues en catégories ou types. Ces types correspondent à un schéma spécifique à la manière selon laquelle les fréquences LSP sont reliées au plus proche formant, F1, F2 ou F3. Les résultats de cette thèse indiquent qu'il est possible de répartir les 16 voyelles de l'anglais selon 9 types différents. L'expérience donne un taux de réussite de 75% quand on utilise l'ensemble test.

La première partie de cette thèse fournit une base théorique pour la compréhension des LSP ainsi que l'essentiel de l'aspect acoustico-phonétique lié à la formation des voyelles. La deuxième partie de cette thèse fournit des détails sur la manière dont les expériences ont été menées ainsi que sur les résultats de l'analyse.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Current methods developed in much of the area of speech analysis and synthesis are predominantly based upon the principle of the conservation of the speech spectrum, and not directly on the conservation of the actual speech waveform. These methods do not attempt to re-create the original waveform itself, but to re-create the original spectrum of the speech waveform. In fact, the conservation of the speech spectrum, along with a reasonably modeled phase, ensures adequate intelligibility and speech quality for most applications. A natural speech waveform contains many redundancies, which is due in part to the fact that good quality is feasible with waveform samples occurring at 10,000 samples/second, whereas the articulator movement can be well coded at 50 samples/second. These redundancies have led to the development of many efficient coding methods used for speech storage and transmission [1][23][39].

In analysis and synthesis methods, certain parameters are extracted during the analysis. These parameters are then utilized at the synthesis end to re-construct the speech. Many methods of feature parameter representation exist, and the quality of the synthesized speech, as well as the coding efficiency, is directly related to the properties of these parameters and their extraction [4]. Most research to date has concentrated on

methods that result in the most efficient transmission of speech [1][19][23][24][39]. The formulation of the Line Spectrum Pair (LSP) method, which is the main subject of this thesis, resulted from the evolution of different methods of feature extraction of speech waveforms. Each of these methods built on each other by compensating for drawbacks in the previous method.

The most popular, and most successful, of these methods is the linear prediction (LP) of speech. Linear predictive coding (LPC) has been the most important technique of parameter extraction to date [33][28]. Ideally, it is required that the models upon which LPC analysis is based are both time-invariant and linear. The structure of the acoustic waveform is, at minimum, complex, and does not satisfy either of these requirements [4].

It is possible, however, to make reasonable assumptions in formulating a linear model in which the continuously time-varying speech waveform is considered to be locally time-invariant over short time periods within significant acoustic events [12]. The LPC model further involves the separation of the smoothed envelope structure from the actual spectrum of speech. Furthermore, a physiological significance is attached to each element of the LPC model. Because the synthesis of speech is based solely on its feature parameters, a high degree of accuracy is required to quantize these parameters to ensure filter stability. One of the more important drawbacks, however, of the LPC direct-form filter is its relatively poor quantization characteristics. Also, LPC parameter values are heavily dependent on the order of analysis [33].

To compensate for the drawbacks associated with the LPC direct-form filter, the partial correlation (PARCOR) lattice filter was invented to perform the same transfer function as the LPC analysis [23]. One of the basic ideas behind PARCOR analysis is to assume that a speech signal can be approximately represented as an output signal from an all pole filter. In the PARCOR speech analysis and synthesis method, feature parameters are composed of excitation source and spectral parameters. The excitation

source parameters represent vocal cord vibration and are composed of fundamental frequency, power, and voicing information. The spectral parameters represent vocal tract frequency transmission characteristics according to articulator movements [29].

The PARCOR filter, however, has the disadvantage of a relatively large spectrum distortion when interpolating parameters over a long interval. That is, the spectral distortion due to parameter interpolation increases rapidly as the parameter refreshing period is lengthened. This is due to the fact that PARCOR parameters are compound parameters and have relatively poor linear interpolation characteristics. The line spectrum pair (LSP) scheme, proposed by Itakura and Sugamura, was devised to correct this weakness in the PARCOR scheme [36].

The LSP method exploits the all-pole modeling of speech, where LSP parameters are interpreted as one of the representations of LPC parameters in the frequency domain. The equivalent line spectrum pair parameter representation of LPCs allows a 25% to 30% lower bit rate than do the LP coefficients while yielding the same quality of speech [25], or a 40% lower bit rate if compared to PARCOR analysis. These enhancements are due to the exploitation of the properties of LSP in the areas of coding and quantization for improved transmission efficiency, continuous speech recognition and speaker recognition [26][29].

1.1 Overview of Thesis

An essential component of any speech system, and where the greatest data reduction (for waveform transmission) occurs, is the feature extraction phase. Since the majority of speech consists of vowels, the feature extraction for vowel sounds is investigated in this research. One of the most important speech characteristics for

vowels is information about the location and distribution of formants[1] in frequency domain. While there has been a great deal of work on the mathematical or statistical aspects of the LSP model and its applications [3][7][13][16][18], there has been very little produced that studies the behavior of LSP frequencies from a cognitive, or knowledge-based, perspective [10][20][25]. The findings of this research represent additional information for the recognition of formants. The output of this research will provide new information that may be used for greater accuracy of speech recognition systems.

This thesis begins by describing the evolution from LPC to PARCOR to LSP analysis. Chapter 2 discusses the mathematical relationships, and provides detail on the special properties of the LSP model.

The results presented here use the LSP model in conjunction with the principles of articulatory phonetics to draw conclusions regarding the relationships between LPC and LSP frequencies. Chapter 3 discusses these principles in greater detail.

This study examines the relationship between frequencies from a speaker-independent perspective, where sixteen vowel phonemes were extracted from 112 audio files. Chapters 4 and 5 describe how data was acquired and preprocessed, the process of training set correlation, and the identification of categories into which the phonemes are grouped. The results show that it is possible to divide the sixteen vowel phonemes into nine distinct categories based on different configurations of LSP and LPC frequencies. The performance on the test set is also described, as well as possible reasons for any inconsistencies. Suggestions for further study are also presented. Finally, Chapter 6 provides a conclusion to this study.

---

[1] Formants correspond to the resonance frequencies of human vocal tract when it is modeled as an acoustic tube.

Chapter 2

Theoretical Background and Properties of Line Spectrum Pairs

This chapter will provide a theoretical background for the study of line spectrum pairs. Specifically, LSP analysis will be shown to be a natural extension of the LP analysis/synthesis model [29]. Some practical methods for the calculation of LSPs will be discussed. Next, some physical properties of the line spectrum frequencies will be presented through a discussion of various applications of the LSP analysis in speech analysis. Finally, an appreciation of the existing information available about the relationships between LSP and LPC frequencies will be provided.

2.1 From LPC to LSP

In the linear predictive analysis, or maximum likelihood estimation, of the speech spectrum, a segment of speech is assumed to be generated as the output of an all-pole digital filter described by $H_p(z)$ as [23]:

$$H_p(z) = 1/A_p(z) \tag{2.1}$$

where:

$$A_p(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p} \tag{2.2}$$

is the inverse filter, p is the order of the LP analysis, $\{a_i\}$ are the LP coefficients, and:

$$z = e^{il} \tag{2.3}$$

$$l = 2\pi fT, \tag{2.4}$$

where T= sampling period and f is the formant frequency. Whereas this speech generation circuit has an extremely simple configuration, a large spectrum distortion occurs when the $\{a_i\}$ are quantized with fewer than 10 bits for each a_i. That is, that the circuit often becomes oscillatory, or otherwise unstable [25][26][29].

In order to mitigate the drawbacks associated with the quantization of the LP coefficients, the PARCOR system was developed, where:

$$A_{n-1}(z) = A_n(z) + k_n B_{n-1}(z), \tag{2.5}$$

and

$$B_n(z) = z^{-1}\left[B_{n-1}(z) - k_n A_{n-1}(z)\right] \tag{2.6}$$

where k_n are the reflection coefficients, and n refers to the order of the coefficient, $n = 1,\ldots,p$. Also, initial conditions are:

$$A_0(z) = 1 \tag{2.7}$$

$$B_0(z) = z^{-1} \tag{2.8}$$

$$B_n(z) = z^{-(n+1)} A_n(z^{-1}) \tag{2.9}$$

where:

$$z = e^{jl} \tag{2.10}$$

$$l = 2\pi f T, \tag{2.11}$$

and, as before, T is the sampling period and f is the formant frequency.

Figure 1 below depicts the PARCOR synthesis circuit. The transfer function from input terminal X to output terminal Y is $H_p(1/z)$ (where p is used to denote the PARCOR transfer function); the transfer function from Y to Z is $B_p(z)$. The system function from X to Z is then:

$$R_p(z) = B_p(z) / A_p(z) \tag{2.12}$$

where the system is stable for $|k_i| < 1$, for all i. The physical importance of the system function between X and Z is described in the next section.

Even though the PARCOR method alleviates the quantization problem, the PARCOR parameters have another significant drawback: they are compound parameters and have relatively poor linear interpolation characteristics. Thus a significant disadvantage for the PARCOR system is its relatively large spectrum distortion when interpolating parameters in a long frame interval. The line spectrum pair (LSP) method has been proposed to deal with this weakness [29]. The following section discusses the line spectrum pair representation of an all-pole spectrum.

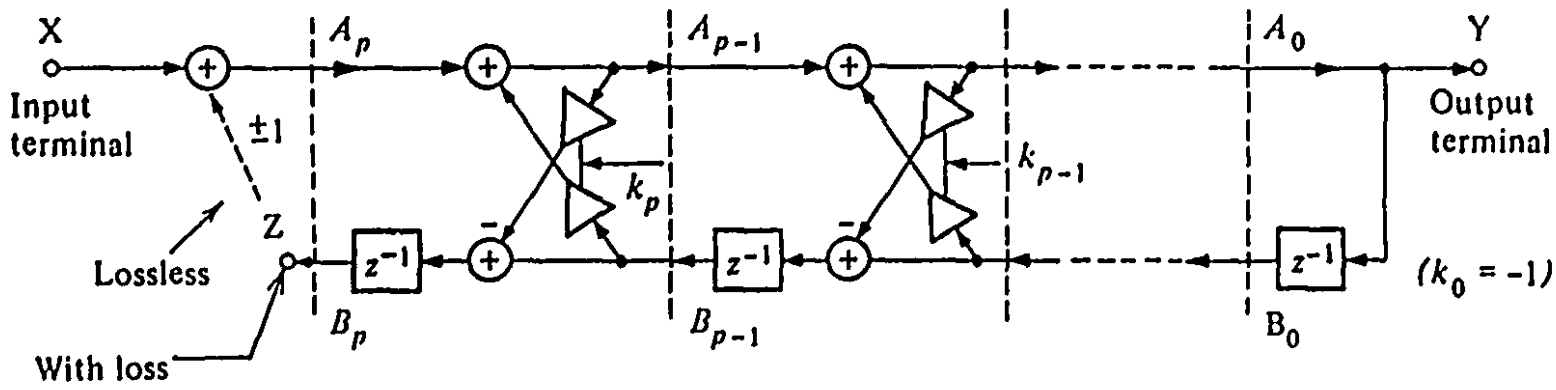

Figure 1: PARCOR Synthesis Circuit [29] [41].

2.2 Line-Spectrum Representation of an All-Pole Spectrum

The PARCOR system, as shown in Figure 1, assumes sound wave propagation through a lossless acoustic tube [23]. The movement of sound in this tube may be described by two components: a forward wave that goes from the glottis to the lips, as well as a backward wave that goes from the lips to the glottis. An interesting result occurs when the reflection coefficients in equations 2.5 and 2.6 are set to 1 and -1. The physical meaning of this is as follows: the acoustic tube becomes perfectly lossless when the output of the Z terminal is fed back to the input terminal through the path with $k_{p+1} = \pm 1$, the value Q of each resonance becoming infinite. The spectrum is then characterized by distributed energy concentrated in several line spectra. The feedback condition corresponding to $k_{p+1} = +1$ corresponds to an opening to infinite free space, while $k_{p+1} = -1$ corresponds to a perfect closure at the input terminal [20].

In the case when an acoustic tube becomes lossless and has an infinite Q of resonance, the root of the p^{th}-order equation which produces the resonance of the spectrum becomes e^{jl}, the root lying on the unit circle of the Z-plane. The order of the equation is then reduced to $p/2$ by projection of the unit circle onto the x axis by $x = \cos(l)$ to eliminate redundancies in calculation. The following shows the residual calculation of $H_p(z)$ at l_i which gives its intensity m_i. The resulting (l_i, m_i) have been called the line spectrum representation of $H(z) * H(z)$ by Sugamura and Itakura. [2]

Referring to Figure 1, the transfer function from X to Y, $H_p(z)$, is given by the equations (2.1) and (2.2). The transfer function from Y to Z is:

$$B_p(z) = z^{-(p+1)} A_p(z^{-1}) \tag{2.13}$$

and the transfer function from X to Z is:

$$R_p(z) = B_p(z) / A_p(z) = z^{-(p+1)} A_p(z^{-1}) / A_p(z). \tag{2.14}$$

The filter will be lossless by the feedback path from Z to X, with $k_{p+1} = \pm 1$. The resonance of Q will be infinite, and the spectrum is reduced to line spectra.

When $k_{p+1} = +1$, $H_{p+1}(z)$ is denoted by the symmetric polynomial, $P_p(z)$, where:

$$P_p(z) = A_p(z) - B_p(z). \tag{2.15}$$

When $k_{p+1} = -1$, $H_{p+1}(z)$ is denoted by $Q_p(z)$, an anti-symmetric polynomial, where:

$$Q_p(z) = A_p(z) + B_p(z), \tag{2.16}$$

and gives rise to the following theorem:

Theorem 1: *If the polynomials of $A_p(z)$ and $B_p(z)$ both satisfy the recursive relation of equations (2.5) and (2.6) respectively, and the roots of $A_p(z) = 0$ all exist inside the unit circle of radius $|k| = 1$, then the roots of $P_p(z) = 0$ and $Q_p(z) = 0$ all exist on the unit circle.*

In order to find the roots of $P_p(z)$ and $Q_p(z)$, many methods may be employed as is discussed in a later section of this report. For the purposes of illustration, we employ the inverse cosine method. We first solve for $P_p(z)$; however, the procedure is also applied to find the roots of $Q_p(z)$. Substituting equation 2.13 into equation 2.15, for p even gives:

$$
\begin{aligned}
P_p(z) &= A_p(z) - z^{-(p+1)}A_p(1/z) = A_p(z) - z^{-(p+1)}A_p(z^{-1}) \\
&= (1 + a_1 z^{-1} + a_2 z^{-1} + \ldots + a_{p-1} z^{-(p-1)} + a_p z^{-p}) \\
&\quad - z^{-(p+1)}(1 + a_1 z + a_2 z^2 + \ldots + a_{p-1} z^{(p-1)} + a_p z^p) \\
&= (1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{p-1} z^{-(p-1)} + a_p z^{-p}) \\
&\quad - (a_p z^{-1} + a_{p-1} z^{-2} + \ldots + a_1 z^{-p} + z^{-(p+1)}) \\
&= 1 + (a_1 - a_p) z^{-1} + \ldots + (a_{p-1} - a_2) z^{-(p-1)} + (a_p - a_1) z^{-p} - z^{-(p+1)}.
\end{aligned} \tag{2.17}
$$

This $(p+1)$-order equation consists of symmetric coefficients around the order of $p/2$, and $(p/2)+1$, with a root at $z=1$. Therefore, the above equation becomes:

$$
(1-z)(1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p}). \tag{2.18}
$$

The circuit is lossless and the spectrum becomes a line spectrum where $z = e^{jw}$. Solving for the real roots only, equation 2.17 is reduced to a $p/2$-order equation of x through repetition of the following procedure:

$$
\begin{aligned}
x = (z + z^{-1}) &= 2\cos(w) \\
(z + z^{-1})(z + z^{-1}) &= z^2 + 2 + z^{-2} \\
z^2 + 1 + z^{-2} &= (z + z^{-1})^2 - 1 = x^2 - 1 \\
(z^2 + 1 + z^{-2})(z + z^{-1}) &= z^3 + 2z + 2z^{-1} + z^{-3} \\
z^3 + z + z^{-1} + z^{-3} &= (x^2 - 1)x - (z + z^{-1}) = x^3 - 2x \\
(z^3 + z + z^{-1} + z^{-3})(z + z^{-1}) &= (x^4 + 2z^2 + 2 + 2z^{-2} + z^{-4}) \\
z^4 + z^2 + 1 + z^{-2} + z^{-4} &= (x^3 - 2x)x - (z^2 + 1 + z^{-2}) \\
&= x^4 - 2x^2 - (x^2 - 1) \\
&= x^4 - 3x^2 + 1
\end{aligned} \tag{2.19}
$$

Its solution, for $i = 1 \ldots p/2$ gives:

$$
x_i = 2\cos(\omega_i). \tag{2.20}
$$

Hence, the line spectrum frequency ω_i is given by the inverse cosine:

$$\omega_i = \cos^{-1}(x_i / 2). \tag{2.21}$$

2.3 LSP Synthesis Circuit

The feedback path of Figure 1 is described by the system function for $k_{p+1} = \pm 1$. In order to find a practical scheme for LSP synthesis we reformulate $P_p(z)$ and $Q_p(z)$, described in equations (2.15) and (2.16), as follows [1] [11]:

For $k_{p+1} = 1$,

$$P_p(z) = A_p(z) - B_p(z) = A_p(z) - z^{(p+1)} A_p(1/z), \tag{2.22}$$

and for $k_{p+1} = -1$,

$$Q_p(z) = A_p(z) + B_p(z) = A_p(z) + z^{(p+1)} A_p(1/z). \tag{2.23}$$

From Theorem 1, the roots of $P_p(z)$ and $Q_p(z)$ exist on the unit circle, and $P_p(z)$ and $Q_p(z)$ may be factored in the following manner [1] [25]:

For p even:

$$P_p(z) = (1 - z)\prod_{\substack{i=2 \\ even}}^{p}(1 - 2\cos\omega_i z + z^2),$$

$$Q_p(z) = (1 + z)\prod_{\substack{i=1 \\ odd}}^{p}(1 - 2\cos\omega_i z + z^2). \qquad (2.24)$$

For p odd:

$$P_p(z) = (1 - z^2)\prod_{\substack{i=2 \\ even}}^{p}(1 - 2\cos\omega_i z + z^2),$$

$$Q_p(z) = \prod_{\substack{i=1 \\ odd}}^{p}(1 - 2\cos\omega_i z + z^2). \qquad (2.25)$$

Hence,

$$A_p(z) = \left[P_p(z) + Q_p(z)\right]/2. \qquad (2.26)$$

The theoretical scheme of line spectrum pair synthesis described in 2.26 would contain a direct feedback path without a delay, and is practically unrealizable. In order to achieve a realizable method, equations (2.24) and (2.25) may be reformatted after expansion to yield the following representation of $A_p(z)$ where $c_o = -z$, $c_1 = -z$, and $c_i = -2\cos(\omega_i)$.

For p even:

$$A_p(z)-1=z/2\left[\sum_{\substack{i=2\\ even}}^{p}(c_i+z)\prod_{\substack{j=0\\ even}}^{i-2}(1+c_jz+z^2)-\prod_{\substack{j=2\\ even}}^{p}(1+c_jz+z^2)\right]$$
$$+\left[\sum_{\substack{i=1\\ odd}}^{p-1}(c_i+z)\prod_{\substack{j=1\\ odd}}^{i-2}(1+c_jz+z^2)-\prod_{\substack{j=1\\ odd}}^{p-1}(1+c_iz+z^2)\right]. \qquad (2.27a)$$

For p odd:

$$A_p(z)-1=z/2\left[\sum_{\substack{i=2\\ even}}^{p-1}(c_i+z)\prod_{\substack{j=0\\ even}}^{i-2}(1+c_jz+z^2)-z\prod_{\substack{j=2\\ even}}^{p}(1+c_jz+z^2)\right]$$
$$+\left[\sum_{\substack{i=1\\ odd}}^{p}(c_i+z)\prod_{\substack{j=-1\\ odd}}^{i-2}(1+c_jz+z^2)\right]. \qquad (2.27b)$$

These equations may now be used to construct a realizable scheme. A practical circuit to implement the line spectrum pair synthesis is shown in Figure 2 for $p=8$ and $p=9$. The circuit may be realized with either (ω_i,ω_{i+1}) or (c_i,c_{i+1}), by using $c_i=\cos(\omega_i)$, $i=1,3,5,\ldots,(p+1)/2$. The (ω_i,ω_{i+1}) terms represent one frequency pair. The (ω_i) and (ω_{i+1}) are called the LSP representation of the speech spectrum. [29]

As a note, however, the circuit does possess some drawbacks. Gretter and Omolog have noted that this structure is more sensitive to the effect of large spectral changes involving adjacent speech frames than the direct filter. Instabilities occur at the beginning of the second frame [10]. A solution proposed by these researchers involves updating LSP parameters at each pitch pulse by utilizing linear interpolation between frames, thus avoiding instabilities.

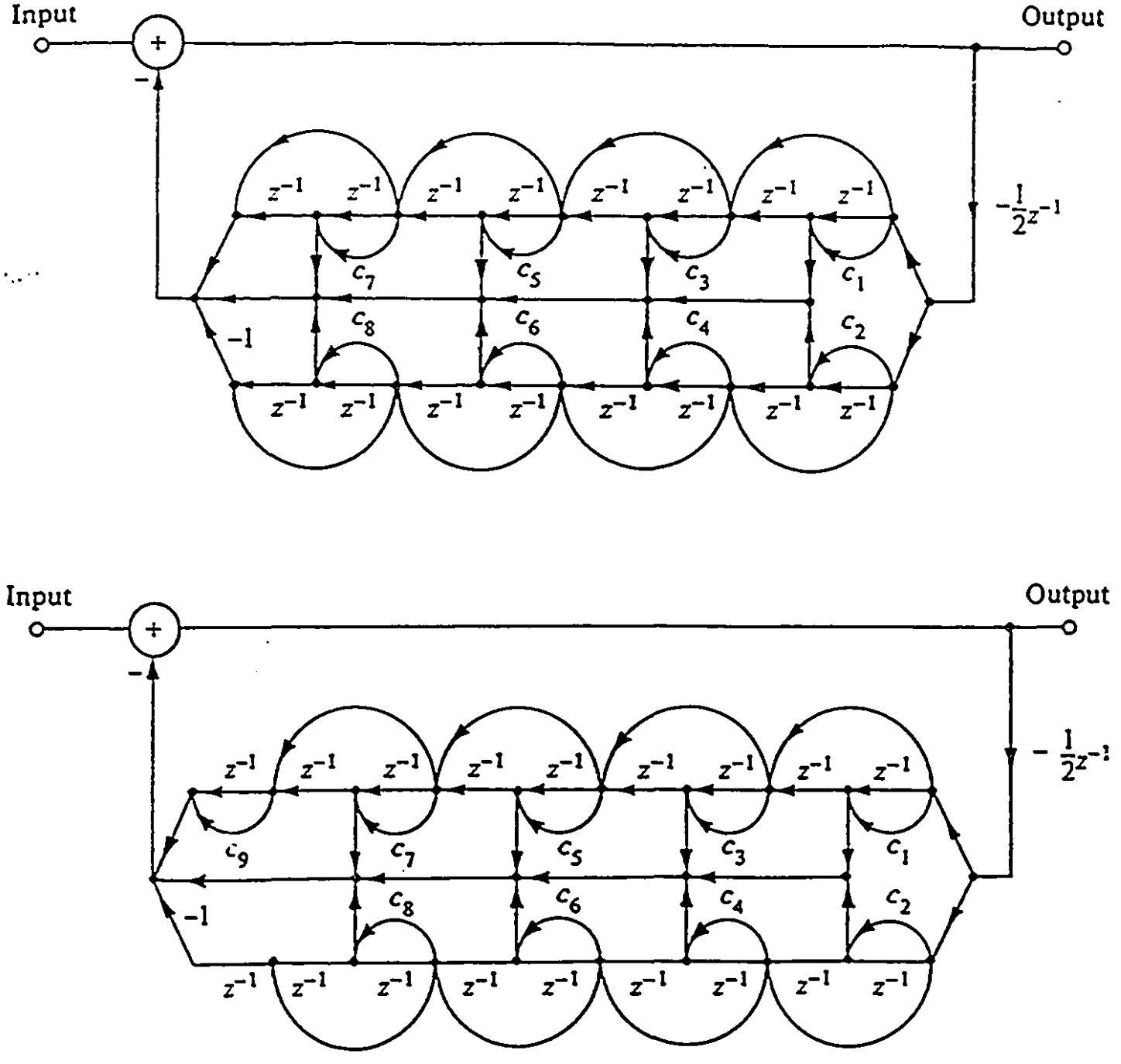

Figure 2: Practical LSP Synthesis Circuit [29]

2.4 Properties of Line Spectrum Pairs

As mentioned in the previous section, LSP polynomials possess many interesting properties that make LSP analysis more efficient than LPC analysis [26][34]. Referring to Figure 3 below, these properties may be summarized as following [14] [15]:

1. all zeros of the LSP polynomials are on the unit circle;
2. zeros of $P_p(z)$ and $Q_p(z)$ are interlaced; and
3. if 1 and 2 are preserved, the minimum phase[2] property of the LPC polynomial $A(z)$ is preserved [19] [36].

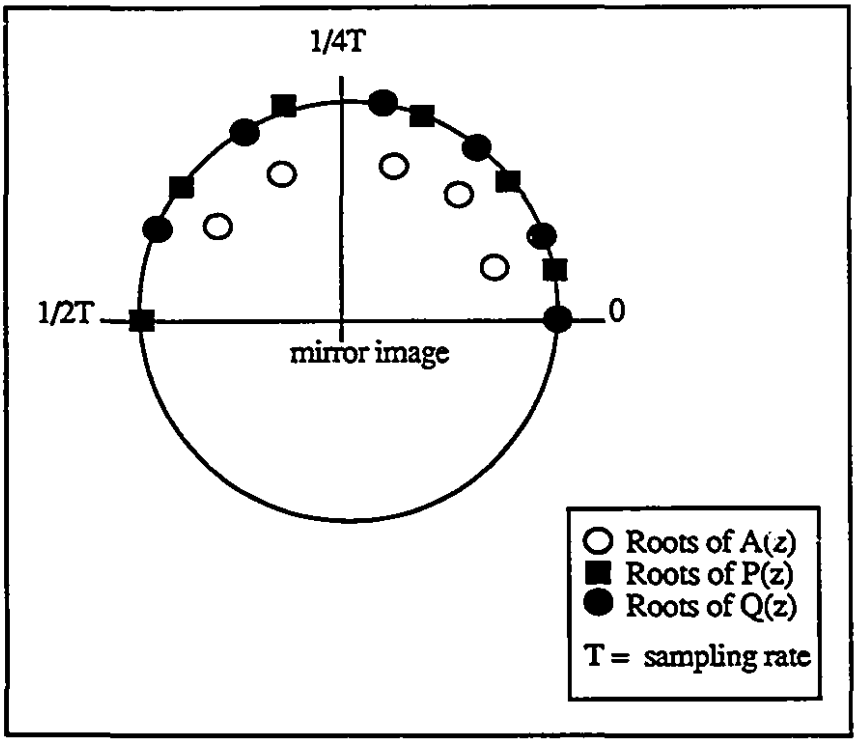

Figure 3: A decomposition of the roots of a tenth-order LPC analysis filter into two sets of roots along the unit circle. These roots are indicated by the filled in circles and squares. The sampling rate is 1/T. Only roots from 0 to 1/2T are shown since identical roots exist between 1/2T to T, with a mirror image about the x-axis. [1] [10] [14]

[2] A transfer function with both its poles and zeros inside the unit circle is called minimum phase. A minimum phase function is the unique transfer function which has a causal and stable inverse.

A mathematical proof of these properties is provided by Soong and Juang in [34].

When compared to linear predictive or PARCOR analysis which correspond to the vocal-tract area function, the line spectrum pair concept has a greater correspondence to the physical relation between the LSPs and the formants of the speech spectrum. A comparison of LSP parameters versus PARCOR parameters reveals a number of advantages. The following represents a summary of the main properties of LSP frequency parameters:

1. The cumulation of two or more frequencies within a certain region in the frequency domain implies the occurrence of formants there. Line spectrum pairs, LP coefficients, and PARCOR coefficients are essentially the same in that they describe the all-pole spectrum. Figure 4 provides an example of a speech spectrum using a Fourier transform and LPC analysis. Vertical lines correspond to the position of line spectrum pairs [8] [15].

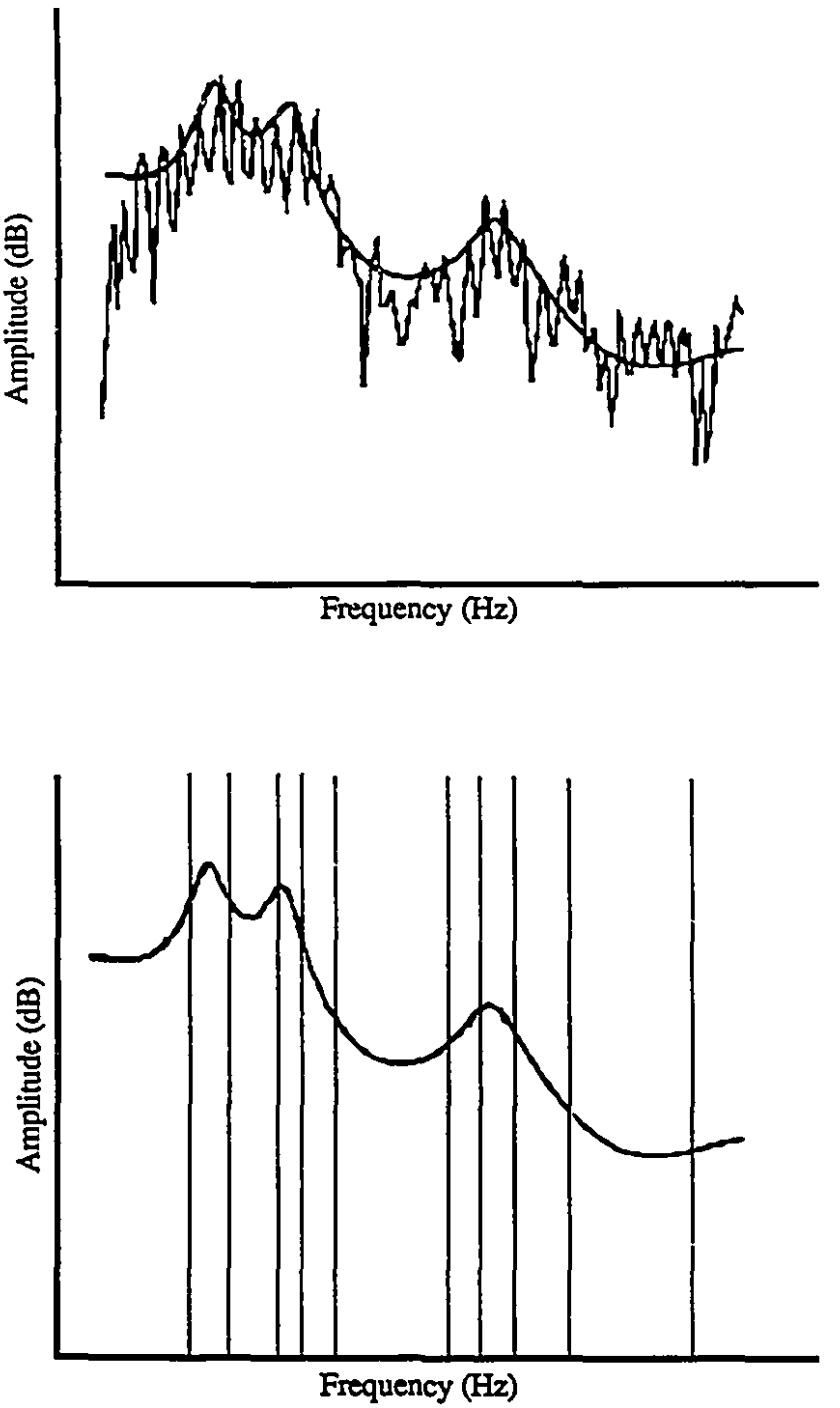

Figure 4: Vowel Spectrum with Corresponding LSPs [29]

2. Minimal spectral distortion by parameter quantization - unlike LPC, LSP are very tolerant of error. LSP quantization error affects the

spectrum only in the immediate frequency region which corresponds to the parameter quantized [6] [14] [15] [27][40];

3. Uniform spectral sensitivity [29];
4. Low spectrum distortion at low bit-rate coding. When LSPs are used for speech coding, they are found to be better than PARCOR by 40% [11][29].
5. Small spectrum distortion for linear interpolation of parameters - LSPs are also tolerant of rapid changes in the speech spectrum - this means that no smoothing or interpolation is required at concatenated speech boundaries [6][40]; hence, LSPs produce less distortion when they are roughly quantized and linearly interpolated [11] [15] [18] [30].
6. Stability for $\omega_1 < \omega_2 < ... \omega_p$ (where ω_p is the LSP frequency for the synthesis filter) [1] [30].
7. Even though they are directly related to LPC parameters, LSP are frequency-domain parameters like formants. Hence the existing vast body of knowledge about the spectral properties of speech can be applied [6].
8. The transformation from LPC coefficients to LSP frequencies is completely reversible; therefore, it is possible to compute exactly the LPC coefficients from the LSP frequencies [26][29].
9. LSP have a well-behaved dynamic range and a filter stability preservation property [11].

However, LSP parameters also possess the following limitations:

1. Dependence of ω_p on the order of LPC analysis [29];

2. A more complicated process of parameter extraction from LPC coefficients for predictor orders greater than 8 [25]; and
3. The requirement of twice the number of registers of one multiplexing lattice PARCOR synthesis system [29].

2.5 Practical Methods for the Direct Determination of LSP Coefficients

LSP frequency domain parameters possess frequency-selective spectral-error characteristics that allow for efficient quantization in accordance with auditory perception. Most methods used to calculate line spectral frequencies require an LPC analysis to produce LP coefficients. This is then followed by an arithmetic procedure requiring the evaluation of the roots of the symmetrical and anti-symmetric polynomials, $P_p(z)$ and $Q_p(z)$. These roots can be found by employing the well known techniques of [31]:

1. solving the polynomials directly;
2. applying the Newton-Raphson method;
3. Fast Fourier Transform method; or
4. Inverse cosine transformation (demonstrated in the previous section).

There are other nonconventional yet effective methods that have been used to find the roots of the symmetric and antisymmetric polynomials. One such method was proposed by Soong and Juang that uses a discrete cosine transform using a fine frequency domain grid [36]. Also, Kabal and Ramachandran employ an iterative root finding technique on a series representation in Chebychev polynomials [13].

Chan and Law provide yet another method to recursively determine the LSP coefficients directly from the LPC reflection coefficients [2]. The line spectrum is generated using a set of reflection coefficients recursively, and the line spectrum frequencies are subsequently found by using a simple zero crossing criterion. The advantage of using reflection coefficients is that they are bounded to ±1, allowing for implementation using fixed point arithmetic [2].

Finally, Saoudi, Boucher and Le Guyader have developed an efficient algorithm to compute LSPs without having to compute the predictor coefficients. In their study, they compared various methods to calculate the LSP coefficients of $P_p(z)$ and $Q_p(z)$. The most efficient method results from using LSP parameters extracted directly from the eigenvalues of tridiagonal matrices whose entries are found using the split-Levinson algorithm. The roots of these polynomials are then found using some of the methods described above [1] [5].

2.6 Discussion of LSP Properties

The properties of the LSP may be used to gain efficiencies in many different areas such as speech coding, speech recognition, and speaker recognition. This section will discuss properties of the line spectrum pair coefficients and frequencies through an examination of some of these applications.

2.6.1 Use of LSP Properties for Low Bit Rate Coding

Low bit rate parametric speech coders generally use a short-time speech spectral envelope to decorrelate the speech source from the vocal tract. The acoustic interaction between source and vocal tract produces only secondary effects, and can thus be

considered independently (the articulatory phonetics associated with the vocal tract are considered in the next chapter) [24]. The model for low bit rate parametric speech coders is usually an all-pole digital filter whose coefficients are derived using Linear Prediction analysis. The linear predictive coefficients are then transformed into an alternative format, quantized, and transmitted as side information. The side information generally accounts for a large portion of the overall bit rate for low bit rate speech coders (typically 25% for Code Excited Linear Predictive coders) [32].

The LSP parametric representation exhibits a number of interesting deterministic and statistical properties, including ordering and limited dynamic range. When these properties are used within source coding, some very efficient schemes are possible. For example, current scalar differential coding schemes can achieve negligible distortion encoding at approximately 30 bits per frame. Vector quantization based schemes can produce good quality speech at approximately 25 bits per frame [32].

LSPs may be quantized more efficiently if the frequency-dependent auditory perception characteristics are exploited. Because the human ear cannot resolve differences at high frequencies as accurately as at low frequencies, the higher frequency LSPs can be more coarsely quantized without introducing audible speech degradation. The amount of frequency variation that produces a just-noticeable difference[3] (JND) is approximately linear from 0.1 to 1 kHz, and increases logarithmically from 1 to 10 kHz.[14][15][24].

LSP parameters have been used to achieve efficient quantization for low to medium transmission bit rates (6.4 to 16 kbps). Equivalent parameter sets such as LPC, PARCOR or LSP correspond to different analysis filter structures, and much work has gone into finding predictor coefficients that best characterize the properties of the speech

---

[3] In psychophysical experiments subjects distinguish between one sound over another. These sounds are varied along one or more acoustic dimensions such as sound pressure levels (intensity) and frequency. The acoustic value at which 75% of responses are identified as different is called the *just noticeable difference*, or JND.

signal. Any parameter set must be chosen in such a way that the corresponding analysis system should have a relatively fast convergence speed in order to deal with the non-stationarity of the speech signal. Also, in order to ensure stability of the synthesis filter, a test should be available to check the phase of the analysis filter (require minimum phase). Finally, the parameter set must not be prone to quantization error to allow for efficient transmission [3].

LPC and PARCOR parameter sets satisfy these requirements; however, LSP parameters are superior to LPC and PARCOR parameters with respect to their quantization and interpolation properties as a function of spectral distortion. LSP parameters can also provide high-quality synthesized speech at low transmission rates. The conventional approach to calculating LSP parameters, however, has been to compute LPC coefficients, followed by a cosine transform. This has made the numerical procedure for finding LSP coefficients using a real-time implementation of an LSP analysis-synthesis speech codec very difficult [3].

Ching and Ho have proposed a split-path adaptive filter configured as a cascade of second-order sections, where the filter parameters are essentially equivalent to the LSP coefficients [3]. The advantage of this type of filter is with respect to its reduced complexity, since LSP coefficients can be obtained by adapting the filter coefficients sample by sample. Another advantage to this method is that a parallel processing architecture could be employed to improve the adaptation speed of the system. The adaptive LSP method proposed by Ching and Ho has potential applications for medium-to-low bit rate speech codecs which could be used for efficient land mobile radio voice communications using a narrowband UHF channel.

While it is easier to quantize due to the uniform sensitivity across the frequency spectrum, there are also disadvantages associated with the propagation of errors across the frequency spectrum, which results in a reduced quantization efficiency. In order to

deal with this, Soong and Juang have proposed an alternative solution that exploits the properties of LSP frequency differences, as is described in the following section.

2.6.2 Use of Transitional LSP Properties for Low Bit Rate Coding

When the LPC spectrum is expressed in terms of predictor coefficients, reflection coefficients, or log area ratios, spectral variation due to the deviation of a single parameter is not usually obvious [29]. The numerical deviation of parameters may lead to a widespread spectral variation in the frequency domain. Because all parameters require quantization in any coding scheme, the use of a sensitivity model for a single parameter becomes ineffective in guiding a scalar quantizer design if error coupling among parameters is found to be severe.

When the LSP model is utilized in speech applications, it is generally the instantaneous LSP frequencies that are employed. The property that the $(i+1)^{st}$ parameter is always greater than the i^{th} parameter implies that LSP parameters are not independent, but correlated. This correlation is sometimes referred to as *intraframe correlation* of LSP parameters. LSP vectors from adjacent frames are also correlated (*interframe correlation*) [7][18]. A differential coding scheme is one in which the differences between adjacent LSP frequencies are encoded instead of the absolute values. In using this method, one observes less divergence in the differential LSP values, thus resulting in a greatly reduced dynamic range for quantization, and hence a greater degree of system performance [8] [14] [35].

Soong and Juang have observed that, unlike other spectral parameters, spectral error caused by single parameter deviation in LSP frequencies displays a localization within a certain frequency range. That is, perturbing any single LSP produces spectral

variation dominated by only neighboring LSP frequencies. The importance of this single parameter sensitivity property may be exploited in efficient scalar quantizer design because variation in spectral patterns due to parameter variation is not significant. Soong and Juang utilized this property in the design of a globally optimal LSP quantizer [35].

2.6.3 Use Of Transitional LSP Parameters In Speech Recognition

As in speech coding applications, when differential LSPs are utilized in speech recognition systems, additional information is provided when the change in the spectral envelope is extracted from the LSP model in the form of first order derivatives of the LSP frequencies [8]. Gurgen, Sagayama and Furui have proposed a new feature vector defined by the linear combination of transitional and instantaneous LSP vectors [11]. This vector is used in speech recognition experiments for speaker-independent isolated word recognition systems and was found to provide superior results over equivalent cepstrum coefficient vectors.

Paliwal has reported that a combination of both transitional and instantaneous spectral parameters produces excellent recognition. In his study of speech recognition, Paliwal found that the best results were achieved using the LSFs as instantaneous parameters, and delta-cepstral coefficients as the transitional parameters [27].

2.6.4 Use of LSP Properties in Speaker Recognition

An interesting utilization of LSP frequencies is with respect to the unique problems related to automatic speaker recognition. Voice variation among people is dependent on many factors such as sex, age, and mother tongue, and may be used to

distinguish one speaker from another. In speaker recognition, the properties of the speech utterance are analyzed. The most important characteristic features, that are also relatively easy to extract, involve the energy, duration, pitch, and the spectrum of the speech waveform. Of these parameters, spectral information is probably the most useful in speaker recognition. Speech signals are segmented into quasi-stationary signals and are modeled using a linear predictive model using parameters such as LPC, PARCOR or LSP coefficients [21][22]. In a paper written by Lin, Liu, Huang, and Wang [21] it was found that the use of LSP parameters provides superior speaker recognition results.

They propose a method for person identification by analyzing speech using LSPs. Characteristics of the speaker's spectral information is represented by line spectrum frequencies. The LSP frequencies are then used to model the spectral distribution of each speaker. Next, the special properties of LSP frequencies are used to build a codebook used in the training phase for a text-independent speaker identification system. LSP coefficients are also extracted in the recognition phase which are compared with the values in the code book. This study provides a comparison of various distance measures that are used to find the minimum distance between the input speech signal characteristics and those found in the codebook, resulting in the possible recognition of the speaker.

2.7 Mathematical and Cognitive Relationships of LPC and LSP Frequencies

This section will discuss some observations of the frequency relationships between LPC and LSP frequencies. Some of the conclusions have been backed by mathematical proof, whereas other relationships are the result of observation and analysis of empirical data. This section will describe some of these conclusions. At the end of this section, the link of this work with previous research will be made clear. An

examination of the power spectra of Figure 5 below reveals that typically two to three LSP frequencies cluster about the formant frequency, and that the bandwidth of the given formant depends on the closeness of the corresponding LSP frequencies. This suggests that the LSP provides much more information about the spectrum than the formant spectrum because other factors about the formants (through LPC and LSP analysis) are used in the characterization of the spectrum [26].

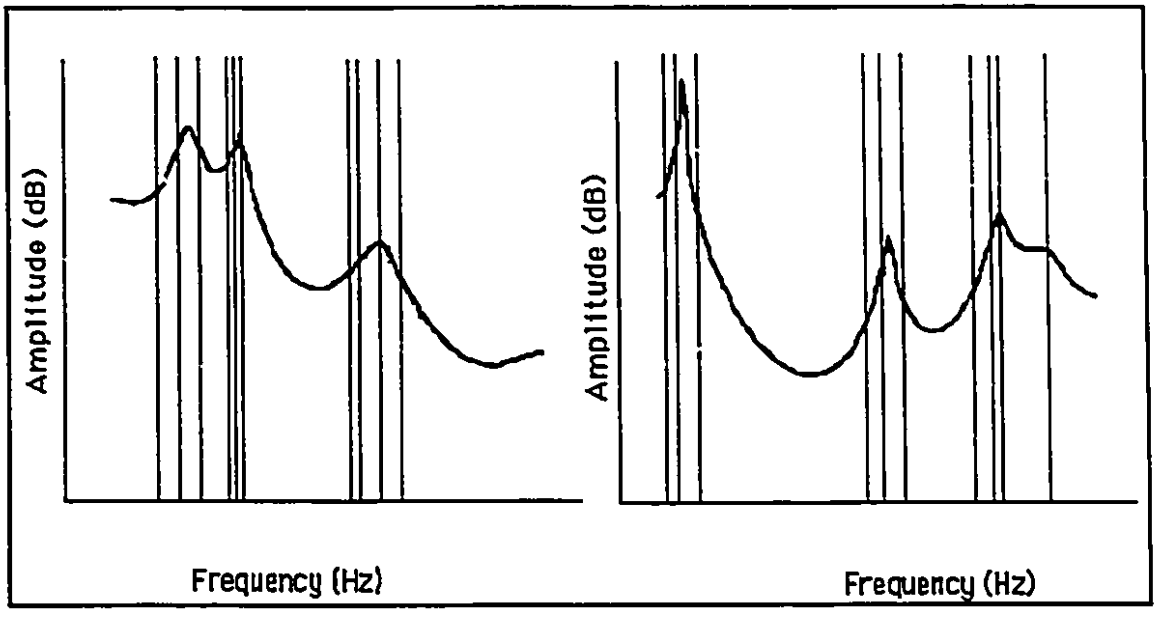

Figure 5: LSP frequencies superimposed on the corresponding formant spectrum [29].

A review of the existing literature indicates that some knowledge-based information exists on how the LSP and LPC frequencies interact in general. In order to understand this further, let us examine the LPC spectrum magnitude, which can be expressed as [10][25]:

$$|S(\omega)|^2 = \frac{\sigma^2}{|A_p(\omega)|^2} = \frac{4\sigma^2}{|P_p(\omega)+Q_p(\omega)|^2} \qquad (2.28)$$

and depends on the whole LSP distribution according to the relation:

$$|S(\omega)|^2 = \frac{2^{-p}\sigma^2}{\left[\cos\left(\frac{\omega}{2}\right)\prod_{\substack{i=1\\ odd}}^{p-1}(\cos\omega - \cos\omega_i)\right]^2 + \left[\sin\left(\frac{\omega}{2}\right)\prod_{\substack{i=2\\ even}}^{p}(\cos\omega - \cos\omega_i)\right]^2} \qquad (2.29)$$

where σ refers to the LPC residual root mean square value.

This equation leads to a number of interesting trends in the distributions of LP and LSP frequencies. From this equation, if two or more lines are close to one another, both denominator terms tend to zero for $\omega \rightarrow \omega_i$ and the magnitude becomes large. Also, when two LSP are clustered the following is true:

1. The average of a pair of line spectrum frequencies does not always approach the corresponding zero of $A(z)$, even though they may be very close to it. Also, if a zero of $A(z)$ is near the unit circle, the corresponding root locus branch remains in its neighborhood. In fact, the LSP frequencies tend to cluster around angular positions of the zeros of $A(z)$[20];
2. It is not surprising that the frequency response of the entire analysis filter is affected by a change in each LPC parameter, whereas a change in each root affects the spectrum near that particular frequency only [14];
3. Closely spaced line spectra are located in the vicinity of the resonant frequencies, whereas sparsely spaced line spectra are located near valleys of the spectral envelope [14];

4. For even predictor order, the odd LSP frequencies (corresponding to the closed glottis model) are located near a low formant center frequency. When the formant center frequency is in the higher part of the spectrum, the role of the odd and even LSP frequencies are interchanged [25];
5. In the medium frequency range, the position of the formant center frequencies is influenced by odd and even LSPs equally [25];
6. If there are three LSP frequencies located near a formant, the middle LSP frequency is generally closest to the formant center frequency [25]; and
7. For an odd predictor order, odd LSP frequencies are dominant in both the high and the low spectrum [25].

Since the existing cognitive information is limited, the present research will carry further this work by investigating how the frequencies tend to behave on a vowel-by-vowel basis. One of the most important speech characteristics for vowels is information about the location and distribution of formants in frequency domain. This research will provide a categorization pertaining to specific patterns in the way LSP frequencies relate to the closest corresponding formant, F1, F2, or F3. In drawing relationships between formants and their corresponding LSPs, yet another indicator is made possible to exploit such that, when incorporated within other statistical or knowledge based techniques, may serve to enhance the effectiveness of speech systems.

This study will use the LSP model in conjunction with the principles of articulatory phonetics to draw conclusions regarding the relationships between LPC and LSP frequencies. This is done by first finding the LSP and LPC frequencies of a training set, and by using articulatory phonetics to draw conclusions regarding the

relationships between the two frequencies. The same procedure is followed for a test set. This is followed by a comparison of the test set results with the results of the training set. The greater the similarity between the test and training set results, the greater the accuracy of the analysis.

The next chapter will provide some theoretical background about the articulatory and acoustic phonetics utilized for the analysis of the LPC and LSP frequencies. This will be followed by a description of how the experiment was conducted, as well as the results found.

Chapter 3

Articulatory and Acoustic Phonetics

Nearly all speech sounds result from the power of the respiratory system which pushes the airstream out of the lungs. Air is pushed up through the trachea, through the larynx, and passes between the two small muscular folds, the vocal cords. The air passages above the larynx are considered to be the vocal tract. The speaker produces a non-stationary signal which changes characteristics based on the contraction and relaxation of the muscles within the vocal tract. Sounds are produced through a complex interworking of the vocal cords, tongue, lips velum and the jaw. When the vocal cords are well apart, air has a relatively free passage through the vocal tract, and the sounds produced are considered to be voiceless, or unvoiced. Most consonants are unvoiced sounds. When the vocal cords are narrowed, the pressure of the airstream causes them to vibrate periodically; the sounds produced are considered to be voiced, producing vowel and sonorant sounds. Because this thesis deals exclusively with voiced sounds, the balance of this chapter will provide background on the aspects of articulatory and acoustic phonetics used in the analysis of vowels (and diphthongs) [17][24].

3.1 Articulator Movement

Components of the vocal tract that participate in the formation of sounds are described as *articulators*. The two main articulators involved in the formation of vowel sounds are the tongue and the lips. The lip muscles, affect speech through the closure of the vocal tract when the lower lip presses against the upper teeth, or when they are pressed together. Also, the lips help control the making of vowels due to their ability to round, protrude, retract or spread [24].

The muscular tongue structure, possessing great freedom of movement, also plays an important role in the creation of the different vocal tract shapes required in the formation of the different vowel sounds. The tongue possesses intrinsic muscles that allow it to change shape. It is extremely flexible and is capable of assuming many different configurations, often in less than 50 ms. The tongue may be divided into four different components, the *apex* (or tip), the *blade*, *dorsum* and the *root*, which function somewhat independently. The most agile part of the tongue is the apex. It is thin and narrow and is able to establish and break contact with the palate up to about nine times per second, and is the quickest part of the tongue. The surface of the tongue is called the dorsum. The anterior portion of the dorsum is called the blade, and is also very mobile. The body of the tongue, or the root, serves to position the dorsum which helps in forming constrictions in different areas of the vocal tract. [6][7] The figure below shows the various parts of the lips and tongue involved in the formation of vowel sounds [17][24].

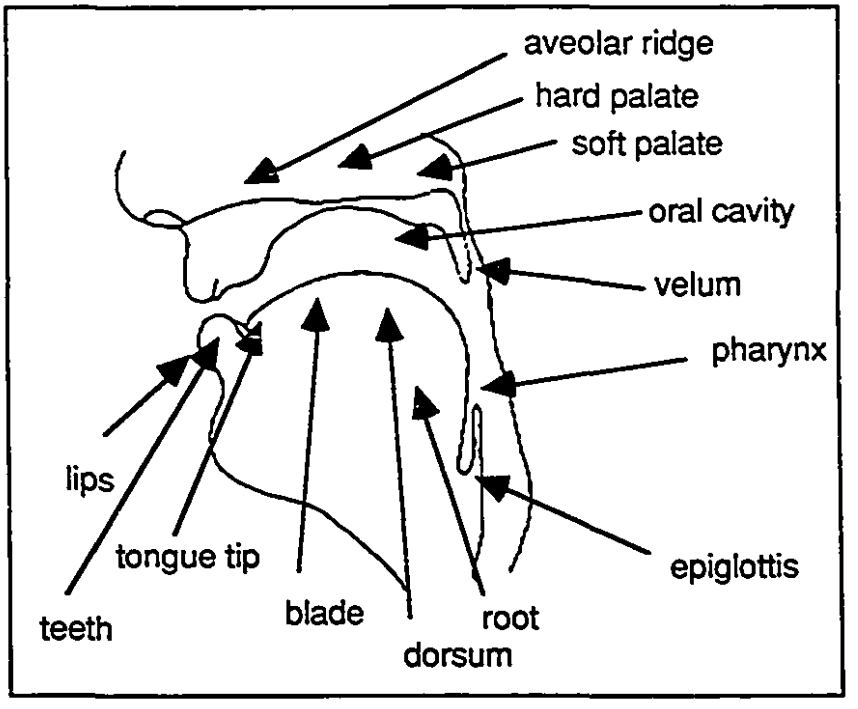

Figure 6: Principal elements of the vocal tract involved in the articulation of vowel sounds. A two dimensional mid-line, or mid-sagittal view of the tongue is shown here. Note that the vocal tract is a tube where the other parts of the tongue or lips may be very different from the center position. The epiglottis, which is attached to the lower part of the root, is included for completeness. [1] [17] [24]

Vowel place of articulation (POA) refers to the physiology related to the making of phonemes. This physiology relates to the horizontal position of the tongue body, which may be described as forward, middle or back, as well as lip constriction. The height of the tongue, described in terms of high or low is also important as is the degree of rounding of the lips. This ranges typically between 1 (for the most rounded lip configuration) to 5 cm^2 (most open lip configuration). The figure below shows typical articulatory positions which describe high and low tongue height, as well as front and back tongue position [24].

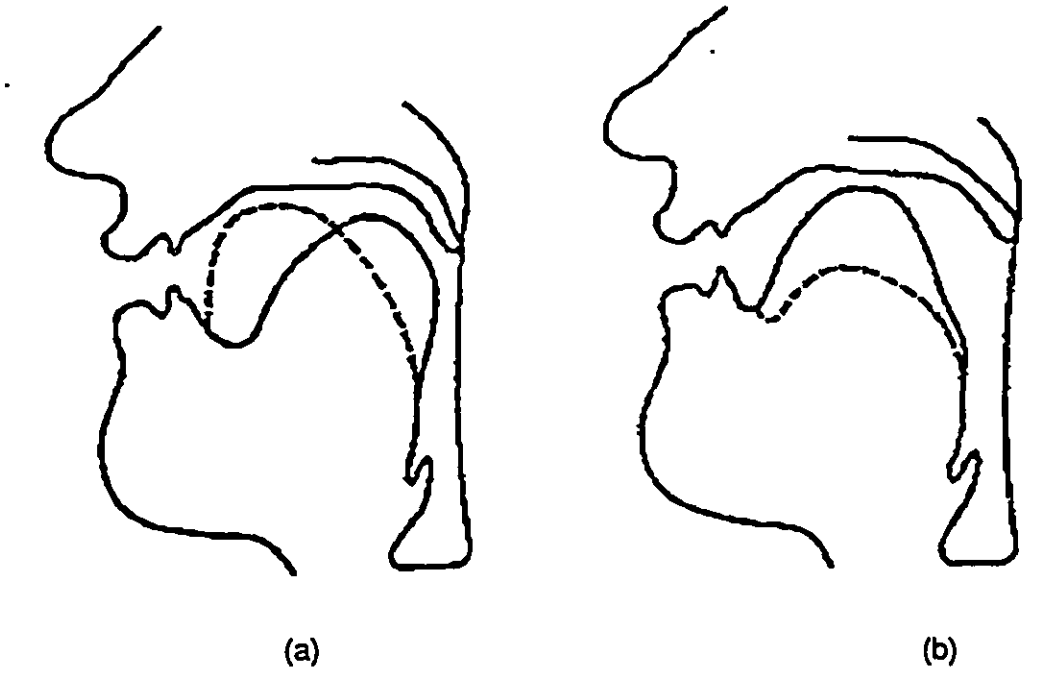

Figure 7: Articulatory positions showing a) tongue position [front, back], and b) tongue height [high, low] [24].

3.2 Articulatory Phonetics - The Vowel Space

Unlike consonants, one of the main problems in describing vowels is that there are no distinct boundaries between different vowel sounds. The movement from one vowel to another is achieved by the movement of the tongue and lips; however, it is very difficult to describe the exact movement of the tongue. This is why vowels are labeled according to their different auditory qualities which are based on the relative position of the tongue or the lips, and not their actual position.

For example, the vowel /iy/ as in "meet" is called high front because it has a high and front articulatory position. Similarly, the vowel /ae/ as in the word "status" has an articulatory quality that may be called low front. The phoneme /eh/ as in

"presented" sounds somewhere between /iy/ and /ae/, but slightly nearer to /ae/, and may be described as mid low front. Also, the vowel /aa/ as in "father" may be described as a low back vowel, whereas the vowel /uw/ in "blue" is a high, but fairly back vowel. As is described by the following diagram, the four phonemes /iy/, /uw/, /aa/ and /ae/ describe four boundaries of a space that describes the relative articulatory position of vowel sounds [17] [24] [39].

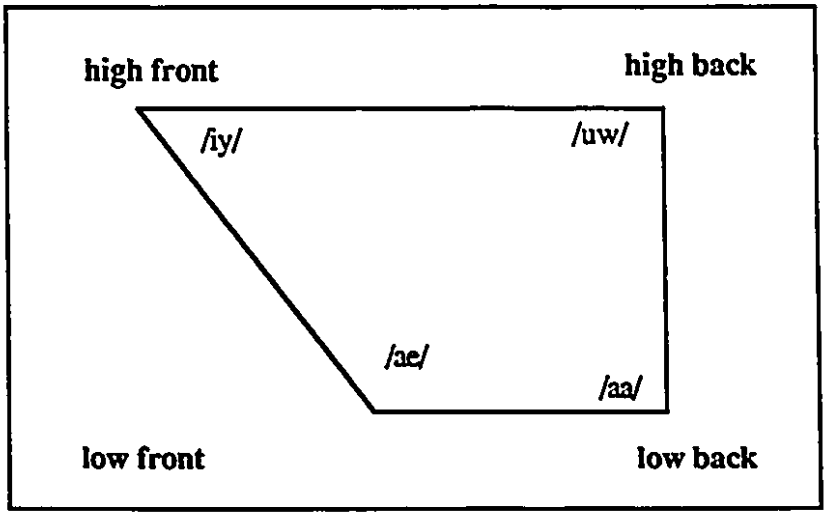

Figure 8: Vowel space showing the auditory qualities of vowel sounds. Sound identifiers do not describe the absolute articulatory position of each vowel; rather, it is the relative position between phonemes that is depicted [17].

The notion of an auditory vowel space can be used to position other vowel sounds relative to the extreme boundaries delineated by /iy/, /ae/, /aa/, and /uw/. Figure 9 below shows the relationship of different phonemes with respect to the vowel space defined in Figure 8. Figure 9 describes the vowel uttered by a typical Midwestern American male or female. The solid points refer to those phonemes that are considered to be monothongs, while the solid lines refer to those phonemes generally considered to be diphthongs. A diphthong is described as a single phoneme that consists of the tongue and mouth migrating from one vowel sound to another. For

the purposes of this study, the phonemes /oy/, /ay/, and /aw/ are considered to be true diphthongs. There are, however, some differences of opinion as to the status of the phonemes /ey/, /ow/, and /iu/.

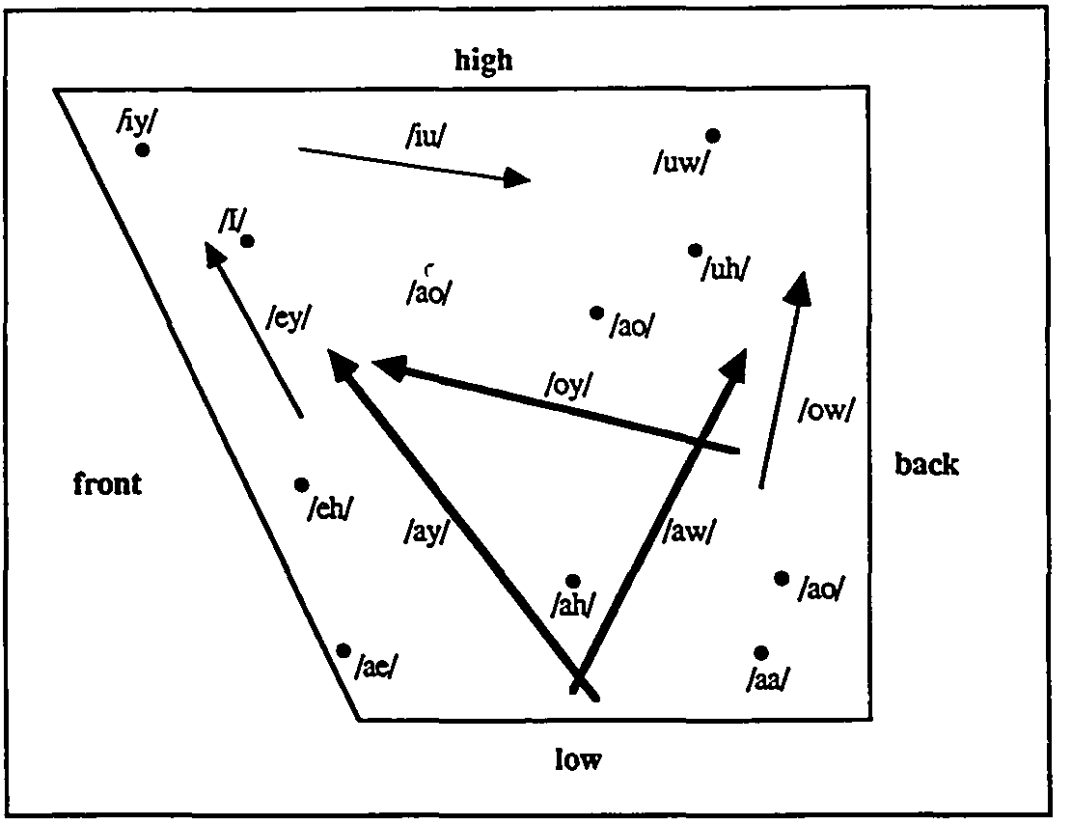

Figure 9: Vowel space showing auditory qualities of certain vowel sounds. Thick, solid arrows indicate true diphthongs (/oy/, /ay, /aw/) [17] [24].

The phonemes /ey/ and /ow/, represented using thin lines in Figure 8, are more difficult to classify as monothongs or diphthongs. While it depends very much on the speaker, the first part of a diphthong is usually more prominent than the latter. In some phonemes, however, the last half of the diphthong is so transitory and brief that it is difficult to determine whether a given phoneme is a diphthong or monothong. In this study, the /ey/ and /ow/ phonemes will be considered to be monothongs. This is because phonemes are averaged over time, and over several speakers, thereby

significantly minimizing the impact of the latter half of the diphthong on the samples studied.

There is some controversy among phoneticians as to whether the phoneme /iu/, as in "cue", which is represented by the weakest arrow at the top of Figure 5, is actually a diphthong or a glide plus a vowel. This phoneme is different from other diphthongs because it is most prominent during the latter part. In this study, /iu/ is not considered as diphthong because of the duration of the phoneme which is very transient compared to that of other vowel sounds. This would reduce the number of samples acquired, and, hence, affect the quality of its analysis. This phoneme is therefore classified as a glide, and was not considered as part of the body of vowel sounds in this thesis.

Finally, the retroflex phoneme, /er/ as in "luxurious", does not fit exactly into the vowel space in Figure 8. This phoneme cannot be described in simple terms such as back, front, high, or low. This vowel is produced using an additional feature unknown as the rhotarization, characterized by the high bunched position of the tongue. The upward curling position of the tip of the tongue is what makes the rhotarized phoneme retroflex. While /er/ is not shown in Figure 8, it was nevertheless considered to be a mid-central vowel, and is included in this study.

The table below summarizes the phonemes used in this study and their corresponding place of articulation. An additional feature of vowel tenseness or laxness was added in the phoneme POA column as another means of identification. Vowels were divided into tense and lax depending on their duration and on the degree of tongue displacement from the central, or schwa, position.

Table 1: Place of articulation of vowel and diphthong phonemes (middle column indicates words used in this study)

Phoneme	Word	POA
iy	meet	high/front/tense
ih	Cliff	high/front/lax
ey	stayed	mid/front/tense
eh	presented	mid/front/lax
ae	status	low/front/tense
aa	yacht	low/back/tense
ao	bought	low/back/rounded
ow	tugboats	mid/back/tense/rounded
uh	books	high/back/lax/rounded
uw	blue	high/back/tense/rounded
ah	enough.	mid/back/lax
er	luxurious	mid/tense/retroflex
ax	the	mid/lax/schwa
ay (diphthong)	item	low/back to high/front
oy (diphthong)	boy	mid/back to high/front
aw(diphthong)	about	low/back to high/back

3.3 Acoustic Phonetics

Vowel sounds are distinguished essentially by the position of their first three formants, with the first two containing the greatest energy. These formants vary from one speaker to another because of the varying physiologies of the vocal tract. Generally,

different speakers manifest different F1 and F2 positions. However, in an experiment described by O'Shaughnessy which studied ten monosylabic words of the form /hVd/ for sixty speakers, it was shown that the same F1 and F2 were perceived as different phonemes for different speakers. This leads to the conclusion that while the first two formants are kept apart from each other for an individual speaker (which is essential in allowing listeners to differentiate between phonemes uttered) the pitch, formant bandwidth, and the position of other formants (especially F3) are also important for this differentiation to occur. This latter point, the relative positions of F1 and F2, also allows listeners to distinguish phonemes uttered by the same speaker.

A plot of the first two formants reveals an interesting similarity with the articulation of phonemes. Figure 8 is superimposed onto Figure 9 in order to demonstrate how F1 and F2 vary with the position of phonemes. Figure 10 shows that F1 is inversely proportional to vowel height. Also, the intensity of energy in vowels decreases as tongue height decreases over 4 to 5 dB. The relatively wide separation between the first three formants, as well as the -6 dB spectrum falloff results in the greatest energy residing in F1, while the lower vowels have significant energy presence in the F2 (F3 is not shown in Figure 10 below). F2 also exhibits a significant correlation with the extent of backness of the vowel; this correlation, however, is not as strong as with F1 and vowel height [17] [24].

This chapter has described articulator movement, articulatory phonetics and the vowel space and aspects of acoustic phonetics. The next chapters will discuss how the experiment was conducted, as well as the results obtained. The articulatory and acoustic phonetic relationships of phonemes described in this chapter will be used to explain these results.

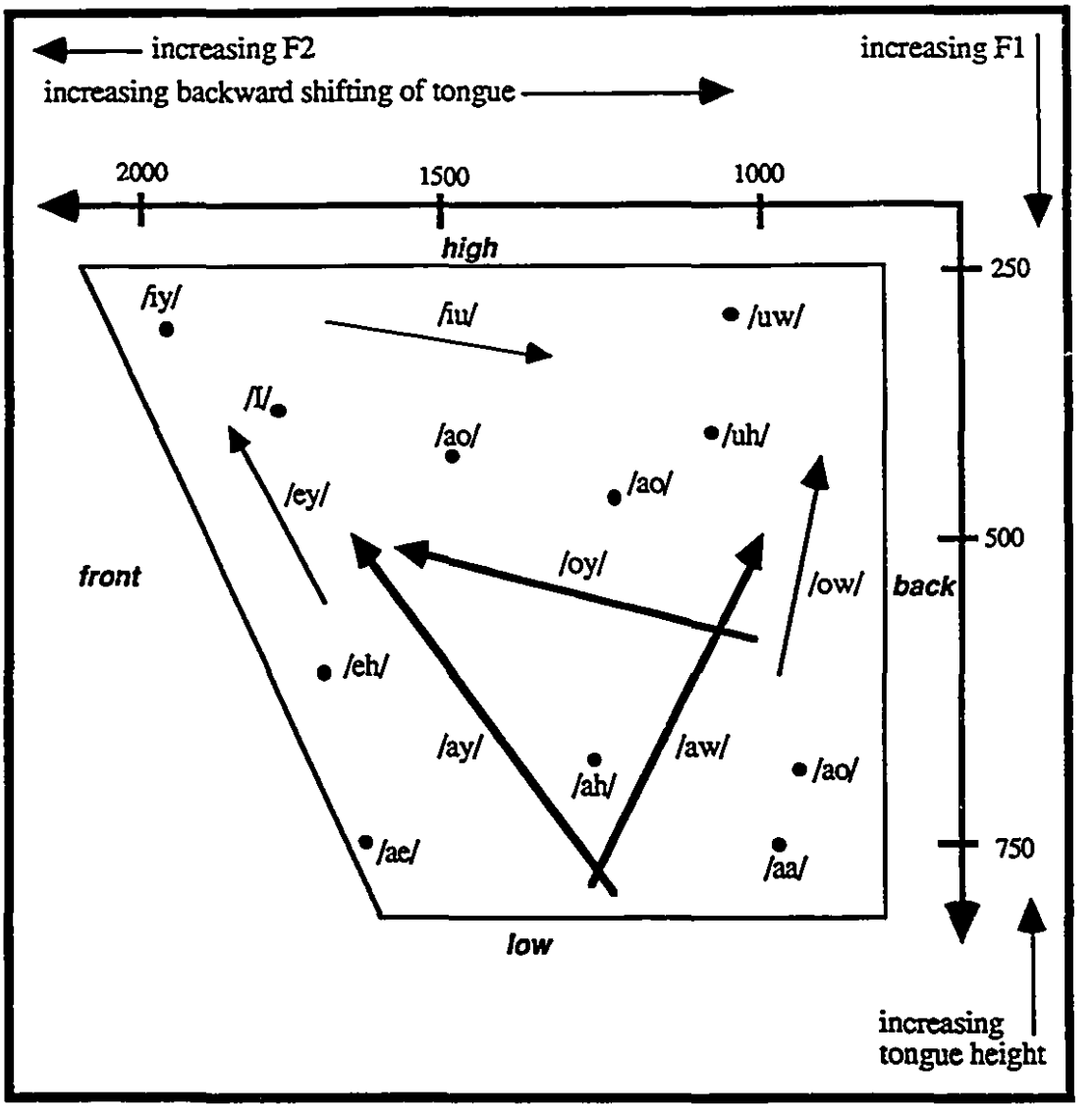

Figure 10: Formant chart indicating the frequencies of F1 and F2 and their relationship to articulatory positions. Generally, different speakers manifest different F1 and F2 positions; therefore, the frequencies described in this figure are not absolute, and indicate the order of magnitude and relative vowel position only [1] [17].

Chapter 4

Data Acquisition and Preprocessing

This chapter deals with how the data used in this study was attained and processed prior to the correlation procedure that is described in detail in Chapter 5. In order to provide the reader with an overall appreciation of the general methodology employed in this experiment, an overview of the procedure used in this experiment is provided in Figure 11 below. This chapter will deal with the first box on the left which deals with finding the LSP and LPC frequencies.

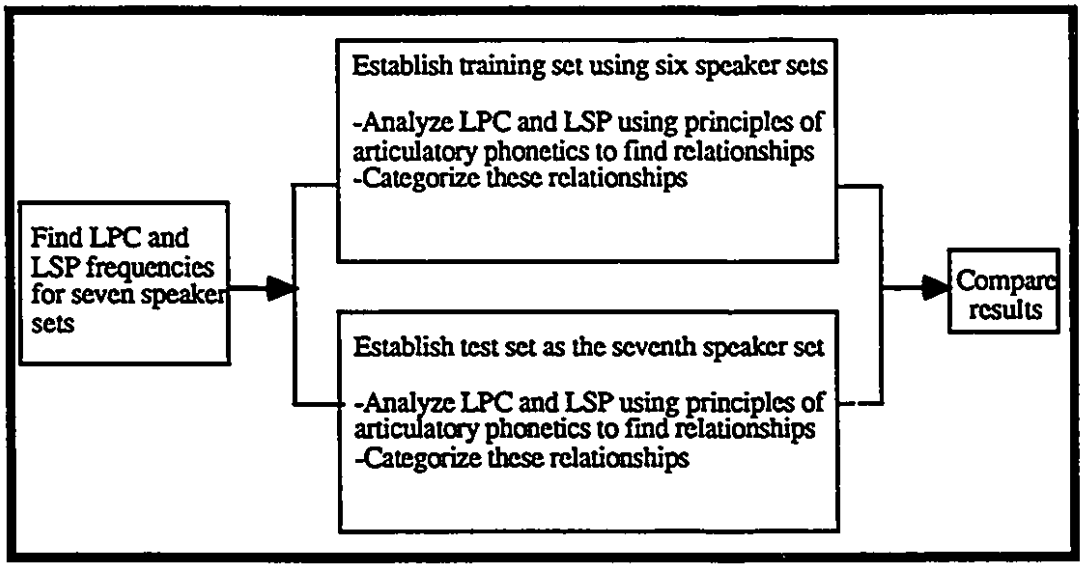

Figure 11: Overview of procedure used in this research.

The speech data used in this study consists of 112 phoneme utterances, involving 7 repetitions by different male and female speakers, of 16 vowel sounds. The vowels were manually acquired from sentences found in the TIMIT[4] database as in Table 2 below:

Table 2: Phonemes and Their Associated Sentences

Phoneme	Sentence
iy	I know I did not meet her early enough.
ih	Cliff was soothed by the luxurious massage.
ey	The groundhog clearly saw his shadow, but stayed out only a moment.
eh	As co-authors, we presented our books to the haughty audience.
ae	Upgrade your status to reflect your wealth.
aa	Many wealthy tycoons splurged and bought both a yacht and a schooner.
ao	Many wealthy tycoons splurged and bought both a yacht and a schooner.
ow	Tugboats are capable of hauling huge loads.
uh	As co-authors, we presented our books to the haughty audience.
uw	We like blue cheese, but Victor prefers Swiss cheese.
ah	I know I did not meet her early enough.
er	Cliff was soothed by the luxurious massage.
ax	The toothpaste should be squeezed from the bottom.
ay	Shaving cream is a popular item in Halloween.
oy	The small boy put the worm on the hook.
aw	George is paranoid about a future gas shortage.

3 The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) corpus of read speech was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT resulted from the joint efforts of several sites under sponsorship from the Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO). The design of the text corpus was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT, and is maintained, verified and prepared for CD-ROM by the National Institute of Standards and Technology (NIST).

Speech utterances from six of the seven sets of speakers were used as a training set. The seventh set of speakers was used as the test set. The distribution of males and females over each set of speakers and dialect region (dr) is shown in Table 2 below. The TIMIT American dialect regions correspond to the following numbering scheme:

1 New England
2. Northern
3. North Midland
4. South Midland
5. Southern
6. New York City
7. Western
8. "Army Brat" (moved around)

The LSP frequencies were found using a FORTRAN program used at INRS, lspgen.for, which was modified for this experiment. This program generated LSP frequencies for each entire sentence using the following parameters:

Order of LPC:	14
Sampling frequency:	16000 Hz
Frame length:	25.6 ms
Frame advance:	5.0 ms
Size of FFT[5]:	512 points

[5] Fast Fourier Transform

The LSP frequencies corresponding to the exact phoneme sample frames were then isolated using another FORTRAN program coded by the author. Steady state vowels were used in this research to enable the study of vowel formants without the effect of other speech components, thus eliminating any effects of coarticulation. Time-aligned word and phonetic transcriptions of each TIMIT sentence were used for this purpose.

Table 3: Distribution of Speaker Set Sex and Dialect Region

Phoneme	Speaker Set Sex and Dialect Region (dr)														Total	
	1	dr	2	dr	3	dr	4	dr	5	dr	6	dr	7	dr	Male	Female
iy	f	1	f	2	m	4	m	4	m	7	m	7	m	8	5	2
ih	f	1	m	3	m	3	f	4	m	5	m	6	m	6	5	2
ey	m	2	m	3	m	4	m	4	f	5	m	6	m	7	6	1
eh	m	2	m	3	m	6	m	6	m	7	m	7	m	8	7	0
ae	f	1	f	1	m	2	m	2	m	3	m	6	f	7	4	3
aa	m	2	f	3	f	4	m	4	m	5	m	5	f	8	4	3
ao	m	2	f	3	f	4	m	4	m	5	m	5	f	8	4	3
ow	m	1	m	1	f	2	m	3	m	3	f	5	m	5	5	2
uh	m	2	m	3	m	6	m	6	m	7	m	7	m	8	7	0
uw	m	1	f	3	m	3	f	5	f	7	f	7	m	7	3	4
ah	f	1	f	2	m	4	m	4	m	7	m	7	m	8	5	2
er	f	1	m	3	m	3	f	4	m	5	m	6	m	6	5	2
ax	f	1	m	3	m	3	m	4	f	6	f	7	m	7	4	3
ay	m	2	m	2	f	5	f	5	f	5	f	7	m	7	3	4
oy	f	1	f	2	m	2	m	3	m	6	m	7	m	8	5	2
aw	m	1	m	1	f	2	m	3	m	3	f	5	m	5	5	2

The exact phoneme samples provided by the TIMIT corpus were also used to identify the LPC frequencies which were extracted using the Waves program using the following parameters:

Order of LPC:	14
Sampling frequency:	16000 Hz
Analysis type:	Autocorrelation
Window type:	Hamming
Window size:	Variable based on size of phoneme duration.

The results from the 112 phoneme utterances were placed in 112 LSP files and 112 LPC files. The training set was obtained using a MATLAB program which averaged the results for each phoneme over 6 speaker sets to yield a total of 16 LPC files (consisting of seven LPC formants for each phoneme), and 16 LSP files (consisting of 14 LSP frequencies for each phoneme). The test set also consisted of the same data uttered by the seventh set of speakers. The diagram below summarizes the data acquisition procedure for this experiment. As mentioned briefly in Figure 11 above, 32 training set and 32 test set files were then correlated and compared. This analysis is described further in the next chapter.

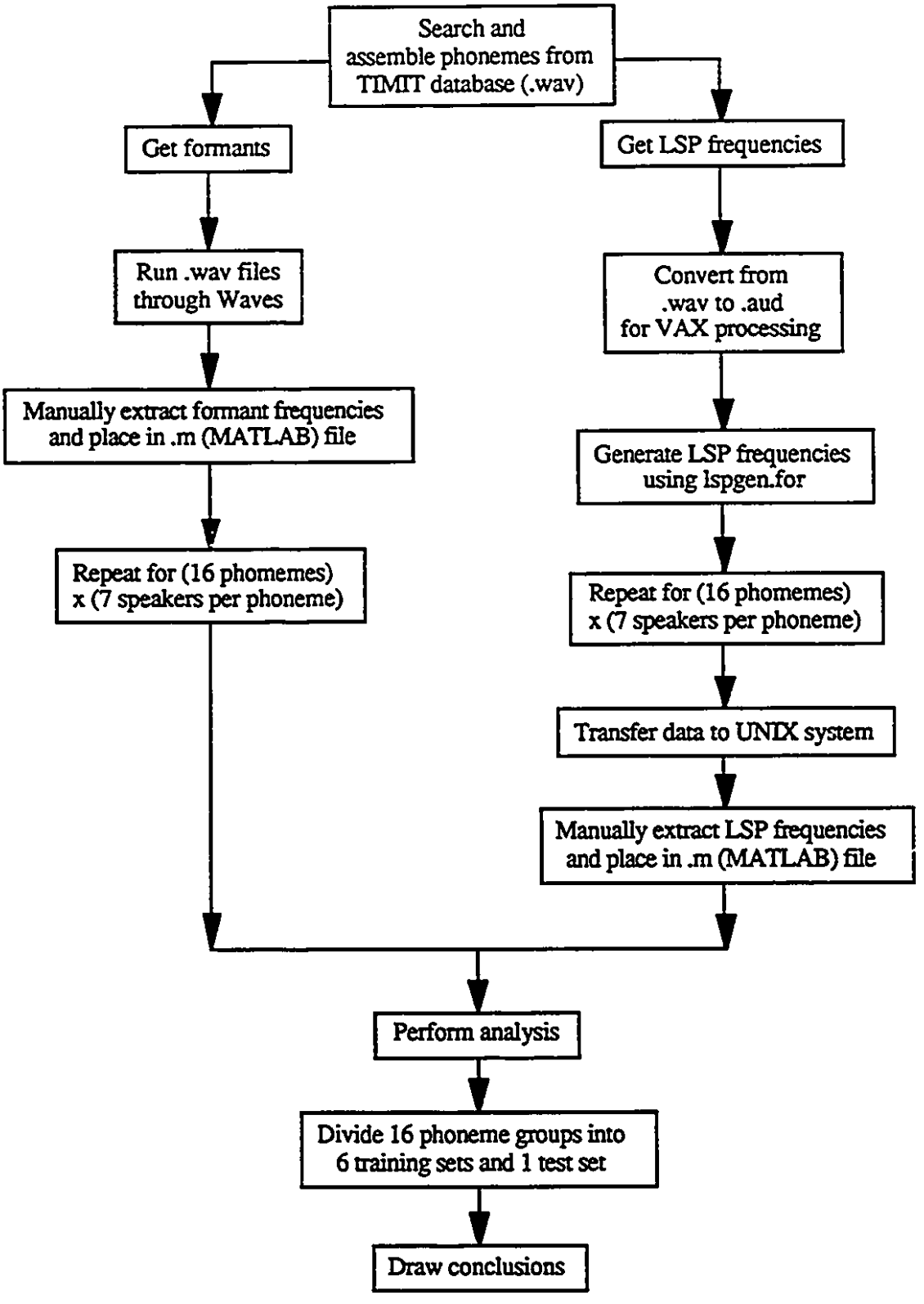

Figure 12: Summary of data acquisition procedure used for this experiment

Chapter 5

Results

This chapter will discuss the results of the experiment and present acoustic-phonetic reasons for the conclusions derived. The next few sections will deal with how the data was correlated prior to its analysis. The procedure was essentially the same for each of the training and the test set data. Initially all frames were considered to be useful. A distance measure was then established that would be used to extract the mean and variance of the data. The mean and variances of the useful frames were examined to establish a bound. A table (described in section 5.1) was used to select useful training frames whose distances were smaller than the bound. Useless frames whose distances are greater than the bound were discarded.

5.1 Training Set Correlation

The training set utilized six speaker sets of files derived from the TIMIT database. At this point, the LPC files which contained seven formants were reduced to three formants, because the first three formants contain the highest energy and are an important distinguishing factor for phonemes [14] [15]. These first three LPC formants were used in the correlation analysis with the LSP frequencies. The LSP frequency files were not shortened at this stage in order to observe the behavior of the 14 LSP frequencies with the first three LPC frequencies.

The LPC and LSP frequencies were studied for any consistencies that may have occurred. It was expected that the first five or six LSP frequencies would be correlated very closely with the LPC formants, but it was not clear how these frequencies would exactly line up. Each LPC formant was compared with its closest LSP frequency, and was noted. The degree of correlation between these frequencies was also identified to eliminate useless data. The bounds used for this elimination were based on observation of the results to include as many cases as possible, and are reported in Table 4 below.

Table 4: Scale used to define degree of correlation between LPC and LSP frequencies.

Frequency	Degree of Correlation	Symbol Used
within ±10 Hz	very very close correlation	vvc
within ±30 Hz	very close correlation	vc
within ±60 Hz	close correlation	c
within ±90 Hz	marginal correlation	mc
greater than ±90 Hz	no correlation	nc

The values for correlation were chosen to be the most subjectively meaningful in the sense that small and large distance correspond to good and bad subjective correlation, respectively. Because of the energy associated with the first formant, and its small bandwidth, the first formant frequency was expected to be extremely close to the first or second LSP frequencies, and hence very tight bounds (within ±10 Hz or ±30 Hz) were chosen. As frequency increases, the bandwidths of the formants and LSP roots also increase. This means that there is a loss of precision in the determination of exact frequencies due to averaging effects. Therefore, the other bounds (within ±60 Hz or ±90 Hz) were chosen to take this into account [9]. The final bound (greater than ±90 Hz) represented a maximum tolerance which produces JND between sounds [14] [15] [24]. Therefore, frequencies separated by more than 90 Hz probably would not add to the effectiveness of the speech system, and were eliminated. Other factors considered in choosing these groupings involved the tractability of the distance

measure such that it would be amenable to mathematical analysis and lead to a practical design, as well as its ease of implementation in a real system [9] [16]. Therefore, the LSP and LPC frequencies were averaged across seven speakers and reported in Table 5 below:

Table 5: LPC/LSP Training Set Frequency Correlations (frequencies in Hz).

	LPC Formant 1 (F1)			LPC Formant 2 (F2)			LPC Formant 3 (F3)		
	LSP	*f*	level	LSP	*f*	level	LSP	*f*	level
iy	LSP1	449	c/vc	LSP4	2372	c	LSP6	2954	vc
ih	LSP1	519	vc/vvc	LSP3	1437	vc			nc
ey	LSP1	502	c	LSP4	1999	mc	LSP5	2506	c
eh	LSP1	557	c	LSP3	1583	c	LSP5	2351	c
ae	LSP1	664	vc	LSP3	1597	c/vc	LSP5	2364	vvc
aa	LSP1	716	vc/vvc			nc			nc
ao	LSP1	666	vc	LSP3	1097	c	LSP6	2897	c
ow	LSP1	528	vc	LSP3	1307	vc	LSP5	2383	vvc
uh	LSP1	485	c			nc	LSP5	2450	mc
uw	LSP1	440	mc	LSP3	1679	vc/vvc	LSP5	2600	mc
ah	LSP1	699	vvc	LSP3	1428	vvc			nc
er	LSP1	439	mc	LSP3	1555	mc	LSP4	1787	nc (very weak)
ax	LSP1	586	vvc	LSP3	1249	vvc	LSP6	2863	c/vc
ay	LSP1	721	vc			nc			nc
oy	LSP1	555	vc	LSP3	1090	c	LSP5	2526	mc
aw	LSP2	668	vvc	LSP3	1183	c	LSP6	2314	vvc

5.2 Training Set Analysis

The correlations shown in Table 5 were based on similarities apparent in the distribution of the LPC frequency correlation with corresponding LSP frequencies. A cognitive approach was used to determine any correlation or similarities. The following summarizes the findings of the analysis and explains possible reasons for the correlation.

5.2.1 Type 1 - Phonemes /aa/, /ay/

One of the most obvious correlations between LPC and LSP frequencies can be observed between the phonemes /aa/ and /ay/ in which F1 is very high, and F2 and F3 possess no correlation (nc). Hence, the overall observation for this Type 1 vowel is:

a) F1 was among the highest for both phonemes, at approximately 720 Hz;

b) There was no correlation for F2 or F3; and

c) Relative positions of all formants and LSP frequencies were the same.

A possible explanation for this correlation between /aa/ and /ay/ is due to similarities in place of articulation (POA). The vowel /aa/ and the diphthong /ay/ are sounds characterized (at least in part for the diphthongs) by low tongue height, and horizontal tongue position towards the back of the mouth. Also, because the articulation of /aa/ requires the maximum mouth opening range (about 5 cm^2), it is not surprising that it also provides one of the most obvious correlations.

5.2.2 Type 2 - Phonemes /ao/ and /ax/

Another type of correlation was observed between the phonemes /ao/ and /ax/. In both of these cases F3 correlated with LSP6, at approximately 2880 Hz. The POA again provides a possible explanation for this correlation. Both the vowel phonemes /ao/ and /ax/ are characterized by the tongue height being in the mid position. Also, these phonemes may be described as lax with respect to their duration and the minimal degree of displacement from the neutral, or schwa, position. Indeed, the phoneme /ax/ is the only schwa phoneme in the data set. The relative positions of all formants and LSP frequencies were also observed to be the same.

5.2.3 Type 3 - Phonemes /iy/ and /ey/

A strong correlation was observed for the phonemes /iy/ and /ey/. These were the only phonemes where in both cases F2 correlated with LSP4. (The two frequencies themselves, however, were not similar). There are many similarities between these vowels that may explain this correlation. Both of these vowels are high and unrounded, with horizontal tongue position towards the front of the mouth. Also, these phonemes are amongst the longest in duration, and are produced with more extreme articulation positions than most other phonemes. Relative positions of all other frequencies, except F2/LSP4 and F3/LSP4, were the same.

5.2.4 Type 4 - Phoneme /er/

The phoneme /er/ stood out from the rest of the training data set. This was the only phoneme where there was a very weak F3 correlation with LSP4, at 1786 Hz. This phoneme is the only retroflex vowel in the set. The curl of the tip of the tongue is responsible for the very low F3 frequency. Also, there were marginal correlations observed for F1/LSP1, and F2/LSP3.

5.2.5 Type 5 - Phonemes /eh/ and /ae/

There was a correlation for the /eh/ and /ae/ phonemes. In both cases:

a) F2 correlated with LSP3 at approximately 1590 Hz; and

b) F3 correlated with LSP5 at approximately 2358 Hz.

The similarities between these phonemes may be attributed to the fact that they are both unrounded, and are front vowels. They are also characterized by a fairly low tongue height. Also, the relative positions of all formants and LSP frequencies were not the same. This was because there was a very minor variation in the exact location of the formant frequencies with respect to LSP1, LSP3 and LSP5.

5.2.6 Type 6 - Phoneme /aw/

The phoneme /aw/ stood out from the rest of the data set because of its F1 correlation:

a) Unlike all other phonemes, F1 correlated with LSP2 at 678 Hz;

b) F2 correlated with LSP3 within ±60 Hz; and

c) F3 correlated very well with LSP6 at 2314 Hz.

The phoneme /aw/ is a diphthong which consists of a changing vowel sound in which the tongue and lips move between /aa/ and /uh/. The /aa/ vowel sound may be responsible for the relatively high F1 frequency. This result was unusual since LSP1 was much lower than F1 at 569 Hz (a difference of approximately 100 Hz); in all other cases LSP1 had at least a marginal correlation with F1.

5.2.7 Type 7 - Phonemes /ih/ and /oy/

A classification for /ih/ and /oy/ was observed due to the presence or absence of F1 and F3 frequency correlations. Specifically,

a) F1 correlated with LSP1 at approximately 537 Hz;

b) F2 correlated to LSP3, however, within different bounds; and

c) There was marginal or no correlation for F3.

It is not surprising to find this similarity since the diphthong /oy/ consists of the vowels /ao/ and /ih/. The vowel /ih/ in both of these phonemes may be described using POA as being characterized by a high tongue height and a horizontal tongue position towards the front of the mouth.

5.2.8 Type 8 - Phoneme /ah/

The phoneme /ah/ is the one of the phonemes that contains an extremely high F1. While there are similarities between Type 1 and /ah/, /ah/ was placed in a special class because of its difference in comparison with the Type 1 F2 frequency. Hence, the main observation for /ah/ is that F1 correlates with LSP1 at one of the highest frequencies, 699 Hz. Note that F1/LSP1 and F2/LSP3 were also found to be very closely correlated.

5.2.9 Type 9 - Phonemes /ow/, /uh/ and /uw/

These phonemes did not fit into any of the above classifications. There were no simple correlations observed for any of these three phonemes. However, the POA similarity among these vowels is interesting. These phonemes are all rounded, with the horizontal tongue position towards the back of the mouth. Another interesting similarity between these phonemes is that all frequencies (except for /ow/) F3 occurred in the same order.

For the purposes of speech recognition, one may speculate that if the phonemes do not follow any of the aforementioned rules, that there is a good chance that the phoneme is one of these rounded, back vowels.

Table 6 below provides a summary of the above observations.

Table 6: Summary of Observations

Type	Phoneme(s)	Observations
1	/aa/, /ay/	• F1 was amongst the highest for both phonemes, at approximately 720 Hz; • There was no correlation (nc) for F2, F3; • Relative positions of all formants and LSP frequencies were the same.
2	/ao/, /ax/	• F3 correlated with LSP6, at approximately 2880 Hz; • Relative positions of all formants and LSP frequencies were the same.
3	/iy, /ey/	• Only cases where F2 correlated with LSP4 (the two frequencies themselves, however, were not similar); • Relative positions of all other frequencies, except F2/LSP4 and F3/LSP4, were the same.
4	/er/	• Weak F3/LSP4 correlation at 1786 Hz; • Marginal correlations for F1/LSP1, and F2/LSP3.
5	/eh/, /ae/	• F2 correlated with LSP3 at approximately 1590 Hz; • F3 correlated with LSP5 at approximately 2358 Hz; • Relative positions of all formants and LSP frequencies were not the same due to a very minor variation in the exact location of the formant frequencies with respect to LSP1, LSP3 and LSP5.
6	/aw/	• F1 correlated with LSP2 at 678 Hz; • F2 correlated with LSP3 within ±60 Hz; • F3 correlated very well with LSP6 at 2314 Hz.
7	/ih/, /oy/	• F1 correlated with LSP1 at approximately 537 Hz; • F2 correlated to LSP3, however, within different bounds; • There was marginal or no correlation for F3.
8	/ah/	• F1 correlates with LSP1 at one of the highest frequencies, 699 Hz; • F2/LSP3 were also found to be very closely correlated; • F3 was not correlated with any LSP frequency.
9	/ow/, /uh/, /uw/	• All frequencies except /ow/-F3 occurred in the same order.

5.3 Test Set Correlation

The seventh speaker set shown in Table 3 comprises the data test set. The frequencies found for the test set were compared with the training set to validate the observations above. The results of the test set correlation are reported in Table 7 below:

Table 7: LPC/LSP Test Set Frequency Correlations (frequencies in Hz)

	LPC Formant 1 (F1)			LPC Formant 2 (F2)			LPC Formant 3 (F3)		
	LSP	*f*	level	LSP	*f*	level	LSP	*f*	level
iy	LSP1	383	c	LSP4	2378	c	LSP6	2914	mc
ih	LSP1	457	c	LSP3	1318	c			nc
ey	LSP1	504	vc	LSP4	1793	vvc	LSP5	2626	vc
eh	LSP1	552	c	LSP3	1523	mc	LSP5	2411	mc
ae	LSP1	559	vvc	LSP4	2113	vvc	LSP6	2944	vvc
aa	LSP1	775	c			nc			nc
ao	LSP1	819	vvc	LSP3	1272	vc	LSP6	2471	vc
ow	LSP1	477	vc	LSP3	1129	vc	LSP5	2527	mc/c
uh	LSP1	491	c	LSP3	1142	mc	LSP5	2567	c/vc
uw	LSP1	392	c			nc			nc
ah	LSP1	705	vc			nc	LSP5	2360	vc
er	LSP1	489	c			nc	LSP4	1691	c
ax	LSP1	554	vc	LSP3	1268	vc/vvc	LSP6	2983	mc
ay	LSP1	665	vvc			nc	LSP4	1552	vc
oy	LSP1	522	vc			nc			nc
aw	LSP1	598	vvc	LSP3	1292	vc	LSP5	2344	mc

5.4 Test Set Performance

The performance of the test set with respect to the training set will be discussed in this section. In order to aid the visualization of the similarities found between the two sets, diagrams of each correlation will be presented at the end of this section.

5.4.1 Phoneme /iy/:

F2 correlated with LSP4 as expected and is correctly within the Type 3 phoneme category.

5.4.2 Phoneme /ih/:

F1 correlated with LSP1 at 457 Hz (as opposed to 537 Hz), which is still in the lower range of frequencies. Also, there was no correlation for F3. Therefore, /ih/ fits within the corresponding training set observations in the Type 7 category.

5.4.3 Phoneme /ey/:

F2 correlated with LSP4 as expected. The vowel /ey/, then, is a Type 3 phoneme.

5.4.4 Phoneme /eh/:

F2 correlated with LSP3 at 1523 Hz, and F3 correlated with LSP5 at 2411 Hz. Hence, /eh/ is a Type 5 phoneme, as expected.

5.4.5 Phoneme /ae/:

F2 correlates very closely with LSP4 at 2113 Hz, unlike the expected F2/LSP3 correlation at 1590 Hz (a 523 Hz difference). Also, F3 correlates very closely with LSP6 at 2944 Hz, unlike the expected F3/LSP5 correlation of 2358 Hz (a 586 Hz difference). It is possible that the test and training set values are uncorrelated because the test case here is a female (thus resulting in higher formant frequencies), unlike the majority of speakers in the training set. Hence, this phoneme does not fall into Type 5 as expected.

5.4.6 Phoneme /aa/:

As expected, there was an F1 was amongst the highest frequencies within the test set at 775 Hz. Also, there was no F2 or F3 correlation. Therefore, /aa/ fits very well into the Type 1 category.

5.4.7 Phoneme /ao/:

F3 for the /ao/ phoneme correlated with LSP6 as expected; however, the correlation frequency was 2471 Hz, and not around the expected 2880 Hz. This difference of 309 Hz may be explained by the fact that this female speaker possesses a South Western American accent, in which the vowel lasts for a longer duration, resulting in a prominent F2, and a relatively more relaxed F3. Given this explanation, the phoneme /ao/ fits into Type 2.

5.4.8 Phoneme /ow/:

While in the training set, /ow/ did not fall into any Type 1 to Type 8 category, the test set put the phoneme /ow/ with F1 correlating with LSP1 at 477 Hz, and F3 possessing a marginal correlation with LSP5. These are the approximate characteristics of a Type 7 phoneme, with a 60 Hz difference in the F1 correlations. The test fails for /ow/ as a Type 9 phoneme because the training set F3 possessed a very close correlation with LSP5. However, the phoneme would pass if /ow/ were classified as Type 7, which may be explained by the similarities in POA (mid back tongue position) between /ow/ and the diphthong /oy/.

5.4.9 Phoneme /uh/:

As expected, /uh/ did not fit into any Type 1 to Type 8 category. Hence, /uh/ is correctly a Type 9 phoneme.

5.4.10 Phoneme /uw/:

As expected, /uw/ did not fit into any Type 1 to Type 8 category. Hence, /uw/ is a Type 9 phoneme.

5.4.11 Phoneme /ah/:

F1 correlates with LSP1 at a very high frequency, 705 Hz. Therefore, /ah/ fits the Type 8 category, as expected.

4.4.12 Phoneme /er/:

There was a weak F3/LSP4 correlation at 1691 Hz. While the expected correlation was at 1786 Hz (a 95 Hz difference), three points are salient here. Firstly, F3 is correlated with LSP4, unlike most other phonemes. Secondly, the third formant had an extremely low frequency due to the retroflex nature of the phoneme. Thirdly, F3 correlations were very weak, hence a 95 Hz difference is acceptable. Thus, /er/ is a Type 4 phoneme.

5.4.13 Phoneme /ax/:

This phoneme fits very well into the Type 2 category since F3 correlated with LSP6 at 2983 Hz.

5.4.14 Phoneme /ay/:

As expected, F1 was very high at 665 Hz. Also, there was no frequency correlation exhibited for F2. However, unlike the training set observations, there was a very strong correlation of F3 with LSP4. Hence, this phoneme fits only partially into the Type 1 phoneme category.

5.4.15 Phoneme /oy/:

As expected, F1 correlated with LSP1 at 533 Hz. Also, there was no correlation for F3. Hence, /oy/ fits well into Type 7, as expected.

5.4.16 Phoneme /aw/:

This correlation failed to fall into the expected Type 6 category. This could be due to the Southern American dialect region of the test male speaker.

5.5 Graphical Representations of Training and Test Sets

The diagrams on the following pages show graphically how the test set performed vis à vis the training set. A discussion of these results follows the graphs.

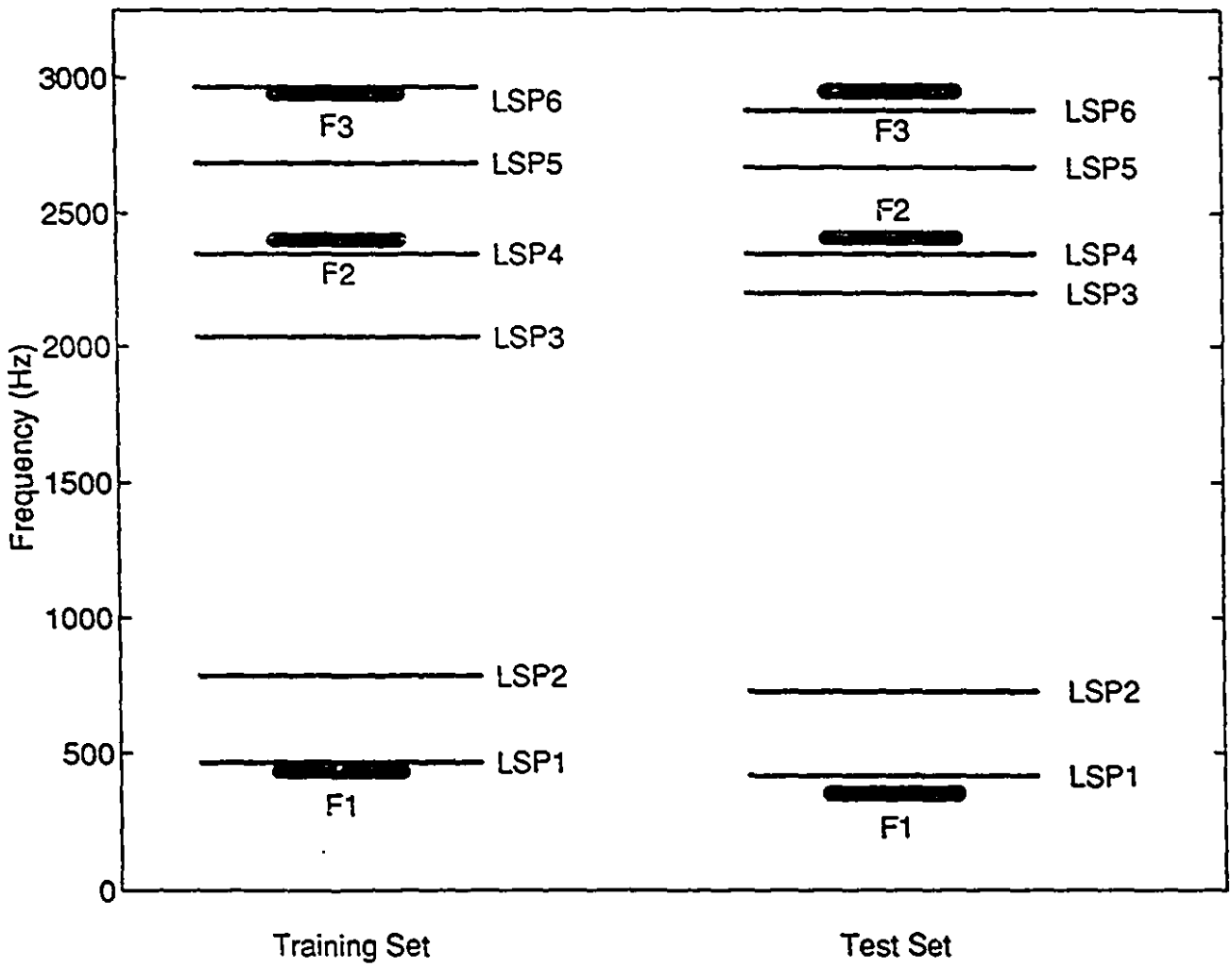

Figure 13: Relationship Between LSP and LPC Frequencies - Phoneme /iy/

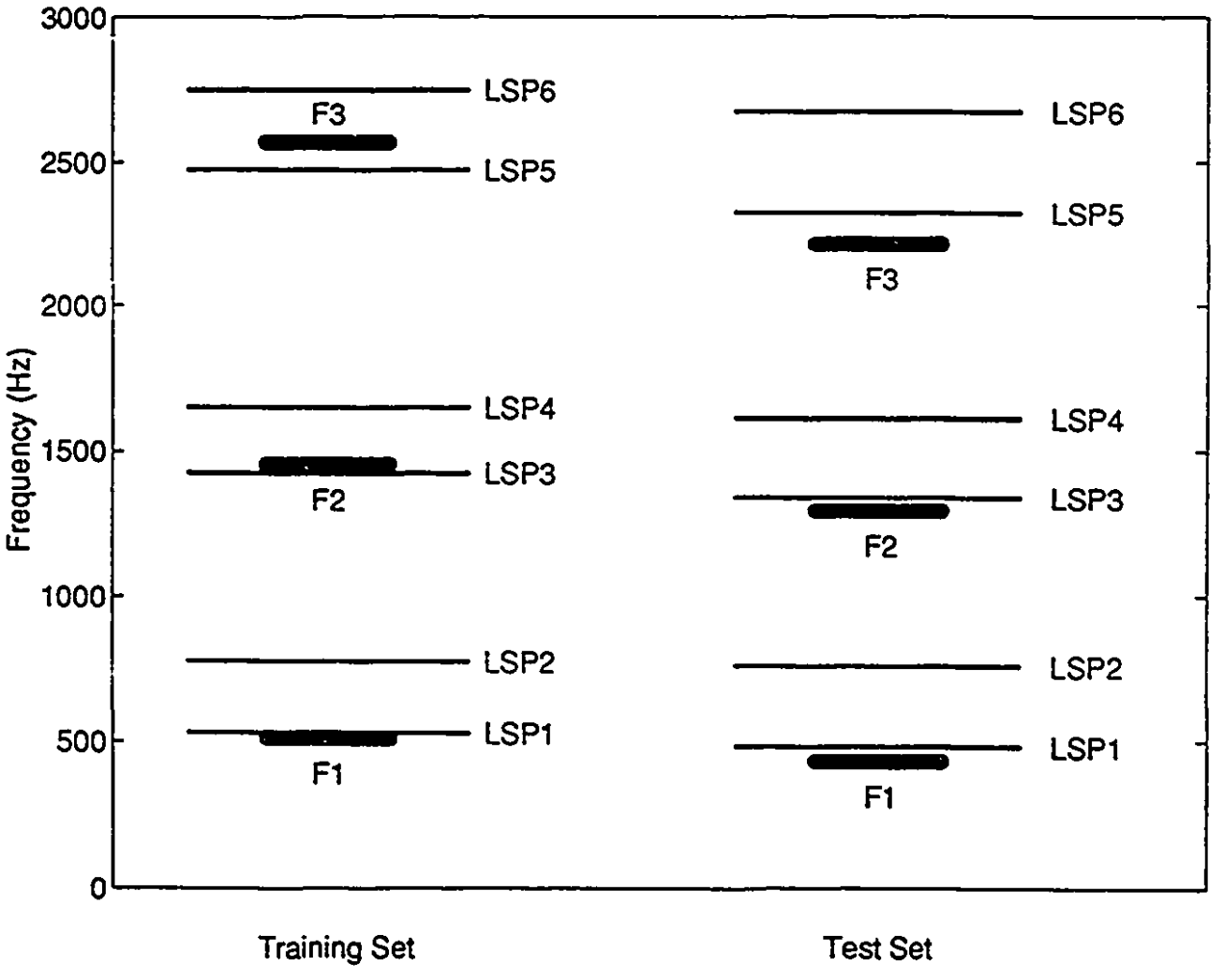

Figure 14: Relationship Between LSP and LPC Frequencies - Phoneme /ih/

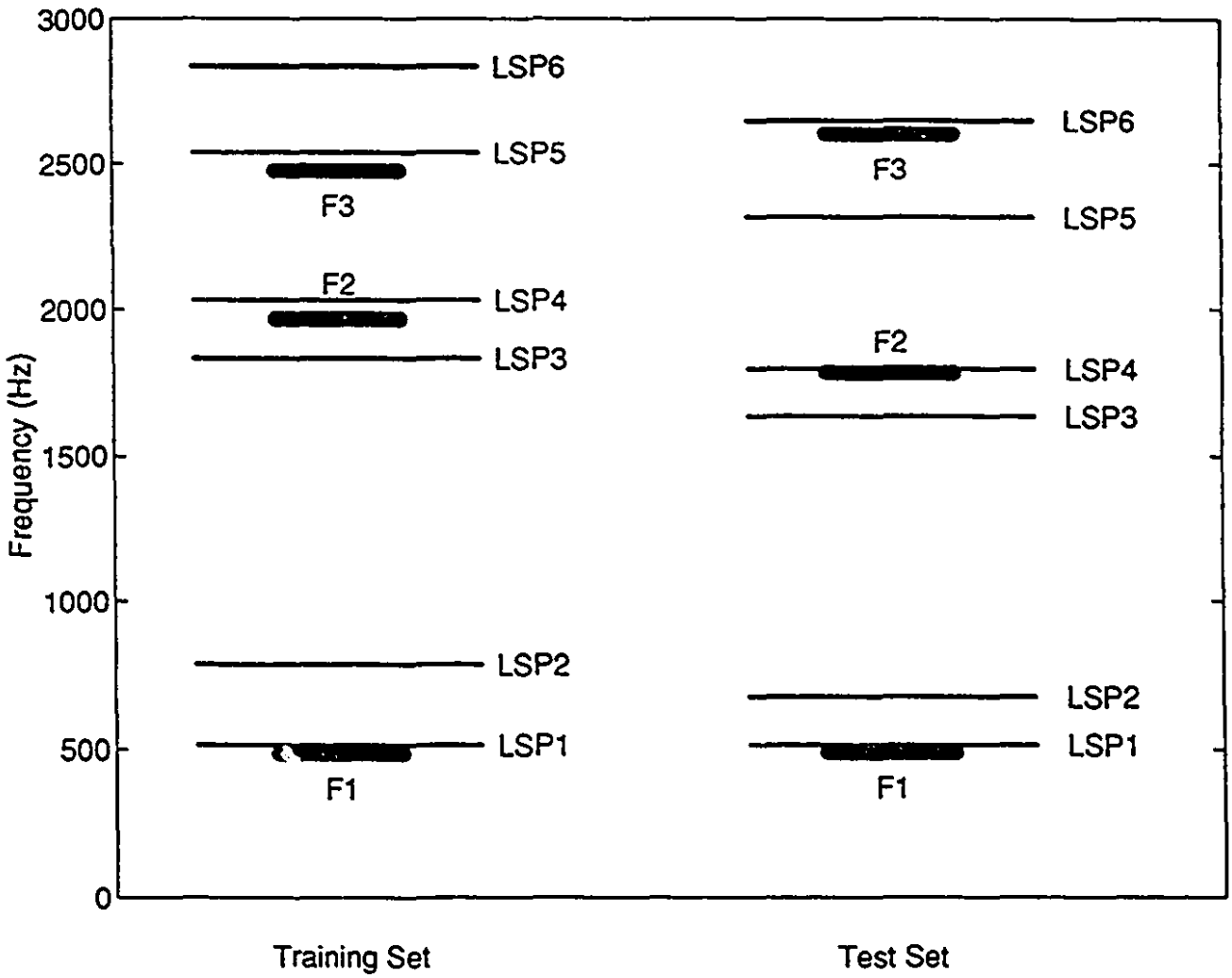

Figure 15: Relationship Between LSP and LPC Frequencies - Phoneme /ey/

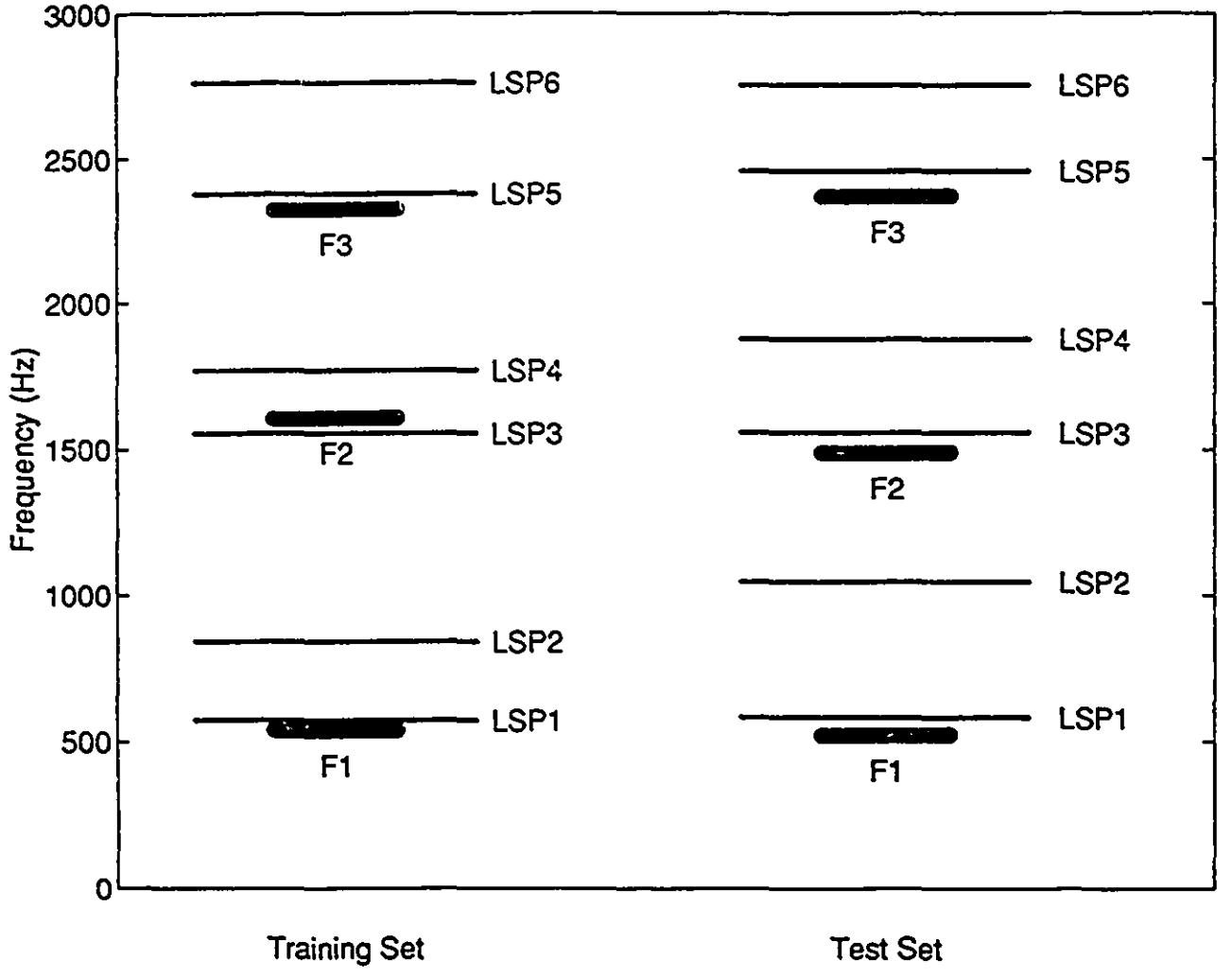

Figure 16: Relationship Between LSP and LPC Frequencies - Phoneme /eh/

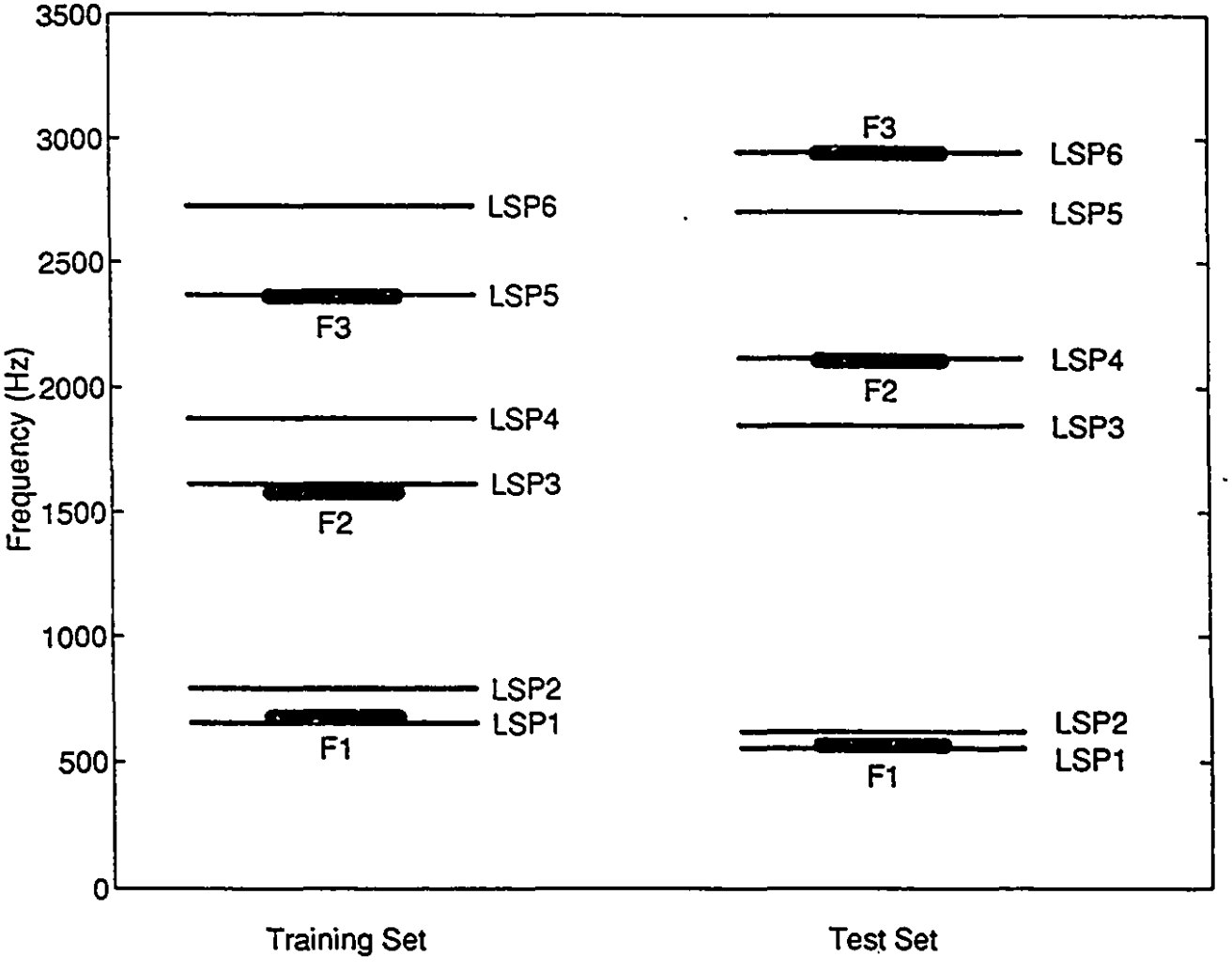

Figure 17: Relationship Between LSP and LPC Frequencies - Phoneme /ae/

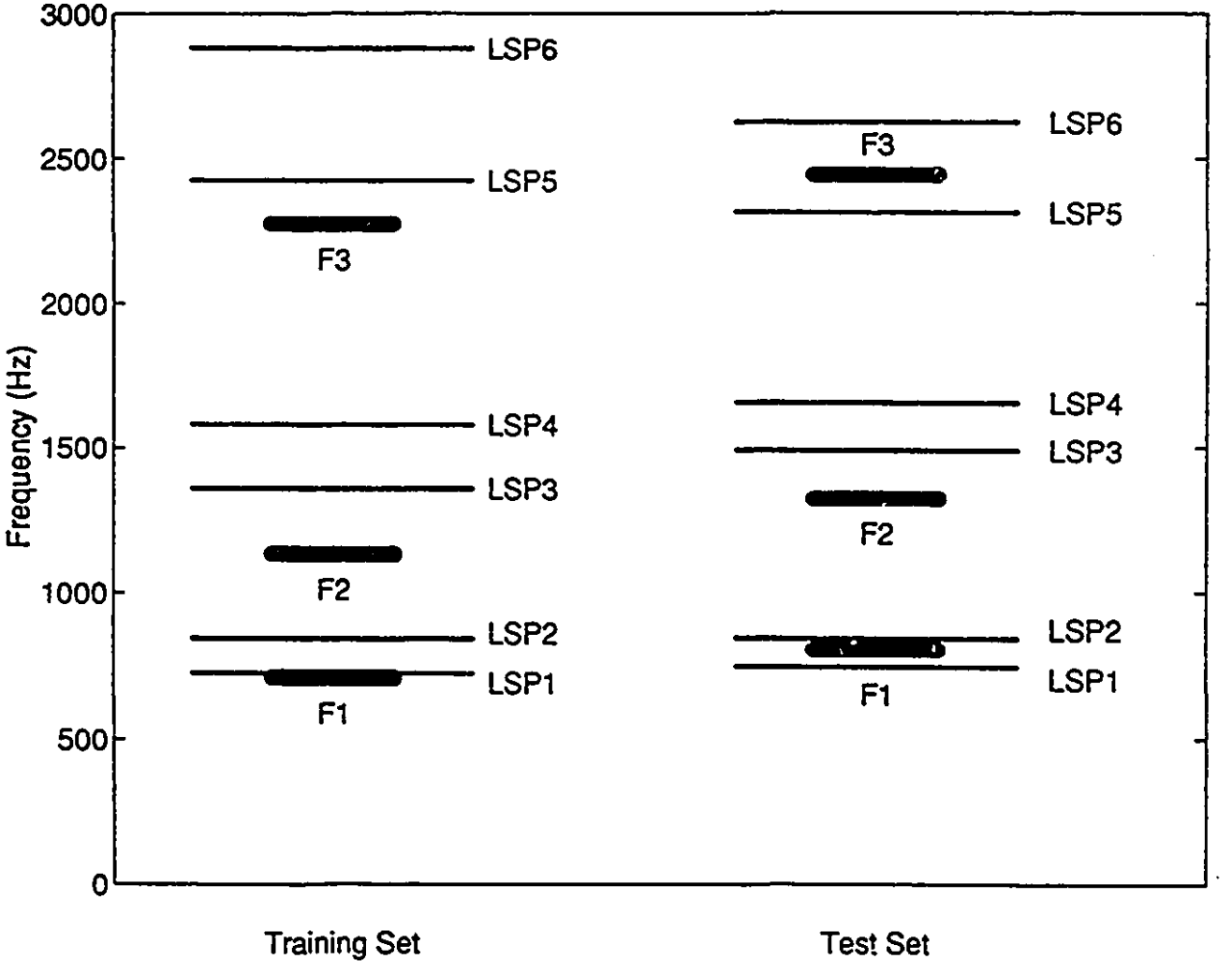

Figure 18: Relationship Between LSP and LPC Frequencies - Phoneme /aa/

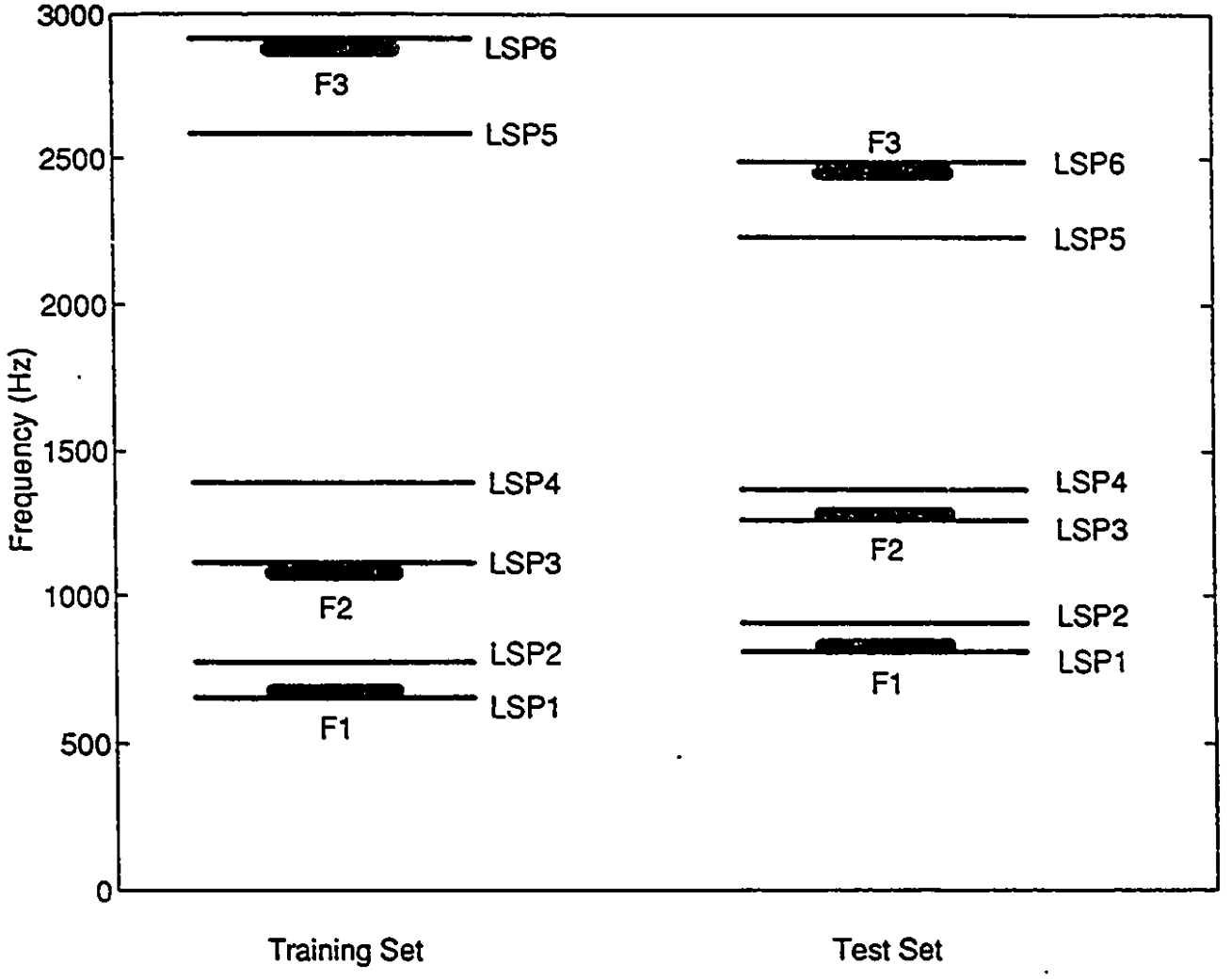

Figure 19: Relationship Between LSP and LPC Frequencies - Phoneme /ao/

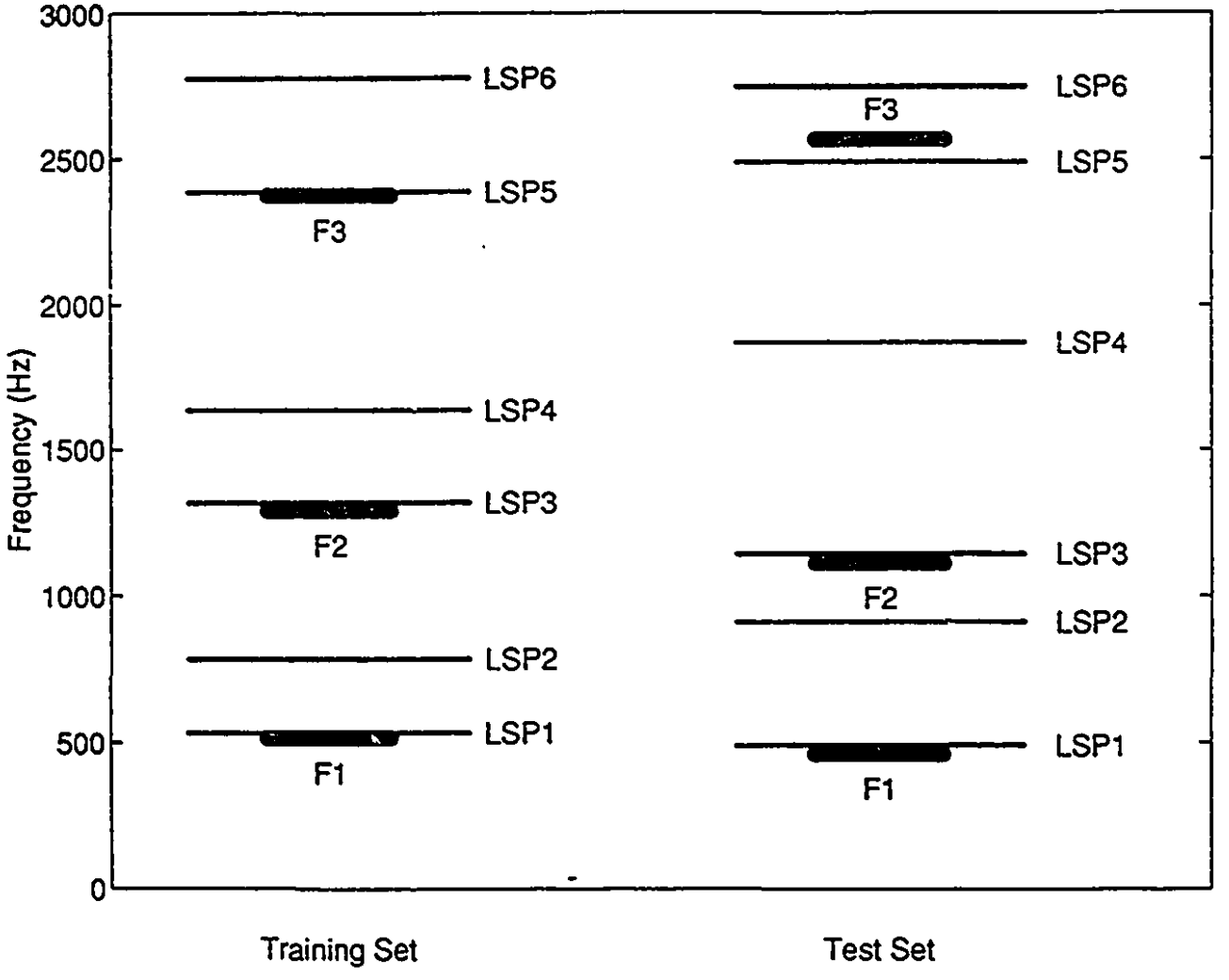

Figure 20: Relationship Between LSP and LPC Frequencies - Phoneme /ow/

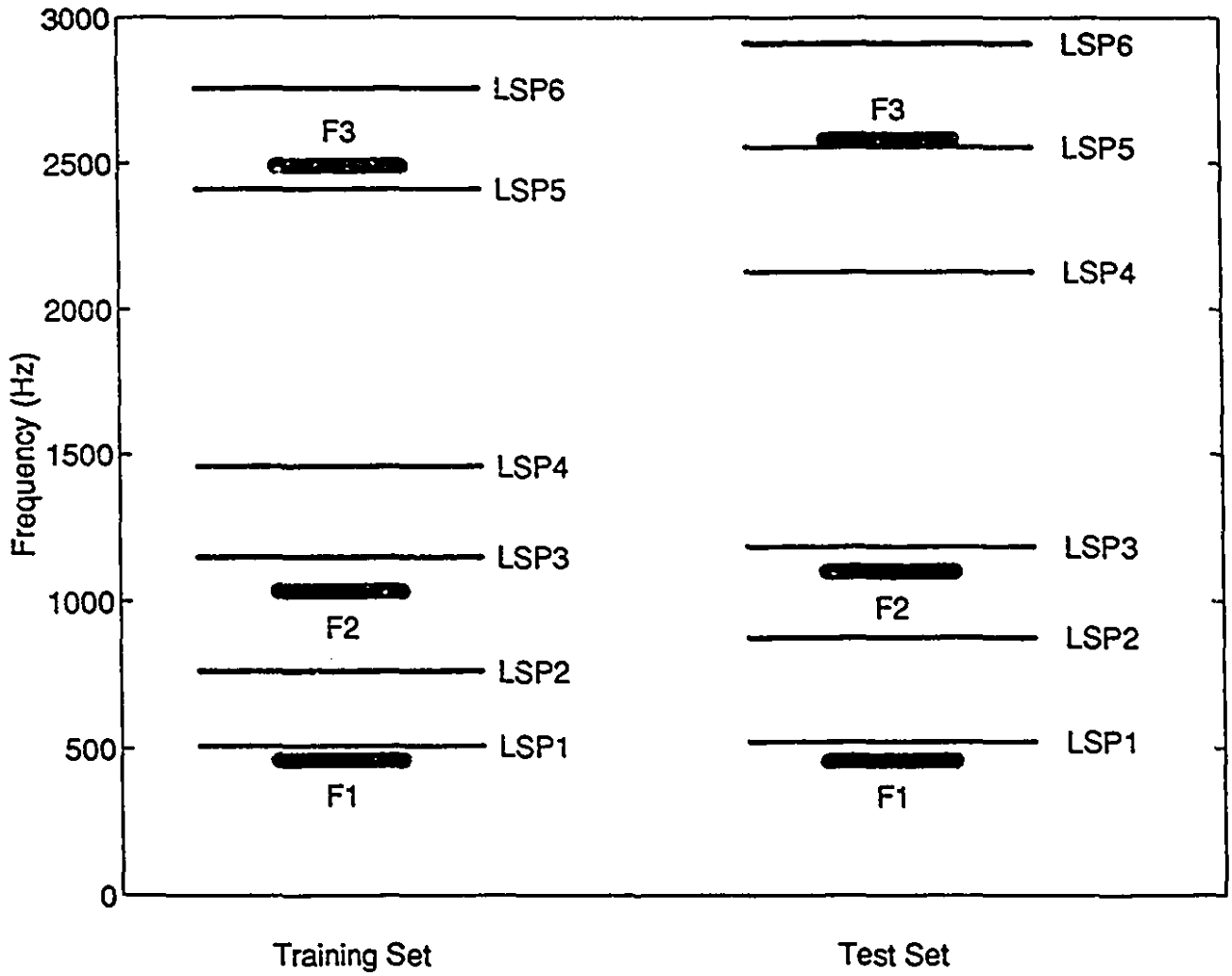

Figure 21: Relationship Between LSP and LPC Frequencies - Phoneme /uh/

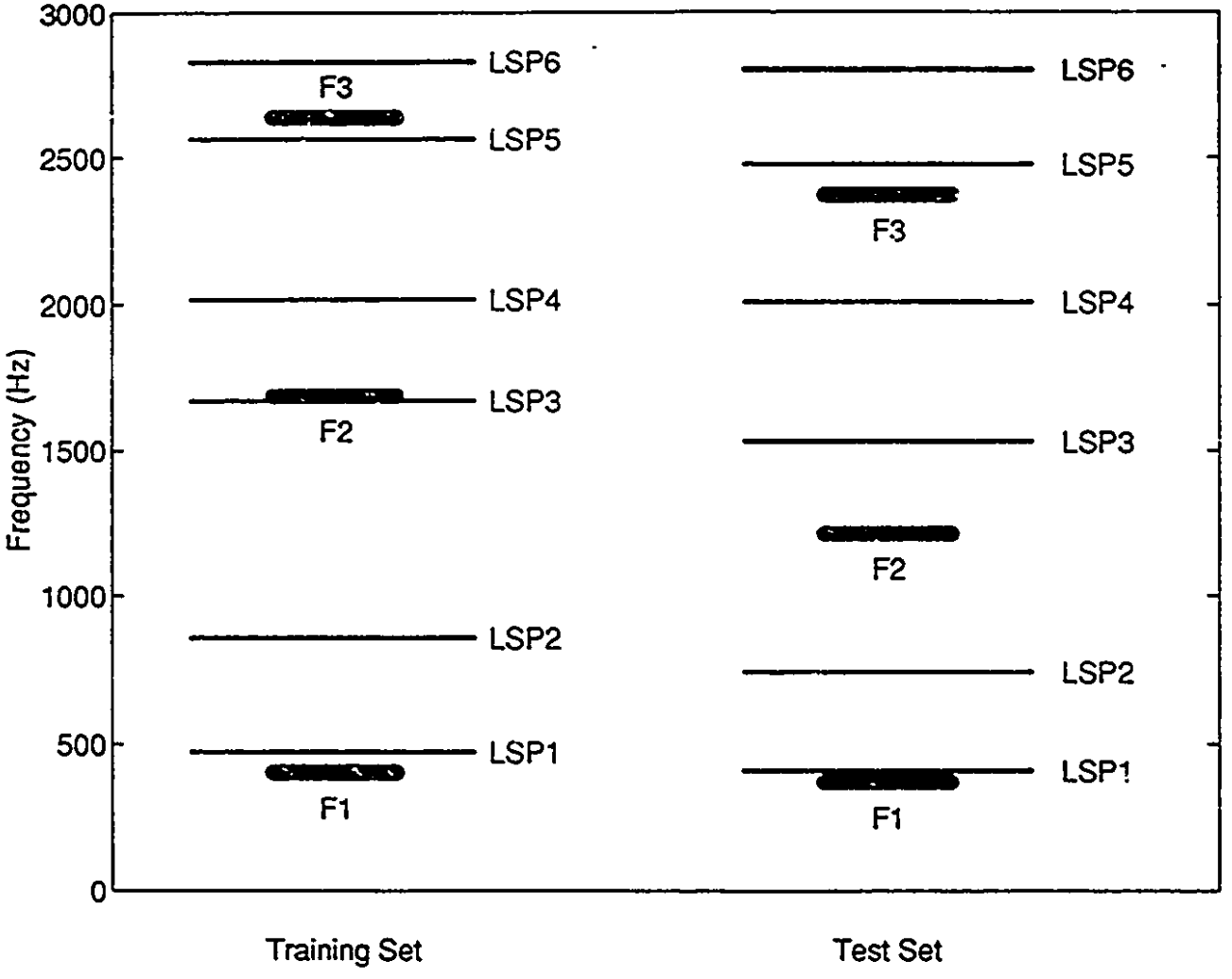

Figure 22: Relationship Between LSP and LPC Frequencies - Phoneme /uw/

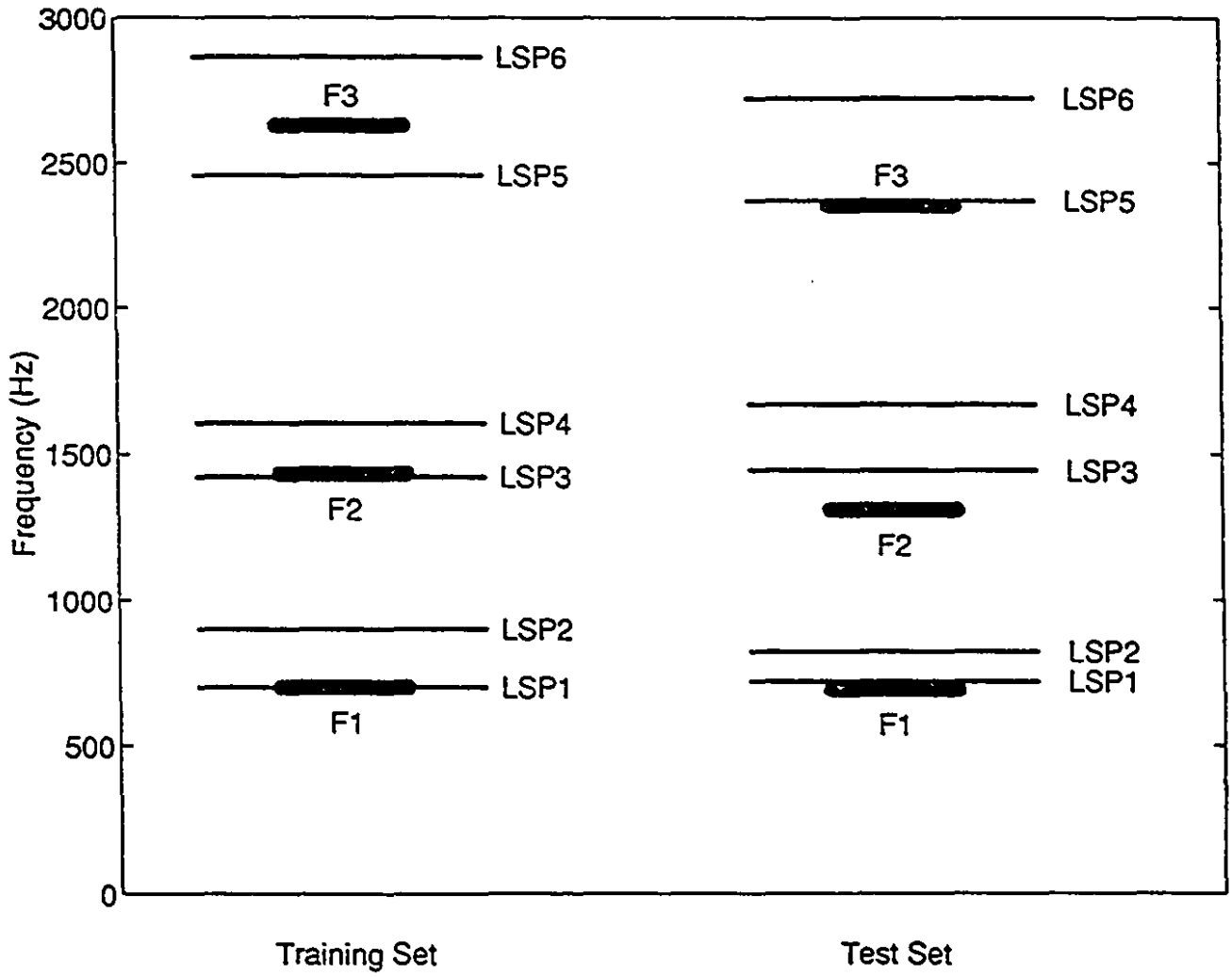

Figure 23: Relationship Between LSP and LPC Frequencies - Phoneme /ah/

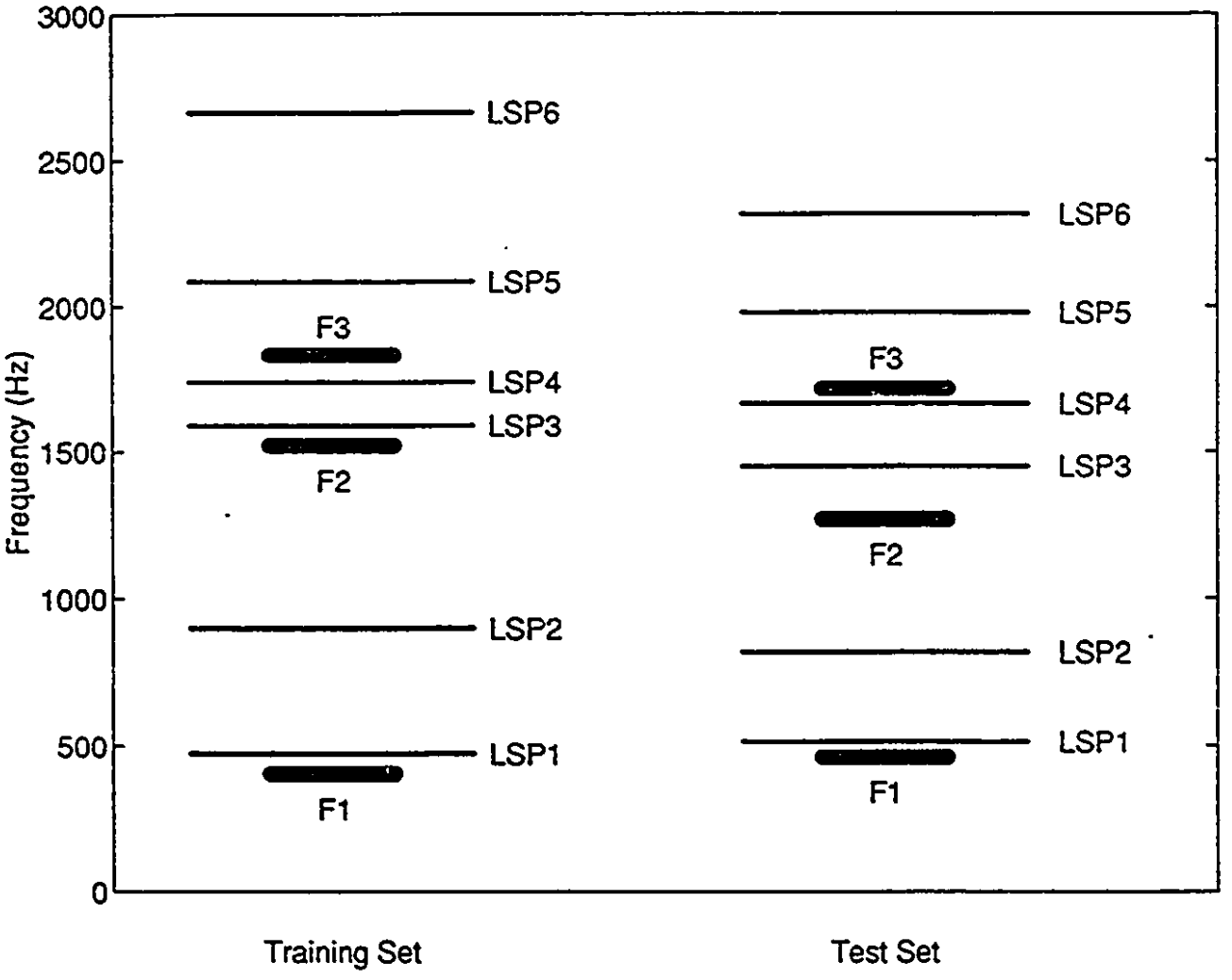

Figure 24: Relationship Between LSP and LPC Frequencies - Phoneme /er/

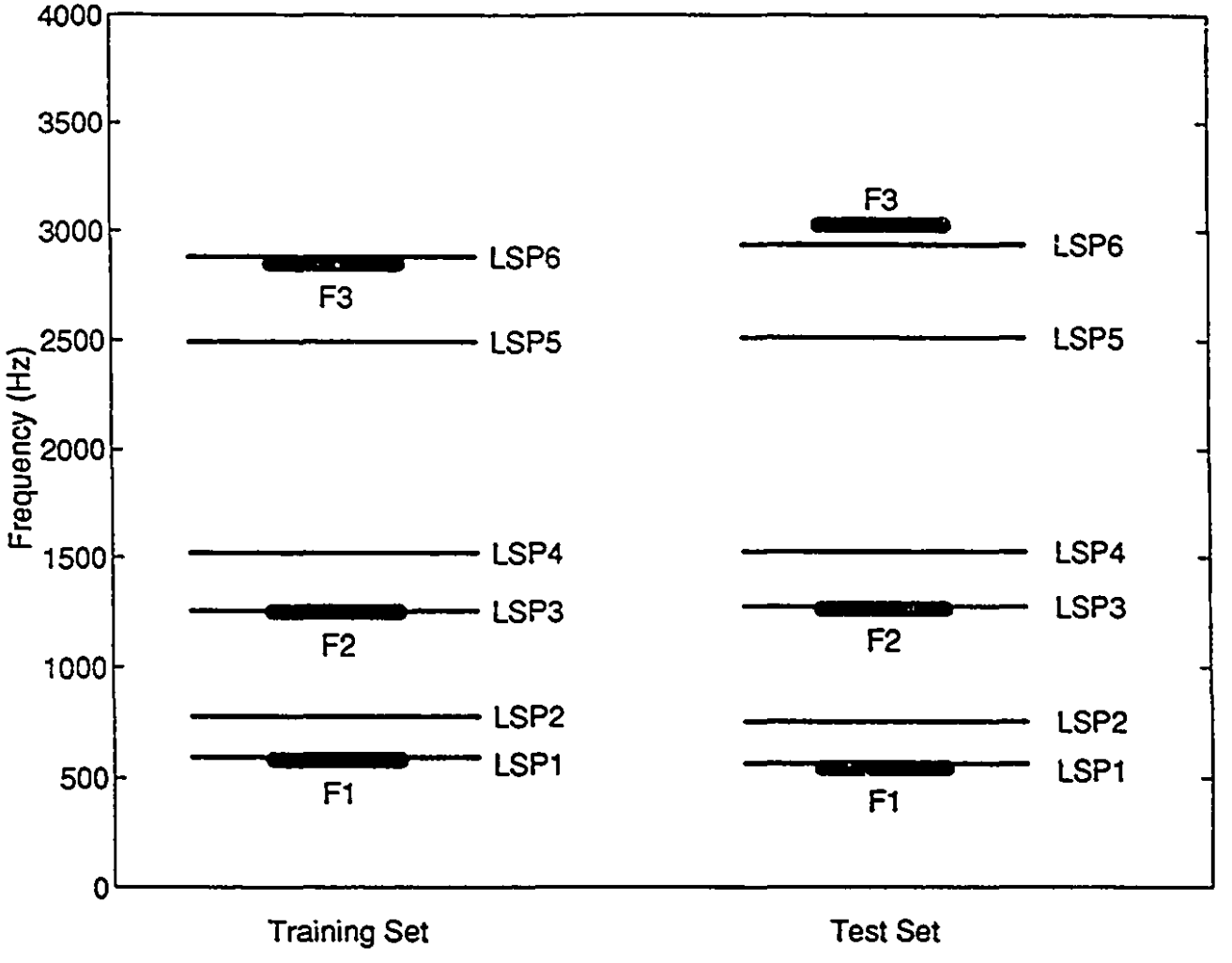

Figure 25: Relationship Between LSP and LPC Frequencies - Phoneme /ax/

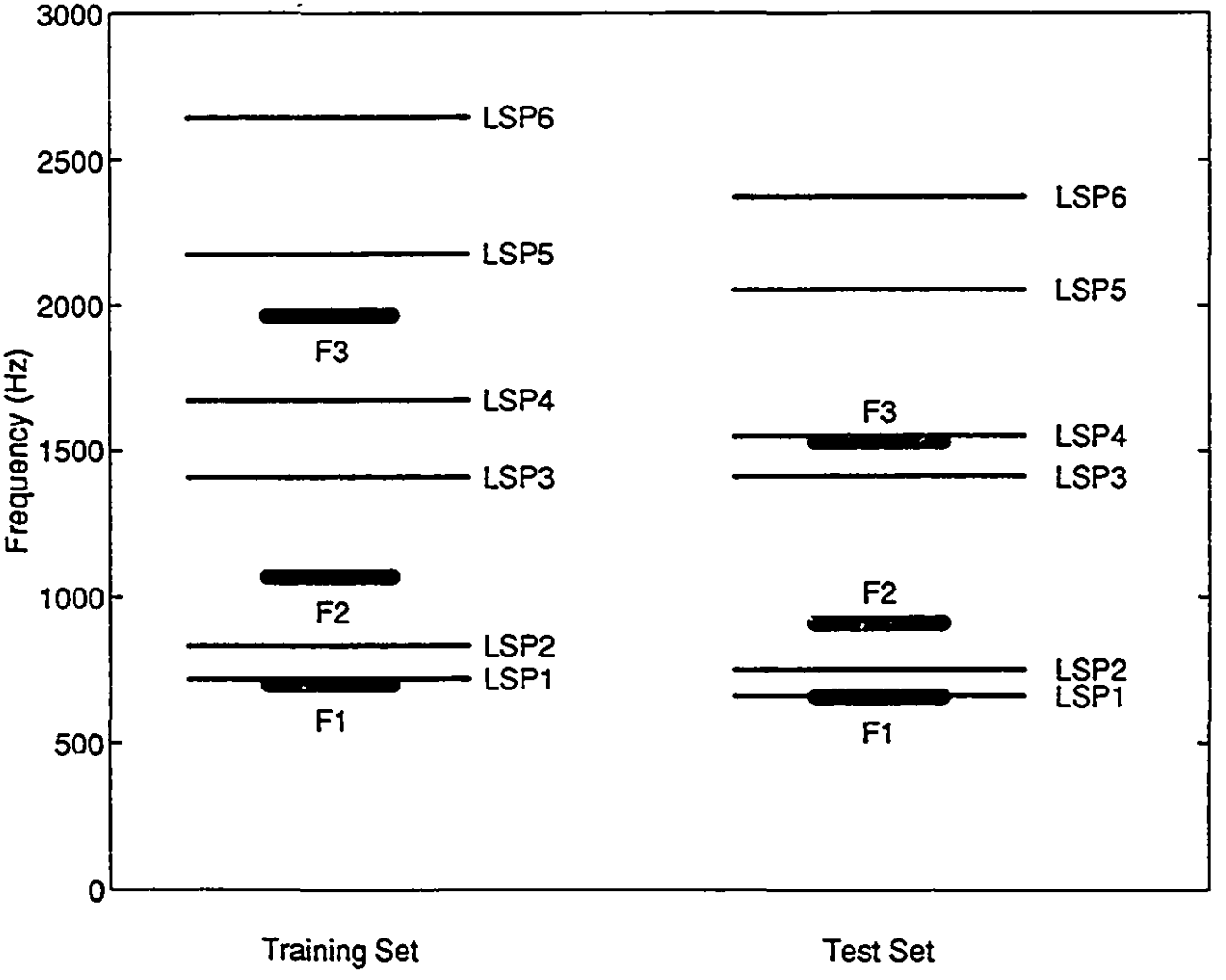

Figure 26: Relationship Between LSP and LPC Frequencies - Phoneme /ay/

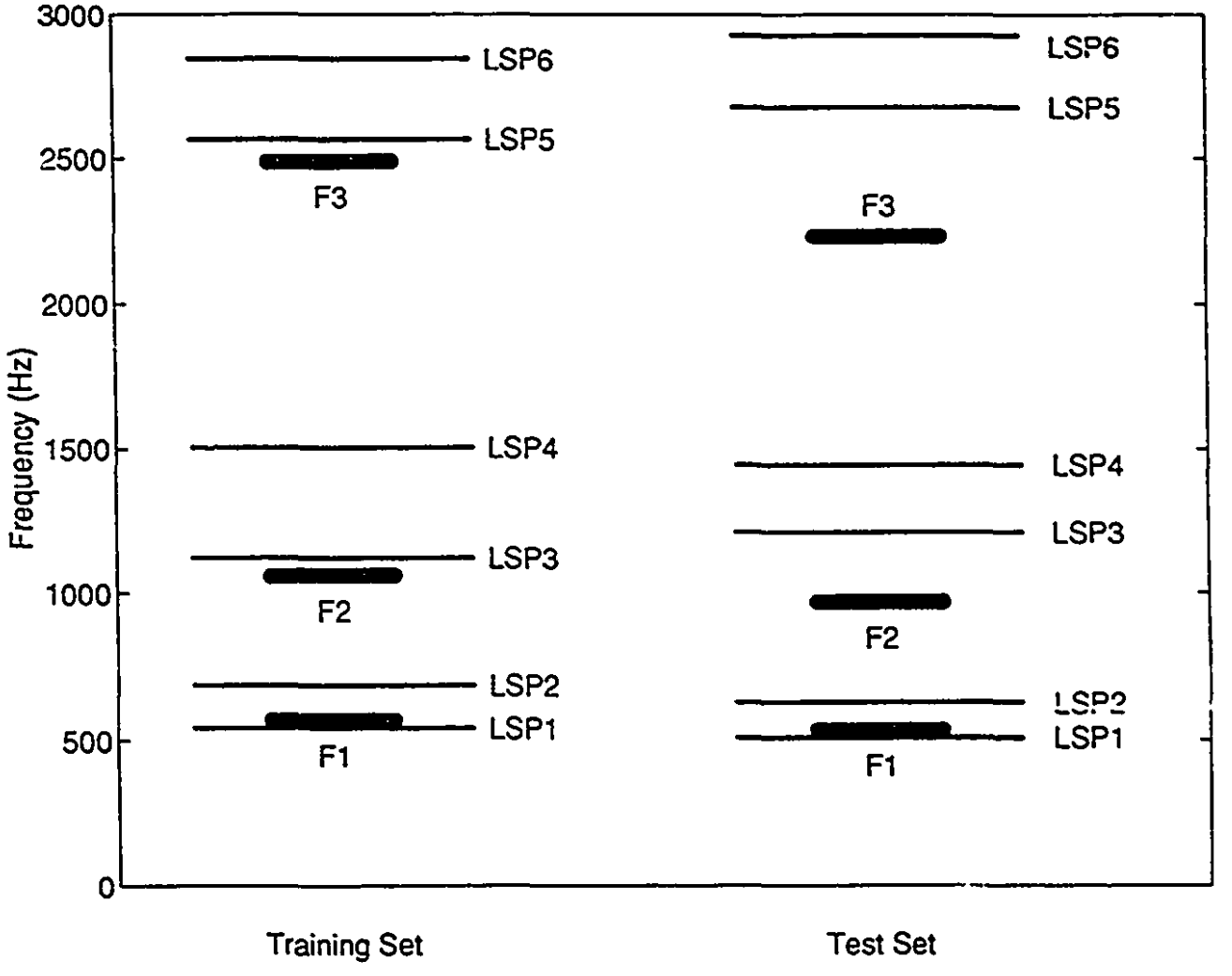

Figure 27: Relationship Between LSP and LPC Frequencies - Phoneme /oy/

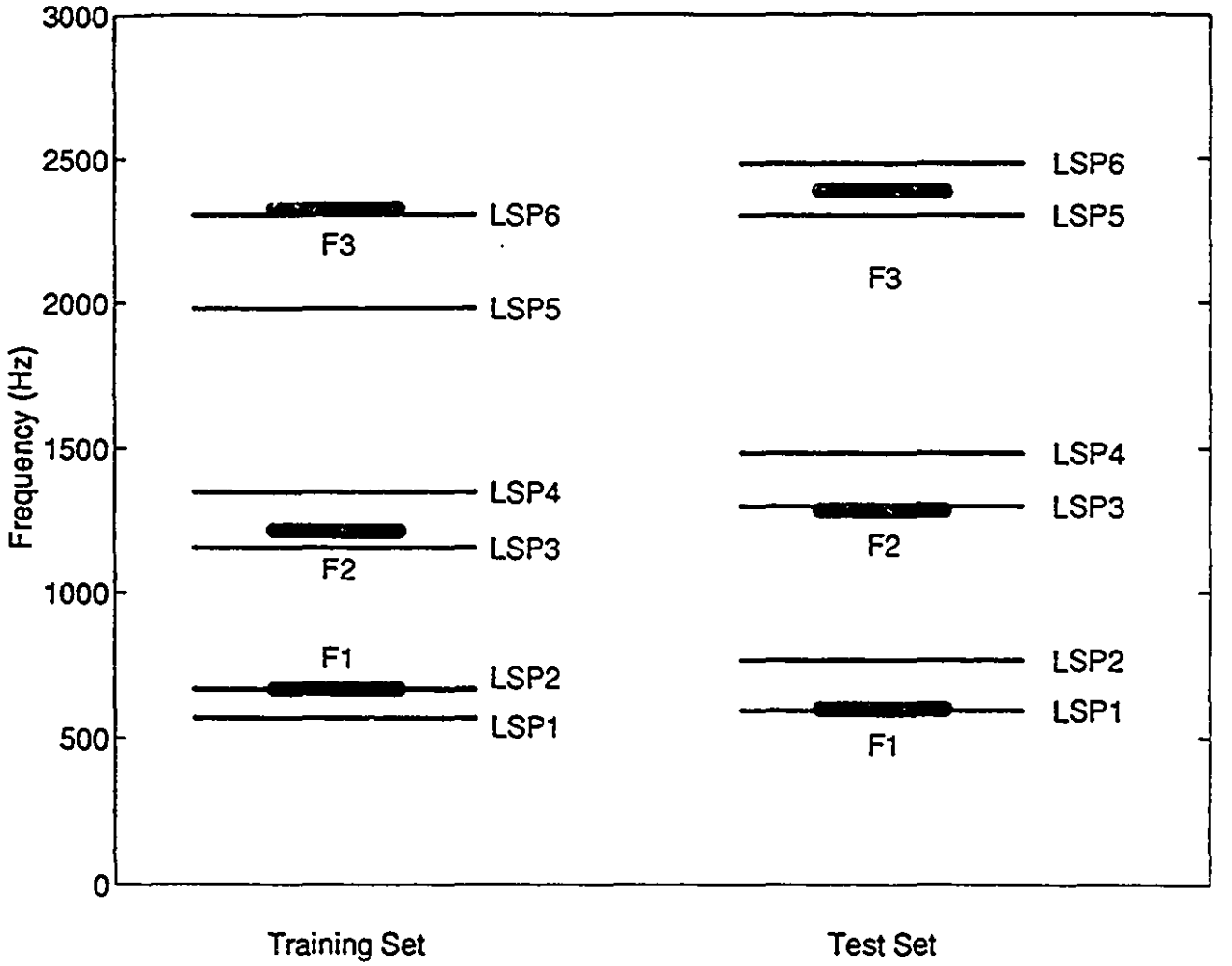

Figure 28: Relationship Between LSP and LPC Frequencies - Phoneme /aw/

5.6 Performance Summary and Results

The table below summarizes the performance of the test set of this experiment:

Table 8: Test Set Performance Summary

Phoneme	Expected Type	Performance
iy	3	pass
ih	7	pass
ey	3	pass
eh	5	pass
ae	5	fail
aa	1	pass
ao	2	pass (with explanation)
ow	9	fail
uh	9	pass
uw	9	pass
ah	8	pass
er	4	pass
ax	2	pass
ay	1	fail (partially)
oy	7	pass
aw	6	fail

In total, 12 of the 16 phonemes fit into the anticipated Types. This represents an experimental success of 75%, and an experimental failure of 25%. Based on the size of the data set used, the results are very good. In general, many of the cognitive observations presented in Section 2.7 are confirmed. Specifically, in most of the phonemes, the odd LSP frequencies are located near the low formant center frequency. Also, a clustering of line spectra occurs in the vicinity of the resonant frequencies,

whereas sparsely spaced spectra are located near the valleys of the spectral envelope. One of the most striking observations in this experiment is the tight correlation between F1 and LSP1 for all phonemes. Another correlation, while not as pronounced, is found between F2 and LSP3 for all phonemes. These factors also serve to prove the reliability of the data obtained.

The reliability of results was expected to be very good because much attention was paid to ensure that the precision and accuracy of the data was high. For example, much care was exercised during manual extraction of steady-state vowels. Also, a number of checks were employed at each stage of the experiment to ensure the validity of the results.

There are some ways in which the reliability may have been improved. For example, if the experiment was conducted using same equipment, operating systems and software. Because of the cutover between the VAX and the UNIX systems at INRS, the same programs could not be used. Also, the reliability of the results would have been greater if a larger, speaker dependent data set were used. This would have possibly decreased the variability between the LSP and LPC frequencies, and also the variability between the training and test set data. Lastly, while every phoneme was extracted manually for the steady-state vowel, the vowels should ideally be part of a vocabulary that provides constant effects on all vowels. For example, a /bVb/ set would have provided a uniform impact of the consonant on the vowel phonemes.

Future work in this area involves a study how LPC formant and LSP bandwidths affect the relationships cited in this research. A correlation between the formant frequency, its bandwidth and the LSP frequencies and their bandwidths could be added to provide another level of comparison. Also, another area for future work involves the analysis using transitional frequencies.

Chapter 6

Conclusion

This thesis commenced with a theoretical background on line spectrum pairs. The mathematics of how the LSP method naturally followed from the PARCOR method was described, as well as a number of special properties of LSP frequencies. The articulatory phonetics of speech were briefly discussed in order to show the physiological link between formants and vowel sounds. The process of data acquisition and preprocessing, as well as the process of analysis, was then described.

The experiment produced a 75% success rate in proving that vowel phonemes may be divided into groups based on similarities between LSP and LPC frequencies. The 16 phonemes were divided into nine types. This work is intended to add to the base of knowledge pertaining to the relationship between LSP and formant frequencies. The research has shown that there is a distinct relationship regarding their behavior, and provides an appreciation of this relationship. The results of this work are intended to increase the effectiveness of speech recognition systems by providing an additional measure as a guide in feature extraction.

Bibliography

[1] Boite, R. and M. Kunt, *Traitement de la Parole*, Switzerland: Presses Polytechniques Romandes, 1987.

[2] Chan, C.F. and K.W. Law, "An Algorithm for Computing LSP Frequencies Directly from the Reflection Coefficients", *EUROSPEECH 91, 2nd European Conference on Speech Communication and Technology Proceedings*, vol. 2, 1991, pp. 913-916.

[3] Ching, P.C. and K.C. Ho, "An Efficient Adaptive LSP Method for Speech Encoding", *IEEE Region 10 Conference on Computer and Communication Systems*, September 1990, pp. 324-328.

[4] Davis S.B and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 4, August 1980, pp. 357-366.

[5] Delsarte, P. and Y. Genin, "The Split-Levinson Algorithm", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-35, no. 5, May 1987, pp. 645-653.

[6] Everett, S.S., "Word Synthesis Based on Line Spectrum Pairs", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 675-678.

[7] Farvardin, N. and R. Laroia, "Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transform", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 1, May 1989, pp. 168-171.

[8] Furui, S., "Recent Advances in Speech Recognition", *EUROSPEECH 91, 2nd European Conference on Speech Communication and Technology Proceedings*, vol. 1, 1991, pp. 3-10.

[9] Gray A.H. and J.D. Markel, "Distance Measures for Speech Processing", *IEEE Transactions on Acoustics, Speech and Signal Processing* , October 1976, vol. ASSP-24, no. 5, pp. 380-391.

[10] Gretter, R. and M. Omologo, "The Use of Line Spectrum Representation in Speech Synthesis", *Alta Frequenza*, LVIII(3), May-June 1989, pp. 293-300.

[11] Gurgen, F.S., S. Sagyama and S. Furui, "A Study of Line Spectrum Pair Frequency Representation for Speech Recognition", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E75-A(1), January 1992, pp. 98-102.

[12] Itakura, F. and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", *Electronics and Communications in Japan*, vol. 53-A(1), 1970, pp. 36-43.

[13] Kabal, P. and R.P. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshev Polynomials," *Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 6, pp 1419-1426, December, 1986

[14] Kang, G.S. and L.J. Fransen, "Experimentation with Synthesized Speech Generated from Line-Spectrum Pairs", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35(4), April 1987, pp. 568-571.

[15] Kang, G.S. and L.J. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encoders", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1985, pp. 244-247.

[16] Kim, H.K., K.C. Kim and H.S. Lee, "Enhanced Distance Measure for LSP-Based Speech Recognition", *Electronics Letters*, 1993, vol. 29(16), pp. 14631465.

[17] Ladefoged, P., *A Course in Phonetics*, 2nd Edition, N.Y.: Harcourt Brace Jovanovich, Inc., 1982.

[18] Laroia, R., N. Phamdo and N. Farvardin, "Robust and Efficient Quantization of Speech LSP Parameters Using Structured Vector Quantizers", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1991, vol. 1, pp. 641-644.

[19] Lee, E.A. and D.G. Messerschmitt, *Digital Communication*, Boston: Kluwer Academic Publishers, 1988.

[20] Lepschy, A., G.A. Mian and U. Viaro, "A Note on Line Spectral Frequencies", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36(8), August 1988, pp. 1355-1357.

[21] Liu, C.-S., C.-S. Huang, M.-T. Lin and H.-C. Wang, "Automatic Speaker Recognition Based Upon Various Distances of LSP Frequencies", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. 104-109.

[22] Liu, C.-S., M.-T. Lin, W.-J. Wang and H.-C. Wang, "A Study of Line Spectrum Pair Frequencies for Speaker Recognition", *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1990, vol. 1, pp. 277-280.

[23] Markel, J.D. and A.H. Gray, Jr., *Linear Prediction of Speech*, Berlin: Springer-Verlag, 1976.

[24] O'Shaughnessy, D., *Speech Communication: Human and Machine*, N.Y.: Addison-Wesley Publishing Company, 1987.

[25] Omologo, M., "The Computation and Some Spectral Considerations on Line Spectrum Pairs (LSP)", *EUROSPEECH 89*, vol. 2, 1989, pp. 352-355.

[26] Paliwal, K.K., "A Study of Line Spectrum Pair Frequencies for Vowel Recognition", *Speech Communication*, vol. 8, 1989, pp. 27-33.

[27] Paliwal, K.K., "On the Use of Line Spectral Frequency Parameters for Speech Recognition", *Digital Signal Processing*, vol. 2, 1992, pp. 80-87.

[28] Quackenbush, S.R., T.R. Barnwell III and M.A. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, N.J.: Prentice Hall, 1988.

[29] Saito, S. and K. Nakata, *Fundamentals of Speech Signal Processing*, Orlando, Florida: Academic Press, 1985, p. 127.

[30] Saoudi, S., J.M. Boucher and A. Le Guyader, "Medium Band Speech Coding Using Optimal Scalar Quantization of LSP", *EUROSPEECH 91, 2nd European Conference on Speech Communication and Technology Proceedings*, vol. 2, 1991, pp. 905-908.

[31] Saoudi, S., J.M. Boucher, A. Le Guyader, "A New Efficient Algorithm to Compute the LSP Parameters for Speech Coding", *Signal Processing*, vol. 28(2), August 1992, pp. 201-212.

[32] Secker, P. and A. Perkis, "Joint Source and Channel Coding of Line Spectrum Pairs", *EUROSPEECH 91, 2nd European Conference on Speech Communication and Technology Proceedings*, vol. 2, 1991, pp. 909-912.

[33] Snell, R.C. and F. Milinazzo, "Formant Location from LPC Analysis Data", *EUROSPEECH 91, IEEE Transactions on Speech and Audio Processing*, vol. 1(2), 1993, pp. 129-134.

[34] Soong, F.K. and B.-H. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1984, pp. 1.10.1-1.10.4.

[35] Soong, F.K. and B.-H. Juang, "Optimal Quantization of LSP Parameters", *IEEE Transactions on Speech and Audio Processing*, vol. 1(1), January 1993, pp. 15-23.

[36] Sugamura N. and F. Itakura, "Speech Data Compression by LSP Speech Analysis and Synthesis Technique", *IECE Transactions*, vol. J64-A, no. 8, 1981, pp. 599-605.

[37] Sugamura, N. and F. Itakura, "Speech Analysis and Synthesis Methods Developed at ECL in NTT - From LPC to LSP", *Speech Communications*, vol. 5, June 1986, pp. 199-215.

[38] Sugamura, N. and N. Farvardin, "Quantizer Design in LSP Speech Analysis and Synthesis", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 398-401.

[39] Tubach, J.P. et al., *La Parole et Son Traitement Automatique*, Paris: Masson and CNET-ENST, 1989.

[40] Umezaki, T. and F. Itakura, "Analysis of Time Fluctuating Characteristics of Linear Predictive Coefficients", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1986, pp. 1257-1260.

[41] Wakita, H., "Linear Prediction Voice Synthesizers: Line-Spectrum Pairs (LSP) is the Newest of Several Techniques", *Speech Technology*, vol. 1, Fall 1981, pp. 17-22.