SPEECH SPECTRUM NON-STATIONARITY DETECTION BASED ON LINE SPECTRUM FREQUENCIES AND RELATED APPLICATIONS

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND
ELECTRONICS ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

BY

ALİ ERDEM ERTAN

OCTOBER 1998

Thesis
TK
7882
.S65
E78
1998

SPEECH SPECTRUM NON-STATIONARITY DETECTION BASED ON LINE SPECTRUM FREQUENCIES AND RELATED APPLICATIONS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND

ELECTRONICS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Ali Erdem ERTAN

October 1998

TK
7882
.S65
E78
1998

043706

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Enis Çetin

A. Enis Çetin, Ph. D (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

M. Demirekler

Mübeccel Demirekler, Ph. D

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

O. Arıkan

Orhan Arıkan, Ph. D

Approved for the Institute of Engineering and Sciences:

Prof. Dr. Mehmet Baray Y.
Director of Institute of Engineering and Sciences

ABSTRACT

SPEECH SPECTRUM NON-STATIONARITY DETECTION BASED ON LINE SPECTRUM FREQUENCIES AND RELATED APPLICATIONS

Ali Erdem ERTAN
M.S. in Electrical and Electronics Engineering
Supervisor: A. Enis Çetin, Ph. D
October 1998

In this thesis, two new speech variation measures for speech spectrum non-stationarity detection are proposed. These measures are based on the Line Spectrum Frequencies (LSF) and the spectral values at the LSF locations. They are formulated to be subjectively meaningful, mathematically tractable, and also have low computational complexity property. In order to demonstrate the usefulness of the non-stationarity detector, two applications are presented: The first application is an implicit speech segmentation system which detects non-stationary regions in speech signal and obtains the boundaries of the speech segments. The other application is a Variable Bit-Rate Mixed Excitation Linear Predictive (VBR-MELP) vocoder utilizing a novel voice activity detector to detect silent regions in the speech. This voice activity detector is designed to be robust to non-stationary background noise and provides efficient coding of silent sections and unvoiced utterances to decrease the bit-rate. Simulation results are also presented.

Keywords: Speech variation measure, spectrum non-stationarity detection, formant estimation, Line Spectrum Frequencies (LSF), speech segmentation, Mixed Excitation Linear Predictive coding (MELP), variable bit-rate vocoder, voice activity detector.

ÖZET

ÇİZGİ İZGE SIKLIKLARININ TEMEL ALINMASI İLE KONUŞMA İZGESİNDEKİ DURAĞANSIZLIĞIN SEZİMİ VE İLGİLİ UYGULAMALAR

Ali Erdem ERTAN
Elektrik ve Elektronik Mühendisliği Bölümü Yüksek Lisans
Tez Yöneticisi: Prof. Dr. A. Enis Çetin
Ekim 1998

Bu tezde, konuşma izgesindeki durağansızlıkların sezimi için iki yeni konuşma değişgenlik ölçüsü önerilmiştir. Bu ölçüler yaratılırken Çizgi İzge Sıklıkları (ÇİS) ve ÇİS konumlarındaki izgesel değerler taban alınmıştır. Önerilen ölçüler öznel olarak anlamlı ve düşük hesaplama karmaşıklığı olacak ve matematiksel olarak izlenebilecek şekilde formüle edilmişlerdir. Durağansızlık sezimleyicisinin yararlılığını göstermek için iki uygulama sunulmuştur: Birinci uygulama, konuşma sinyalindeki durağansız bölgeleri bulan ve bu bölgelerde bulunan konuşma parçalarının sınırlarını sezimleyen bir kesin konuşma bölütleyicisidir. Öteki uygulama ise konuşmadaki sessiz bölgeleri sezimleyen yeni bir konuşma faaliyet sezimcisini kullanan Değişken İkil Hızlı-Karışık Tahrikli Doğrusal Öngörülü (DİH-KTDÖ) kodlama ses kodlayıcısıdır. Bu ses faaliyet kestirimcisi, durağansız arka plan gürültüsüne dayanıklı olacak şekilde tasarlanmıştır ve ikil-hızın düşürülmesi için sessiz bölgelerin ve sessiz harflerin verimli kodlanmasına olanak sağlamaktadır. Test sonuçları da tezde sunulmuştur.

Anahtar Kelimeler: Konuşma değişgenlik ölçüsü, izgideki durağansızlıkların sezimi, formant kestirimi, Çizgi İzge Sıklıkları (ÇİS), konuşma bölütleme, Karışık Tahrikli Doğrusal Öngörülü (KTDÖ) kodlama, degişken ikil-hızlı ses kodlayıcısı, ses faaliyet sezimi.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to Prof. Dr. A. Enis Çetin for his supervision, guidance, suggestions and patience throughout this study, and also for encouraging me to attend the international conferences, which provide me motivation and experience.

I would also express my special thanks to Prof. Dr. Mübeccel Demirekler for her enormous help and guidance for the last two years, from whom I really enjoy learning the basics and the theory behind speech processing.

I would like to thank Dr. Orhan Arıkan for reading and commenting on the thesis.

I would like to acknowledge to TÜBİTAK-BİLTEN who supported our work.

I would also like to thank to all of my friends in TÜBİTAK-BİLTEN and Bilkent University who have been with me during my M.Sc. study and MELP project, especially Emre, Levent, Önder, Haydar, Murat and Taner for their great friendship and sharing long working nights in TÜBİTAK, and Dr. H. Gökhan İlk for the long discussions on my thesis.

My special thanks go to Cem Baydar and Yılmaz Acar for their continuous ”*Eye of the Tiger*” style morale support. I also want thank to S. Muzaffer for providing our security throughout the development of this thesis!

Finally, it is a pleasure for me to express my sincere thanks to my family for their continuous morale support throughout my graduate study, and my sweetheart, Didem Öztürk, for her endless morale support and patience.

Contents

List of Figures

List of Tables

To My Family and My Beloved...

Chapter 1

Introduction

Communication is defined as *the imparting or interchange of thoughts, opinions, or information by speech, writing, or signs* in Webster Dictionary. Every living organism which can move likes to communicate with its own race and its living environment. Even bacteria, a one celled organism, communicate with each other to exchange DNAs to gain more immunity for the changing environment conditions. As the most complex and developed organism on the earth, human race also enjoys to communicate each other to share information and tell their emotions.

For centuries, communication methods of humans are getting complicated starting from body language and simple sounds to highly structured spoken and written languages, consist of syntactic, semantic and linguistic rules. Among these methods, speech is the most used one in the daily life of a human. Humans use their articulatory and auditory systems to generate and perceive speech. The rules for this communication method is described by language. Speaker produces different sounds and concatenates them to generate meaningful words. On the other side, the listener receives generated sound by her/his auditory system and this incoming signal is processed within brain to extract the meaning of this signal.

The advances in the digital signal processing area make speech also a serious communication media between human and machine. As a result, speech becomes a central component in digital communication. For several decades, considerable research focused on several areas of speech processing. These areas can be summarized as compression, recognition, enhancement and synthesis [1].

Speech segmentation is an important first step in coding, recognition and synthesis. The primary purpose is to segment continuous speech signal into phonetic units [2]. The simplest way is to divide the speech for fixed, non-overlapping time intervals. This type of segmentation is mostly used in variable-rate speech compression algorithms [3]. In continuous speech recognition, end-points of the utterances within the continuous speech is extracted first [4]. After this step, various segmentation algorithms may be applied to the signal to obtain the boundaries of the segments prior to the recognition algorithms.

In this thesis we propose new speech spectrum variation measures, which are similar to speech distortion measures. They are used to detect the amount of change in the speech spectrum according to the variation of selected parameter set. Since our purpose is to detect the speech spectrum variations among analyzed frames rather than quantization effects, we use '*speech variation measure*' instead of '*distortion measure*'.

Segmentation algorithms can be classified into two major groups: Implicit segmentation, in which no prior information is used about signal [5–8], and explicit information, in which phonetic transcription is also available [9–13]. Success of the implicit segmentation algorithms depends primarily on the selection of the parameter set and the speech variation measure. In literature, following parameters are reported to be used:

- Modeling error of linear prediction filter [6],
- Linear Prediction Coefficients (LPC)-smoothed log amplitude spectra [2],
- parametric filtering [5],
- energies in subbands [8],
- auditory model [12], and

- Line Spectrum Frequencies (LSFs) [14].

These parameters are used in various speech variation measures, whose values are generally compared with a threshold to detect the non-stationarity between consecutive frames or to obtain exact change point. In [15], it is stated that in order to be useful a distortion measure, following conditions must be satisfied:

1. It must be subjectively meaningful in the sense that small and large distortion corresponds to good and bad subjective quality, respectively.
2. It must be tractable in the sense that it is amenable to mathematical analysis and leads to practical design techniques.
3. It must be computable in the sense that the actual distortions resulting in a real system can be efficiently computed.

The most commonly used distortion measure is the traditional mean squared error. However, in speech processing, this distortion measure does not provide any subjective meaning. Therefore, in speech processing algorithms, usually Itakura-Saito distortion measure (1.1), Itakura's likelihood ratio (1.2) or L2 distance of log spectrum are used (1.3). Excellent review of distortion measures for speech processing can be found in [15].

- Itakura-Saito distortion measure:

  $$D_{IS} = \int_{-\pi}^{\pi} \left[\frac{S_1(w)}{S_2(w)} - \log \frac{S_1(w)}{S_2(w)} - 1 \right] dw \qquad (1.1)$$

  where $S_1(w)$ and $S_2(w)$ are the estimated spectral density functions of the two speech frames.

- Itakura likelihood ratio:

  $$D_I = \frac{a_1^T R_1 a_1}{a_2^T R_1 a_2} \qquad (1.2)$$

  where a_1, R_1 and a_2 are LPC coefficient vector of the reference frame, autocorrelation matrix of the reference frame and LPC coefficients vector of the comparison frame, respectively.

- L2 distance of log spectrum or spectral distortion:

$$D_{LL} = \int_{-\pi}^{\pi} \mid \log S_1(w) - \log S_2(w) \mid^2 dw \tag{1.3}$$

where $S_1(w)$ and $S_2(w)$ are the estimated spectral density functions of the two speech frames.

Besides these well-known measures, following measures are reported to work successfully in literature:

- Brantd's generalized likelihood ratio test [6],
- divergence test [6],
- the pulse method: A modified divergence test with a priori unvoiced-voiced detection [6],
- the normalized correlation between selected parameters of two frames [2],
- time-correlation based speech variation measures [5], and
- weighted Euclidean distance measure [14].

In Erkelens work [16], it is proved that spectral distortion can be approximated by weighted square distance of the coefficients of LPC filter and derived parameters, if cubic and higher terms of Taylor series expansion is neglected. The weighting matrix is equal to the inverse of the theoretical covariance matrix of the coefficients. Furthermore, LSFs are found to be uncorrelated, and only main diagonal entries of the weighting matrix are non-zero. Therefore, with the usage of LSFs, this equation is also reduced to a weighted Euclidean distance measure. Due to this fact, it is possible to derive LSF based speech variation measures which not only provide a meaningful comparison of the spectrum of two speech frames in a low computational complexity way, but subjectively meaningful as well.

The main contribution of this thesis is the new speech variation measures based on the estimated peaks of the spectrum from LSF locations, angular difference between consecutive LSFs, and usage of these measures in the detection of spectrum non-stationarity between consecutive frames. Both of these measures show high correlation between speech spectrum and LSF displacement. These measures are formulated to be subjectively meaningful, mathematically tractable and have low computational complexity. In order to demonstrate the usefulness of the non-stationarity detector, two applications using this detector are presented: A speech segmentation system and a variable rate speech vocoder.

In Chapter 2, a novel spectrum non-stationarity detection algorithm based on the new speech variation measures is presented. In this algorithm, peaks of the spectrum are estimated from LSFs by a two-step algorithm, and these estimated peaks and the angular difference between consecutive LSFs are used to form new perceptually meaningful speech variation measures. Non-stationarity detection of speech spectrum is obtained by comparing the computed values of these measures with thresholds, which may be adjusted for different algorithms.

Chapter 3 and Chapter 4 presents applications using this non-stationarity detector: A frame based speech segmentation algorithm is presented in Chapter 3. This algorithm is of implicit type and only finds the non-stationary regions in the speech signal.

In Chapter 4, a Variable Bit-Rate Mixed Excitation Linear Predictive Coding (VBR-MELP) system is described. This new coder is based on federal standard fixed-rate MELP coder [17]. In order to reduce bit-rate, unvoiced frames are encoded with fewer bits, sufficient to synthesize these sections and parameters for the frames including only silence or background noise are not transmitted. Silent parts and background noise sections are detected by a novel voice activity detector (VAD) which has non-stationary noise immunity.

Conclusions and discussions are given in Chapter 5.

In following sections, the theoretical background information including vocal tract modeling, linear prediction, modeling of human speech production system are presented. Also, detailed information about Line Spectrum Frequencies (LSFs) is given.

1.1 Linear Modeling of Vocal Tract

In this section, description of linear modeling of the vocal tract is presented and parameters, required for this model, is defined. The model should be mathematically tractable, while imitating the actual system as much as possible. This task can be achieved via following assumptions:

1. The effects of nasal tract can be ignored.
2. The vocal tract can be assumed to consist of N interconnected sections where each individual section is of uniform cross-sectional area.
3. The transverse dimension of each section is small enough compared with a wavelength, so that the sound propagation through an individual section can be treated as a plane wave.
4. The internal losses due to wall vibration, viscosity and heat conduction are negligible.
5. A linear, time varying acoustic tube model of the vocal tract, uncoupled from the glottis can be constructed.

A typical example for this modeling can be seen in Figure 1.1.

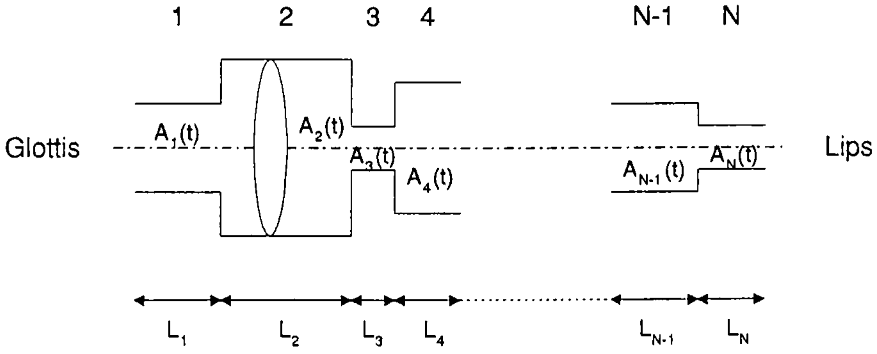


Figure 1.1: Simplified interconnected acoutic tube modeling.

In digital modeling of speech, length of every tube is assumed to be equal. Also τ is defined as the required time for traveling of a wave from one junction to another. If the system is excited with an impulse, it propagates down the tubes being partially reflected and partially propagated at the junctions. The soonest that the impulse can reach output is $N\tau$ seconds. The successive impulses due to the reflections can reach to the output at multiples of 2τ seconds later. As a result, impulse response of the system can be written as:

$$h(t) = \alpha_0 \delta(t - N\tau) + \sum_{k=1}^{\infty} \alpha_k \delta(t - N\tau - 2k\tau) \tag{1.4}$$

and the system function is:

$$H(s) = \sum_{k=0}^{\infty} \alpha_k e^{-s(N+2k)\tau} \tag{1.5}$$

The term $e^{-sN\tau}$ is pure time delay. Furthermore, the resonances of the system in Figure 1.2 is defined as follows:

$$\hat{H}(s) = \sum_{k=0}^{\infty} \alpha_k e^{-s2k\tau} \qquad \text{where } \hat{h}(t) = h(t + N\tau) \tag{1.6}$$

and

$$\hat{H}(\Omega) = \sum_{k=0}^{\infty} \alpha_k e^{-j\Omega 2k\tau} \tag{1.7}$$

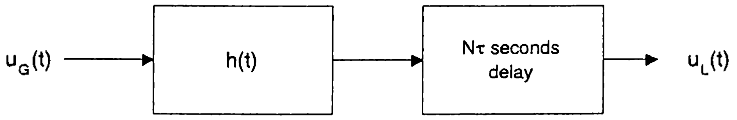


Figure 1.2: Simplified system.

Note that $\hat{H}(\Omega)$ is periodic with $\frac{2\pi}{2\tau}$ resembling the frequency response of a discrete time system like in Figure 1.3. If $u_G(t)$ is bandlimited with $\frac{\pi}{2\tau}$, we can sample it with period 2τ and filter it with a digital filter whose impulse response is $\hat{h}(n) = \alpha_n, n \geq 0$ and obtain $u_L(n)$, from which $u_L(t)$ can be reconstructed with an appropriate filter. Notice that, delay of $N\tau$ seconds corresponds to a shift of $\frac{N}{2}$ samples.

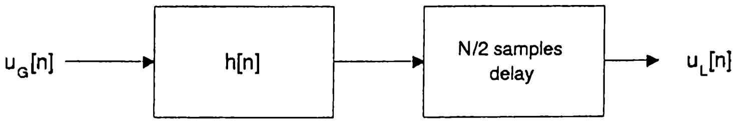


Figure 1.3: Discrete-time equivalent of the system.

The system function, corresponding to $\hat{h}(n)$, is

$$\hat{H}(z) = \sum_{k=0}^{\infty} \alpha_k z^{-k} \tag{1.8}$$

This transfer function can also be written in this form:

$$\hat{H}(z) = \frac{U_L(z)}{U_G(z)} \tag{1.9}$$

In order to make derivations for the transfer function, first, consider a single stage as shown in Figure 1.4 to find chain parameter representation of a 2-port network:

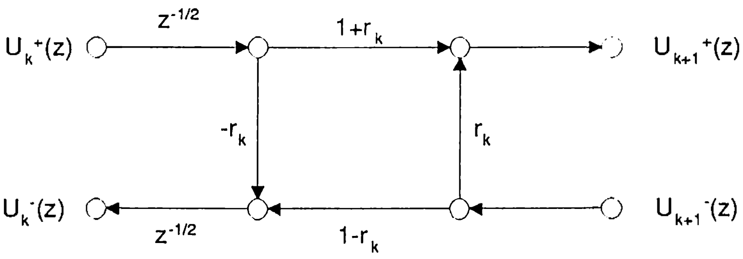


Figure 1.4: Single stage of transfer system.

$$\begin{aligned}
U_{k+1}^+(z) &= (1+r_k)z^{-1/2}U_k^+(z) + r_k U_{k+1}^-(z) \\
U_k^-(z) &= -r_k z^{-1} U_k^+(z) + (1-r_k)z^{-1/2}U_{k+1}^-(z) \\
U_k^+(z) &= \frac{z^{1/2}}{1+r_k}U_{k+1}^+(z) - \frac{r_k z^{1/2}}{1+r_k}U_{k+1}^-(z) && (1.10) \\
U_k^-(z) &= \frac{-r_k z^{-1/2}}{1+r_k}U_{k+1}^+(z) + \frac{z^{-1/2}}{1+r_k}U_{k+1}^-(z) && (1.11)
\end{aligned}$$

where

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{1.12}$$

and A_k is the cross-sectional area of the k^{th} junction.

The parameter, r_k, used in modeling is called reflection coefficients and can be used as a representation of the vocal tract. Furthermore, they are more suitable for quantization purposes, since their values are bounded with -1 and 1 for non-negative cross-sectional areas.

The chain matrix, Q_k and $\bar{U}_k$ is defined as:

$$Q_k = \begin{bmatrix} \frac{z^{1/2}}{1+r_k} & \frac{-r_k z^{1/2}}{1+r_k} \\ \frac{-r_k z^{-1/2}}{1+r_k} & \frac{z^{-1/2}}{1+r_k} \end{bmatrix} \tag{1.13}$$

$$\bar{U}_k = \begin{bmatrix} U_k^+(z) \\ U_k^-(z) \end{bmatrix} \tag{1.14}$$

so (1.10) and (1.11) can be expressed in matrix form:

$$\bar{U}_k = Q_k \cdot \bar{U}_{k+1} \tag{1.15}$$

To eliminate these half sample delays, a small modification can be performed on Figure 1.4. With an additional half sample delay at end of each stage, the half sample delay in the lower branch can be moved to upper branch to eliminate the usage of half sample delays. Result of the modification is illustrated in Figure 1.5:

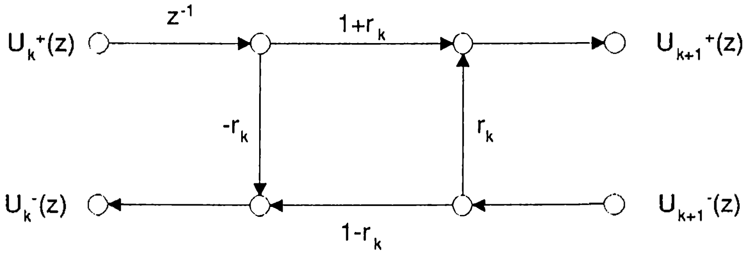


Figure 1.5: Modified single stage of transfer system.

$$Q'_k = \begin{bmatrix} \frac{z}{1+r_k} & \frac{-r_k z}{1+r_k} \\ \frac{-r_k}{1+r_k} & \frac{1}{1+r_k} \end{bmatrix} \tag{1.16}$$

Note that, $Q'_k = z^{1/2} Q_k$. Apart from the delay term, Q'_k and Q_k are equal, completing the discussion on the half-sample delays. Now, if N stages are considered like in Figure 1.6:

$$\bar{U}_1 = Q_1 \cdot Q_2 \cdots Q_N \cdot \bar{U}_{N+1} = \prod_{k=1}^{N} Q_k \cdot \bar{U}_{N+1} \tag{1.17}$$

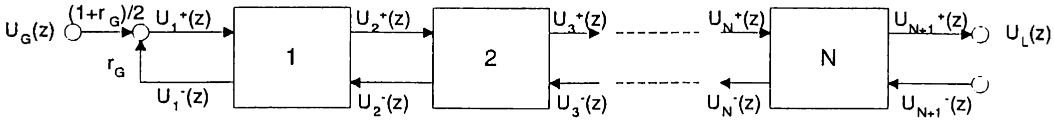


Figure 1.6: Interconnected N stages.

The equations for the boundaries are as follows:

$$\begin{aligned} U_G(z) &= \frac{2}{1+r_G}U_1^+(z) - \frac{2r_G}{1+r_G}U_1^-(z) \\ U_G(z) &= \left[\frac{2}{1+r_G}\frac{-2r_G}{1+r_G}\right]\bar{U}_1 \end{aligned} \tag{1.18}$$

and

$$\bar{U}_{N+1} = \begin{bmatrix} U_L(z) \\ 0 \end{bmatrix} \tag{1.19}$$

If we write transfer function, we obtain final formulation as follows:

$$\frac{U_G(z)}{U_L(z)} = \left[\frac{2}{1+r_G}\frac{-2r_G}{1+r_G}\right]\prod_{k=1}^{N} Q_k \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{H(z)} \tag{1.20}$$

where

$$Q_k = z^{1/2}\begin{bmatrix} \frac{1}{1+r_k} & \frac{-r_k}{1+r_k} \\ \frac{-r_k z^{-1}}{1+r_k} & \frac{z^{-1}}{1+r_k} \end{bmatrix} = z^{1/2}\hat{Q}_k \tag{1.21}$$

The elements of $\hat{Q}_k$ are either constants or z^{-1}, implying that complete matrix product will reduce to a polynomial in z^{-1} of order N. So transfer function can be written as:

$$H(z) = \frac{0.5\cdot(1+r_G)\cdot\prod_{k=1}^{N}(1+r_k)\cdot z^{N/2}}{D(z)} \tag{1.22}$$

where

$$D(z) = [1 - r_G]\cdot\begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix}\cdots\begin{bmatrix} 1 & -r_N \\ -r_N z^{-1} & z^{-1} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{1.23}$$

or

$$D(z) = 1 - \sum_{k=1}^{N}\alpha_k\cdot z^{-k} \tag{1.24}$$

Neglecting the delay term, $z^{N/2}$, transfer function can be expressed as:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k \cdot z^{-k}} \tag{1.25}$$

As a result, an all-pole model of vocal tract where the poles of $H(z)$ are the resonance frequencies, so called formants, of the acoustic tube system is obtained.

1.2 Linear Prediction

The linear prediction estimate $\hat{s}(n)$ of $s(n)$ from previous samples of $s(n)$ is defined as:

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_p(k) s(n-k) \tag{1.26}$$

where $\alpha_p(k)$ are the weights.

The error signal, $e(n)$, between the original and the predicted signal is defined as:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_p(k) s(n-k) \tag{1.27}$$

Now the problem is to select the coefficients, $\alpha_p(k)$, so that an error criterion is minimized. Generally, this error criterion is chosen to be the total squared error:

$$\varepsilon_p = \sum_{n=0}^{\infty} e^2(n) \tag{1.28}$$

$$\varepsilon_p = \sum_{n=0}^{\infty} (s(n) - \hat{s}(n))^2 = \left(s(n) - \sum_{k=1}^{p} \alpha_p(k) s(n-k) \right)^2 \tag{1.29}$$

To find $\alpha_p(k)$'s, which minimize ε_p, the derivative of ε_p with respect to $\alpha_p(j)$ must be equated to 0 and resulting equations must be solved:

$$\frac{\partial \epsilon_p}{\partial \alpha_p(j)} = \sum_{n=0}^{\infty} -2s(n-j) \left(s(n) - \sum_{k=1}^{p} \alpha_p(k) s(n-k) \right) = 0 \qquad \text{for j = 1 to p} \tag{1.30}$$

so:

$$\sum_{n=0}^{\infty} s(n)s(n-j) = \sum_{k=1}^{p} \alpha_p(k) \sum_{n=0}^{\infty} s(n-k)s(n-j) \qquad \text{for j = 1 to p} \qquad (1.31)$$

If we define $\omega(k,j)$ as:

$$\omega(k,j) = \sum_{n=b_l}^{b_h} s(n-k)s(n-j) \qquad (1.32)$$

and set b_l and b_h to zero and infinity, respectively, the final equation reduces to:

$$\omega(0,j) = \sum_{k=1}^{p} \alpha_p(k)\omega(k,j) \qquad \text{for j = 1 to p} \qquad (1.33)$$

The steps up to this point is the same for all modeling methods. The assumptions on the boundaries for the summation term in (1.32) make the difference for the finite sample modeling methods.

1.2.1 Autocorrelation Method

In this method, the signal is first windowed (generally with a Hamming window) and then it is used in the calculations above. Since boundaries, b_l and b_h, are set to 0 and ∞, respectively, $\omega(k,j)$ reduces to $\omega(|k-j|)$. After this modification, (1.33) reduces to:

$$\omega(j) = \sum_{k=1}^{p} \alpha_p(k)\omega(|k-j|) \qquad \text{for j = 1 to p} \qquad (1.34)$$

Now, we have p equations and p unknowns, hence $\alpha_p(k)$'s can be found by solving Equations (1.34) which can also be written in matrix form:

$$\begin{bmatrix} \omega(0) & \omega(1) & & \omega(p) \\ \omega(1) & \omega(0) & \dots \dots & \omega(p-1) \\ \vdots & \vdots & & \\ & & & \\ \omega(p) & \omega(p-1) & & \omega(0) \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \alpha_p(1) \\ \vdots \\ \\ \alpha_p(p) \end{bmatrix} = \begin{bmatrix} \varepsilon_p \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \qquad (1.35)$$

or

$$\mathbf{\Omega} \cdot \mathbf{a} = \varepsilon_p \cdot \mathbf{u_p} \tag{1.36}$$

where $\mathbf{a} = [1, \alpha_p(1), \ldots, \alpha_p(p)]^T$ and $\mathbf{u_p} = [1, 0, \ldots, 0]^T$.

The $\mathbf{a}$ vector can be found easily by:

$$\mathbf{a} = \varepsilon_p \cdot \mathbf{\Omega}^{-1} \cdot \mathbf{u_p} \tag{1.37}$$

Inverse of $\mathbf{\Omega}$ can be obtained by Gaussian elimination method. Since $\mathbf{\Omega}$ is a Toeplitz matrix, a recursive formulation, called *Levinson-Durbin recursion*, can also be used to obtain $\mathbf{a}$. The complexity of this algorithm is $O(n^2)$, compared to $O(n^3)$ of Gaussian elimination. The whole algorithm is reviewed in [18]. In addition to $\alpha_p(k)$'s, this recursion also produces reflection coefficients, r_k, and modeling error, ε_p, as side products. Note that the gain parameter in the all-pole formulation in (1.25) is equal to the square root of ε_p. As a final remark, the filter produced by this method is always stable [18].

1.2.2 Covariance Method

In covariance method, there is no assumptions made on the sequence. Lower boundary, b_l, and upper boundary, b_h, is set to p and the last element of the sequence, N, in (1.32), respectively. The matrix form of these equations is as follows:

$$\begin{bmatrix} \omega(1,1) & \ldots & \ldots & \omega(p,1) \\ & & & \vdots \\ & & & \\ \omega(1,p) & \ldots & \ldots & \omega(p,p) \end{bmatrix} \cdot \begin{bmatrix} \alpha_k(1) \\ \vdots \\ \\ \alpha_k(p) \end{bmatrix} = \begin{bmatrix} \omega(0,1) \\ \\ \\ \omega(0,p) \end{bmatrix} \tag{1.38}$$

The disadvantage of this method is that the positive definiteness of this matrix is not guaranteed, hence the filter, obtained by this method, may not be stable. Since matrix is not Toeplitz, these equations can not be solved by Levinson-Durbin equation. As there is no assumption on the input, the energy of the residual signal is smaller than the one extracted by autocorrelation method. Therefore, it provides better modeling of the input signal, especially for deterministic sequences.

1.3 Modeling of Human Speech Production System

As discussed in Section 1.1, the human vocal tract can be modeled by an all-pole filter. Furthermore, human speech production system uses two types of excitation to produce desired sounds:

1. Vocal folds make quasi-periodic movements to produce an air flow from lungs through glottis, which has an impulsive nature. This type of excitation can be modeled with impulse train whose periods are same as the period of this quasi-periodic movement.
2. Vocal folds are completely open to produce noise like sounds. This type of excitation can be modeled with uniform distributed white noise.

The block diagram of this basic model can be seen in Figure 1.7. In this model, the human vocal tract is modeled with an all-pole filter as discussed in Section 1.2. This filter is excited with either impulse train or white noise for voiced and unvoiced speech, respectively. Finally, a gain term is applied to the synthesized speech to amplify the signal to a desired level. Generally, it is assumed that statistical characteristics of speech signal do not vary for 20-30 ms periods. Hence, frames, whose lengths are between 20 to 30 ms, can be used to obtain synthetic speech. Most of the speech processing algorithms, mostly speech coders, use these facts to obtain synthetic speech with acceptable quality: Encoder only extracts the state of the voiced/unvoiced switch, pitch period for voiced speech, coefficients of LPC parameters and gain of the input speech and transmits these parameters with efficient quantization methods. The decoder uses these parameters to synthesize the desired speech signal.

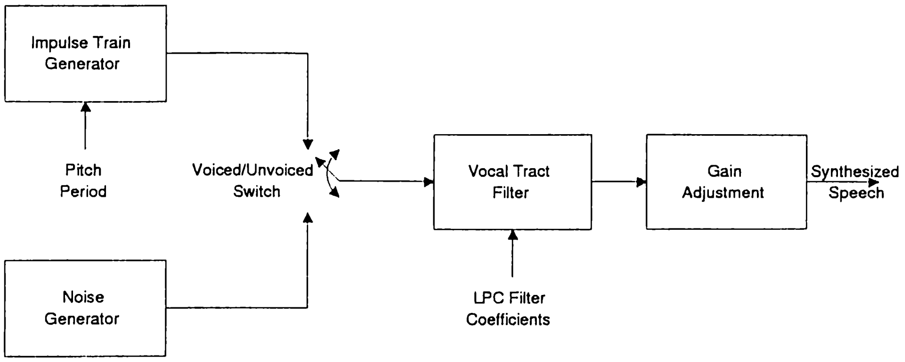


Figure 1.7: LPC vocoder synthesizer.

The vocal tract filter coefficients are not generally directly quantized: Dynamic range of these coefficients are high and quantization error may yield to an unstable filter. To solve these problems, new sets of parameters derived from LPC filter coefficients are used. These parameters can be summarized as follows:

1. Reflection coefficients,
2. log-area ratio parameters,
3. inverse sine transform of reflection coefficients, and
4. Line Spectrum Frequencies (LSF).

As stated before, reflection coefficients are the side products of the Levinson-Durbin recursion and they are used in the lattice form of the same all-pole filter. For the stable filters, these coefficients are bounded by 1 and -1, hence they are more suitable for quantization than LPC coefficients. Furthermore, it is possible to obtain reflection coefficients directly with Schur recursion without computing direct form of the LPC filter. Usage of this recursion enables the computation of these coefficients with fixed point arithmetic without considering loosing the stability of the filter.

Although the reflection coefficients are bounded by -1 and 1, the spectrum becomes very sensitive to quantization errors when the coefficients are close to the boundaries. To overcome this problem, two new sets of coefficients are introduced. Both of these transformations warp the scale of parameters and then uniform quantization of these parameters becomes non-uniform quantization for reflection coefficients.

Log-area ratio (LAR) is defined as,

$$LAR_i = \log \frac{1+r_i}{1-r_i} \tag{1.39}$$

and the inverse sine transform is defined as follows:

$$g_i = arcsin(r_i) \tag{1.40}$$

where r_i is the i^{th} reflection coefficient.

Both of these coefficients has good performance in quantization and hence they are widely used in vocoders. And also note that second one is also bounded between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$.

Besides these parameters, another type of parameters, called line spectrum frequencies, are used to quantize speech spectrum efficiently. These parameters have some unique features and have excellent performance in quantization purpose.

1.3.1 Line Spectrum Frequencies (LSF)

The linear prediction filter coefficients can be represented by Line Spectrum Frequencies (LSFs). This parameter set is first introduced by Itakura [19]. For a minimum phase, m^{th} order polynomial, $A_m(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_m z^{-m}$, one can construct two $(m+1)^{st}$ order LSF polynomials, $P_{m+1}(z)$ and $Q_{m+1}(z)$, by setting the $(m+1)^{st}$ reflection coefficient to 1 and -1 in Levinson-Durbin algorithm:

$$P_{m+1}(z) = A_m(z) + z^{-(m+1)} A_m(z^{-1}) \tag{1.41}$$

and

$$Q_{m+1}(z) = A_m(z) - z^{-(m+1)} A_m(z^{-1}) \tag{1.42}$$

This is equivalent to setting the vocal tract acoustic tube model completely closed or completely open at the $(m+1)^{st}$ stage. It is clear that $P_{m+1}(z)$ and

$Q_{m+1}(z)$ are symmetric and anti-symmetric polynomials, respectively. There are three important properties of these two polynomials:

1. All of the zeros of the LSF polynomials are on the unit circle and can be represented by only their angles,
2. the zeros of symmetric and anti-symmetric LSF polynomials are interlaced, and
3. the reconstructed linear prediction all-pole filter maintains its minimum phase property, if the first two properties are preserved during quantization.

Since these parameters can be represented by only angles, they are called line spectrum frequencies. Therefore the LSFs are also bounded between 0 and 2π, similar to reflection coefficients.

Besides these properties, in recent studies [16], it is found that LSFs are uncorrelated. This property of LSFs makes them a suitable parameter for quantization. It is also observed that LSFs are closely related to the speech formants as shown in Figure 1.8 and hence they provide a spectrally meaningful representation of the linear prediction filter [20]. Furthermore, it is observed that the spectral changes due to the perturbation of any LSF frequency is highly localized around the specific frequency [21].

Due to above reasons, the LSFs are widely used in speech coding [22] and speech recognition as speech feature parameters [23]. For example, it is possible to quantize coefficients of LSFs for 10^{th} order LPC filter by 21 bits for a speech frame of duration 20 ms without introducing any audible distortion [24]. Various quantization methods can be found in [25] for a review of scalar quantization methods, [26–30] for various vector quantization methods and [31–37] for different interframe quantization methods.

Several methods for the computation of LSFs are reported: The simplest way to compute LSFs are obtaining the root locations of these two polynomials by complex arithmetic. However, this method is obviously very complex and due to the iterative nature of the complex root finding algorithms, the time

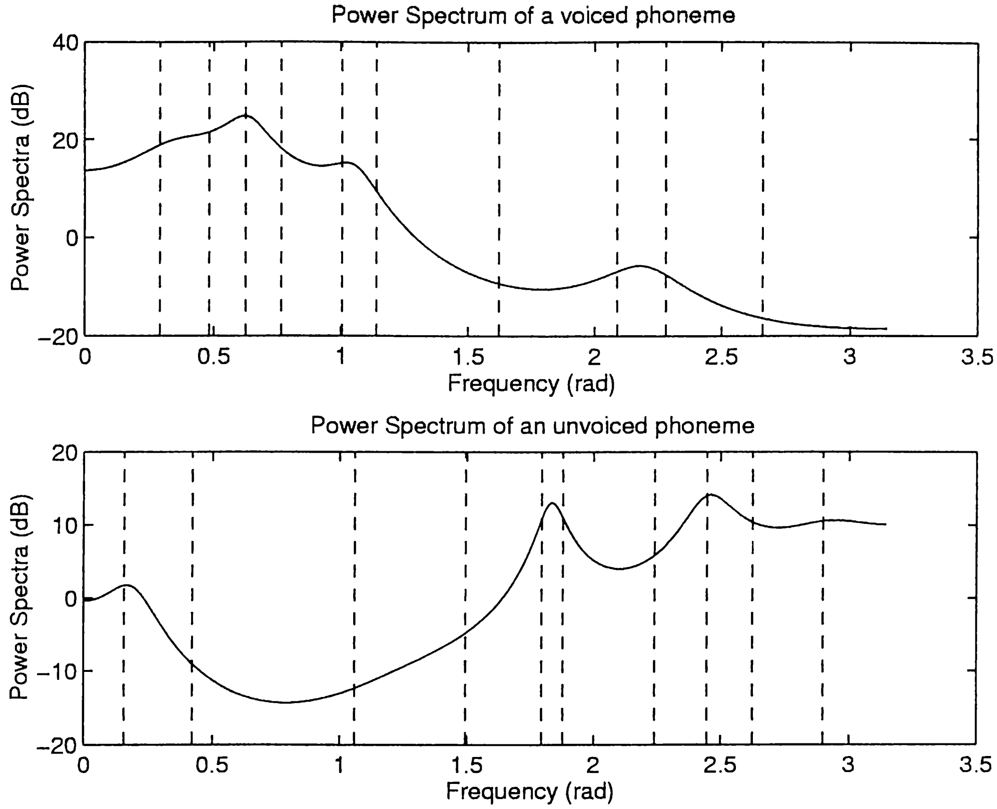


Figure 1.8: LP power spectrum and the associated LSFs for voiced and unvoiced speech.

required for the evaluation of this algorithm can not be estimated [38]. To overcome this problem, several methods are proposed: Soong and Juang have adopted a discrete cosine transform to evaluate cosine functions on a fine grid [39]. Furthermore, an all-pass ratio filter can also be used to extract locations of LSFs [38]. However, all of these methods require large number of computation of trigonometric functions. Therefore, Kabal and Ramachandran [40] presented a backward recursive formulation to determine the values of the cosine function on a fine grid by Chėbyshev's expansion and bisection method. Wu and Chen reported a similar method which uses a modified Newton-Raphson technique for faster convergence [41]. These latter two methods are widely used in real-time speech coding algorithms. Besides these methods, Goalic and Saoudi utilizes Split Levinson algorithm to compute LSFs independently [42]. Finally, an LMS based adaptive algorithm, applied in sample-by-sample basis, to find LSFs are reported by Cheetham [43].

Chapter 2

Spectrum Non-Stationarity Detection Algorithm Based on Line Spectrum Frequencies

One of the well-known properties of Line Spectrum Frequencies (LSFs) is that their locations are closely related with the peaks of the speech spectrum: Two or three consecutive LSFs are generally clustered to represent a peak, also called formant frequency, in the spectrum. As the formant frequencies change the LSF locations also change. By taking advantage of this fact, LSFs can be used to tract the changes in spectrum. We introduce two definitions which we use in the rest of this chapter:

1. The area between two consecutive LSFs is defined as an LSF region.
2. The difference between two consecutive LSFs is defined as the bandwidth of that LSF region.

The displacement of the LSFs usually gives a clue about the formation of the spectrum [25]: If the bandwidth of an LSF region is higher than its neighboring

LSF regions, usually a valley is present in the spectrum at this LSF region. Similarly, if the bandwidth of an LSF region is smaller than that of previous neighboring region and higher than that of next neighboring region, the energy of the spectrum is said to be increasing by increasing frequency. Besides, if the bandwidths of two consecutive LSF regions are almost the same, usually three LSFs come together to form a peak between these three LSFs. Coetzee and Barnwell [44] used the above relations to make a speech quality measurement algorithm based on LSFs.

However, the generalization described above may not be true in all cases. Sometimes formants become much closer to each other, and two consecutive LSF regions may both contain peaks, or sometimes an LSF region whose bandwidth is smaller than its neighboring regions may not contain a peak due to its wide bandwidth.

In addition to the bandwidth of the LSF regions, the spectral values at the LSF locations, extracted by evaluating prediction filter on the unit circle at the LSF locations, can be used to characterize the spectrum formation: If the energy of an LSF is larger than its neighboring LSFs, a peak is said to be present in the neighborhood of that LSF. In other words, it is easier to characterize the region by using both the difference between LSF locations and corresponding spectrum values.

In this chapter, a new and simple spectrum non-stationarity detector based on LSF related speech variation measures is introduced. In Section 2.1, an algorithm which estimates peaks of the spectrum, or the so-called formant frequencies for voiced speech, is presented. This section is divided into two parts: The first part describes the algorithm which makes the decision whether an LSF region contains a peak or not, and the second part describes the algorithm used to estimate the location of the peaks precisely. The non-stationarity detection algorithm, using speech variation measures based on the bandwidths of the LSF regions and peak locations are presented in Section 2.2. Simulation studies are given in Section 2.3.

2.1 Peak Estimation

In this section, a two-step peak estimation algorithm is presented. In the first step, the LSF regions which contain the peaks are detected and in the second step, the location of the peaks are calculated. Details of the first step and second step is presented in Section 2.1.2 and Section 2.1.3, respectively. In Section 2.1.1, the speech database used in this thesis is described.

2.1.1 Experimental Data

In this work, all required statistical data are obtained from two databases, owned by TÜBİTAK-BİLTEN, which contain 50 male and 50 female people in each database. The first database contains telephone speech with various hand-sets, while other one is formed by digitizing speech from close microphone talk. The databases include 12 words from each speaker - the numbers from zero to nine and '*yes*' and '*no*' in Turkish. Total number of frames for the voiced and the entire speech is approximately 40000 and 100000, respectively. Sampling rate is 8000 sample/sec and number of bits per sample is 16.

LSFs used in this work is extracted from the coefficients of 10^{th} order vocal tract filter, calculated by the autocorrelation method followed by Levinson-Durbin recursion [18]. This recursion uses Hamming windowed 200 samples, previously filtered with 4^{th} order type II-Chebyshev's high-pass filter with cut-off frequency at 60 Hz. Although it is known that covariance method gives better results, it is not used because of its computational complexity. The method used in extraction of LSFs is defined in [41], which uses a modified version of Newton-Rapson method for faster convergence in root finding algorithm. Before computation of LSFs, a bandwidth expansion of 15 Hz is applied to the poles of all-pole filter to increase bandwidth of the peaks.

2.1.2 Detection of a Peak in a LSF Region

Initially, two separate algorithms, running parallel, based on the bandwidths of the LSF regions and energies of LSFs are used to make initial peak estimation assignments to LSF regions.

In the bandwidth based method, the bandwidths are calculated and the LSF regions whose bandwidths are smaller than the bandwidths of their neighboring LSF regions are assigned to contain peaks in them. The bandwidth for the i^{th} LSF region, ψ_i, is defined as $f_i - f_{i-1}$, where f_i is the i^{th} LSF.

In the energy based method, first, logarithmic energies of line spectrum frequencies, P_i^{τ}, are calculated as follows:

$$P_i^{\tau} = \left(\left| \frac{1}{1 + \sum_{k=1}^{10} a_k e^{2jk\pi f_i}} \right| \right)^{2\tau} \tag{2.1}$$

In (2.1), τ is selected to be 0.15 as $P_i^{0.15}$ approximates the logarithm of the spectrum as shown in Figure 2.1. Also it can be observed that more emphasis is given to low frequency regions.

Equation (2.1) is also used in peak location estimation algorithm described in Section 2.1.3, where value of τ is varied for different LSF regions.

To find a region containing a peak, an energy-bandwidth based measure, E_{li} is defined for the LSF region before the i^{th} LSF:

$$E_{li} = \frac{P_{i-1}^{0.15}}{f_i - f_{i-1}} \tag{2.2}$$

where f_i represents the i^{th} LSF.

To detect the LSF region containing peak, (2.2) is applied to the previous and next LSF regions of the LSF, whose spectral value is larger than its neighboring LSFs. It is experimentally observed that the region, which gives the highest score, contains the peak in it.

After finding the peak locations with both algorithms, a merging strategy which reduces misclassification of the state of regions is applied to obtain final states of the regions.

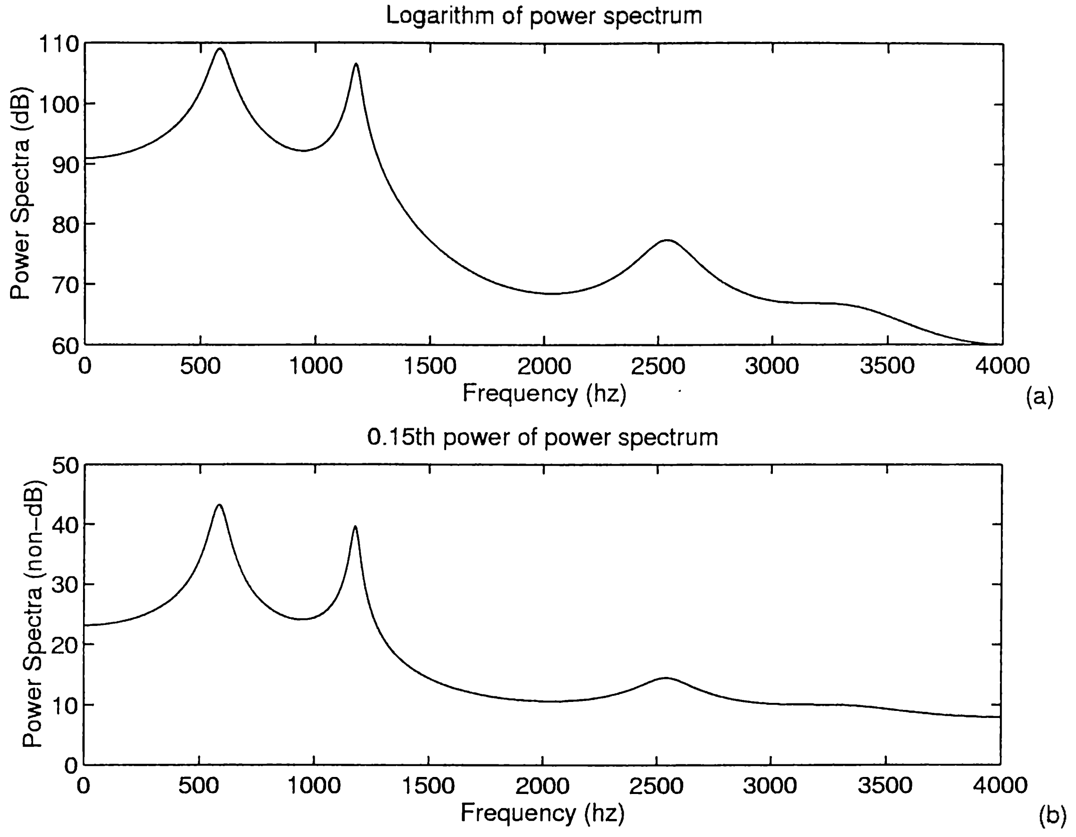


Figure 2.1: Logarithm and 0.15^{th} power of power spectrum of utterance /a/ is plotted in (a) and (b), respectively

Success rate of both algorithms is obtained from the data set described in Section 2.1.1. Since formants are tried to be estimated, part of the database, which contains only voiced speech, is used. For the voiced/unvoiced estimator, the one, based on normalized autocorrelation of the input sequence, described in detail in [17] is used with an exception that only the frames whose first bandpass voicing strength is larger than 0.8 is considered to be voiced speech. In other words, only strong voiced frames are considered in calculations. After deciding voiced speech, power spectrum is calculated with 1 Hz step size and a peak picking algorithm is applied to find peaks and also the LSF regions which contains these peaks are located. Statistics for correct classification and misclassification rates are calculated for bandwidth based method and energy based method separately and are tabulated in Table 2.1 and Table 2.2, respectively.

Table 2.1: Statistics about classification of LSF regions in bandwidth based method for voiced speech. P_C stands for the percentage of the correct classified LSF regions which contains a peak in it. P_M stands for the percentage of the misclassified LSF regions which contains a peak in it. P_{FP} stands for the percentage of the false peak assigned LSF regions with respect to total peak assigned LSF regions. P_{FNP} stands for the percentage of false peak assigned LSF regions with respect to the LSF regions which does not contain a peak.

	Regions								
	1	2	3	4	5	6	7	8	9
P_C	97.79	95.38	93.90	96.08	97.54	94.24	95.50	90.69	99.60
P_M	2.21	4.62	6.10	3.91	2.46	5.76	4.50	9.31	0.40
P_{FP}	4.33	23.98	63.41	33.37	24.33	21.32	26.10	28.70	35.42
P_{FNP}	22.13	2.79	18.54	10.63	18.90	5.97	22.77	6.60	45.75

Table 2.2: Statistics about classification of LSF regions in energy based method for voiced speech.

	Regions								
	1	2	3	4	5	6	7	8	9
P_C	98.77	91.56	67.98	93.45	93.99	97.85	90.76	98.28	93.00
P_M	1.33	8.44	32.02	6.55	6.01	2.15	9.24	1.72	7.00
P_{FP}	2.86	13.25	25.18	19.31	9.95	19.16	11.94	24.35	13.65
P_{FNP}	14.53	1.29	2.61	4.94	6.26	5.42	8.30	5.72	11.86

In these experiments, bandwidth based method is observed to detect approximately 96% of the regions containing peak, while in the energy based method, this number reduces to 94%. However, the critical problem in both of these methods are the large number of false peak assigned regions: In the third row of both tables, where percentage of false peak assigned regions with respect to total peak assigned regions are shown, nearly 25% and 12% of the peak assigned regions are false alarms for bandwidth based method and energy based method, respectively. Although some of these false alarms occur due to the selection of neighboring LSF regions of the LSF regions containing peak, remaining large number of false peak assignments other than this problem must be eliminated. The best solution for the elimination of these regions are found to be the selection of the only LSF regions detected by both methods.

Furthermore, sometimes three LSFs are clustered together to form a peak. In this case, bandwidths of the two LSF regions formed by these three LSFs are almost equal to each other, and usually peak location is around the middle LSF. If this formation is occurred, the proposed methods detect only one of these LSF regions and sometimes, they detect different LSF regions from these two LSF regions. Because of our merging criteria, such peaks are missed. In order to get rid of this problem, if detection of one LSF region by one method and its neighboring LSF region by other method is encountered, a small test is applied to both LSF regions to select the correct one: If the absolute difference between the bandwidths of these two LSF regions are smaller than 8 percent of the bandwidth of the detected LSF region by bandwidth based method, the LSF region detected by energy based method is considered to be true. Otherwise decision of the bandwidth based method is accepted.

Unfortunately, still large number of false estimations occur in some regions and also some peaks, estimated by one method but missed by other one, are remaining. To eliminate these false peaks, a bandwidth threshold, T_i, is assigned to i^{th} LSF region. If bandwidth of the i^{th} LSF region is larger than T_i, this detected peak is assumed to be a false peak. Also to include missing peaks, detected by only one method, following two tests are applied to those regions:

1. If bandwidth of the i^{th} region is smaller than a threshold then a peak is assigned to the i^{th} LSF region. This lower bandwidth threshold for the i^{th} LSF region is γ_i for bandwidth based method and α_i for energy based method.

2. Let us define the energy-bandwidth based measure, ϵ_i, as follows:

$$\epsilon_i = \frac{(P_i^{0.15} + P_{i-1}^{0.15})}{(f_i - f_{i-1})} \tag{2.3}$$

   for the i^{th} LSF region. If ϵ_i is larger than a threshold then a peak is assigned to the i^{th} region. This higher energy-bandwidth based measure threshold for the i^{th} LSF region is β_i for bandwidth based method, and ζ_i for energy based method.

The thresholds used in this algorithm are also estimated from the same database. In order to obtain the thresholds for false peak assignment elimination, the distribution of the percentage of correct estimation and false estimation versus the threshold, T_i, is calculated and the point, which introduce minimum loss of correct detected LSF regions and provide maximum false peak assignment elimination is selected. Similar calculations are performed for the thresholds that are used to catch the misclassified LSF regions which actually contains peaks. These thresholds are selected so that maximum number of misclassified LSF regions are corrected, while minimum number of false peak assignment is introduced. The flowdiagram of the algorithm is given in Figure 2.2, and final thresholds are tabulated in Table 2.3. The distribution of the percentage of correct estimation and false estimation versus thresholds for the LSF regions is described in Appendix A in detail.

Table 2.3: Thresholds for elimination and detection of misclassified regions.

	Regions								
	1	2	3	4	5	6	7	8	9
T_i	0.251	0.236	0.251	0.259	0.251	0.267	0.255	0.267	0.314
γ_i	0.102	0.083	0.129	0.102	0.126	0.110	0.149	0.116	0.130
α_i	0.129	0.102	0.168	0.196	0.196	0.134	0.196	0.094	0.196
β_i	N/A	N/A	99949	N/A	101858	40744	31831	95493	40744
ζ_i	127324	127324	79578	26738	22282	50930	22282	95493	40744

Final statistics about proposed classification method is given in Table 2.4, Table 2.5 and Table 2.6. In these tables, it can be seen that number of false peak assigned regions reduces dramatically. Furthermore, the highest percentage of false alarm, which occurs in the 3^{rd} region, only contains 3.5% of whole peaks. If all regions are considered together, false alarm rates goes down to 5% and 10% for voiced speech and entire speech, respectively. It must be noted that since bandwidths are wide in unvoiced speech, increase in the false alarm rate is expected. Also, total percentage of misclassified regions, which contain peaks, remains at 7%. Overall results can be seen in Table 2.6: Approximately 95% of regions are classified correctly by proposed method for both voiced and whole speech.

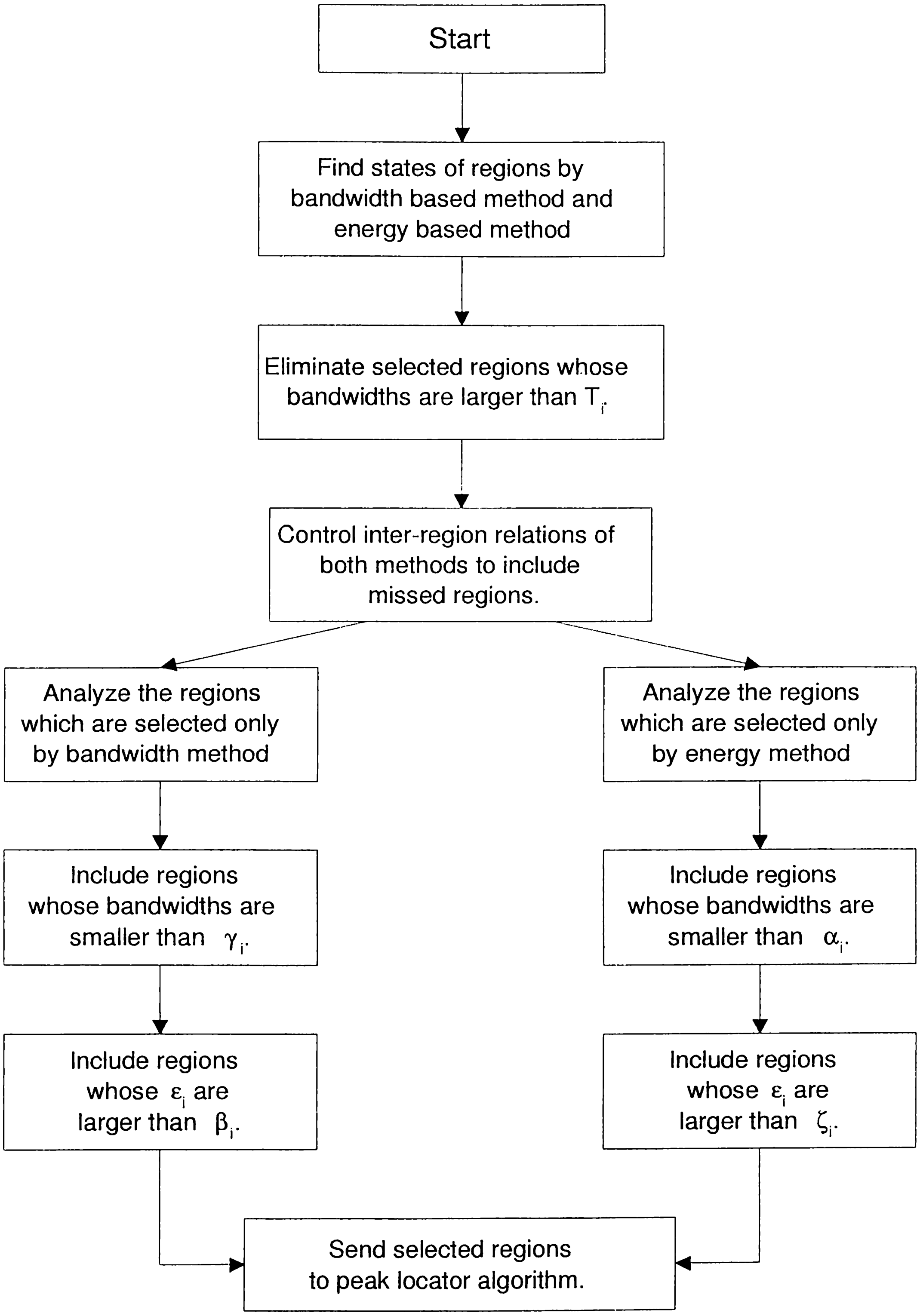

Start
Find states of regions by bandwidth based method and energy based method
Eliminate selected regions whose bandwidths are larger than T_i.
Control inter-region relations of both methods to include missed regions.
Analyze the regions which are selected only by bandwidth method
Analyze the regions which are selected only by energy method
Include regions whose bandwidths are smaller than γ_i.
Include regions whose bandwidths are smaller than α_i.
Include regions whose ε_i are larger than β_i.
Include regions whose ε_i are larger than ζ_i.
Send selected regions to peak locator algorithm.

Figure 2.2: Algorithm for classification of the LSF regions.

Table 2.4: Statistics about classification of LSF regions in both methods for voiced speech.

	Regions								
	1	2	3	4	5	6	7	8	9
P_C	98.39	81.23	70.11	86.84	93.22	92.64	90.54	85.85	92.67
P_M	1.61	18.77	29.89	13.16	6.78	7.36	9.46	14.15	7.33
P_{FP}	1.78	4.80	11.20	7.13	3.78	6.85	5.92	10.53	9.80
P_{FNP}	10.22	0.56	0.96	1.63	2.38	1.79	4.30	2.12	8.29

Table 2.5: Statistics about classification of LSF regions in both methods for entire speech.

	Regions								
	1	2	3	4	5	6	7	8	9
P_C	95.38	80.93	74.45	90.95	91.10	92.30	90.56	86.07	94.69
P_M	4.62	19.07	25.55	9.05	8.90	7.70	9.44	13.93	5.31
P_{FP}	3.06	7.00	14.03	12.72	10.03	15.39	12.56	14.69	16.34
P_{FNP}	8.64	0.30	2.20	2.98	5.09	3.11	7.54	2.20	15.99

Table 2.6: Overall performance of the system for the proposed method. Percentage of correct classification of the LSF region state is tabulated.

	Regions								
	1	2	3	4	5	6	7	8	9
Entire speech	94.34	98.82	94.22	95.90	93.64	96.16	91.76	96.28	88.97
Voiced speech	97.11	97.25	96.22	96.10	95.89	97.06	93.47	95.79	92.14

2.1.3 Accurate Estimation of Peak Location

After finding the regions which contains the peaks, another algorithm is applied to find the exact location of the peak. In Coetzee and Barnwell's work [44], peak location is estimated by the mean of the two LSFs which form the region. This estimate gives acceptable peak locations only if the bandwidth of that region is sufficiently small - smaller than 150 Hz. Since bandwidth of the LSF regions may become as large as 300 Hz, this estimate will not give satisfactory results and the difference between actual peak and estimated peak may be as large as 150 Hz. As an alternative, weighted means, whose weights are the same as the one used in quantization of LSFs in [30], may be used to estimate peak locations as follows:

$$\rho_i = f_{i-1} + (f_i - f_{i-1})\frac{P_i^{\tau}}{P_i^{\tau} + P_{i-1}^{\tau}} + \mu_i \tag{2.4}$$

where ρ_i represents the location of the peak in the i^{th} region and and μ_i is the correction term for the peak in the i^{th} LSF region. For this alternative procedure, τ is chosen as 0.15 and μ_i is set to zero for all LSF regions.

Unfortunately, this energy weighted mean only makes a small improvement in the peak location estimation. In order to get better results, different values of τ may be considered. For this purpose, mean, standard deviation, percentage of peak estimation error smaller than 25 Hz and percentage of peak estimation error smaller than 50 Hz are calculated for different values of τ, ranging from 0.15 to 1.75, for both voiced and entire speech data for all LSF regions. Based on this experiment, different τ values are selected for different LSF regions such that the standard deviation would be minimum or closer to minimum and percentage of peak estimation error smaller than 25 Hz is maximum. This criterion is selected, because if we try to maximize the percentage of peak estimation error smaller than an error range lower than 25 Hz, the percentage of peak estimation error larger than 50 Hz is also increased, which exceeds the acceptable range. Furthermore, Schafer *et al.* reported that 25 Hz error is negligible for formant estimation [45]. Besides, the correction term in (2.4), μ_i, is selected as the mean of the error between actual and estimated peak location for the selected τ value for the i^{th} LSF region.

After selecting optimum exponents for all LSF regions, (2.4) is used to estimate peak locations with different τ for different LSF regions. Selected τ's and μ's are tabulated in Table 2.7. Number of occurrence versus error in estimation of peak location for voiced speech and whole speech is given in Figure2.3a and Figure2.3b, respectively. Nearly 97% and 95% of the peaks are estimated within 25 Hz error range for the voiced and the entire speech, respectively. These values increases to 99.5% and 99%, if 50 Hz is also accepted as a tolerable error range.

The figures for standard deviation, percentage of peak estimation error smaller than 25 Hz and percentage of peak estimation error smaller than 50

Table 2.7: Selected τ_i and corresponding mean of the error in peak estimation, μ_i, in radian.

	Regions								
	1	2	3	4	5	6	7	8	9
τ_i	1.25	1.3	1.125	1.3	1.175	1.5	1.175	1.45	1.15
μ_i	0.0057	0.0058	0.0019	−0.0009	0.0009	−0.0001	0.0018	0.0011	0.0040

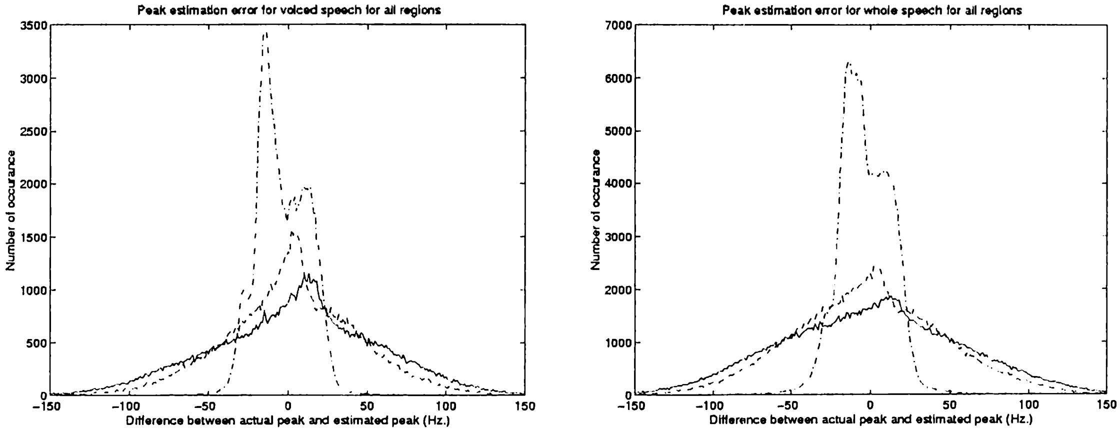


Figure 2.3: Number of occurance versus difference between original and estimated peaks. The solid, dashed and dashed dotted lines corresponds to simple mean, energy weighted mean with $\tau = 0.15$ and energy weighted mean with $\tau = \tau_i$, respectively.

Hz versus different τ values and tabulated form of statistics for simple mean estimation method, energy weighting method for $\tau = 0.15$ and energy weighting method for $\tau = \tau_i$ are given in Appendix B. Proposed method decreases standard deviation four times compared to simple mean method and increase accuracy of the estimation dramatically.

Without considering the non-stationary region detector, the output of this algorithm may be used as a formant tracker in conjunction with an voiced/unvoiced estimator. An example for this formant estimator is illustrated in Figure 2.4 for the Turkish sentence "Şanslı adam kaybettiği mücheveri buldu.".

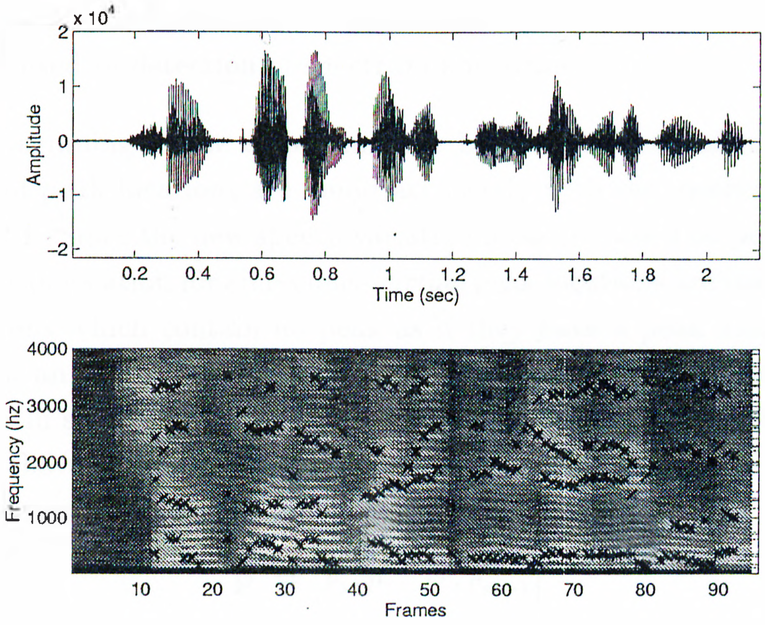


Figure 2.4: Time sequence and spectrogram of Turkish word ”Şanslı adam kaybettiği mücheveri buldu”. Formant frequencies are plotted with ”*” on the spectrogram.

2.2 Non-stationarity Detection

The non-stationarity of the speech spectrum can be detected using the LSF based peak estimation method described in Section 2.1. The simplest way is to examine the L_2 norm of the difference between the peak locations. Unfortunately, since the peak estimation algorithm may miss some peak locations or the number of peaks may change especially in transient regions, direct application of L_2 norm to estimated peak location will not give good results. Furthermore, L_2 norm lacks of incorporating perceptual information to the speech variation measure, which is essential in speech. A weighted Euclidean distance measure whose weights are selected according to the nature of peaks is more suitable. Furthermore, instead of using only the estimated peaks, all data related with the LSF regions are used in the calculation of speech variation measure for better results. In this section, two speech variation measures, one based on the bandwidth of the LSF regions and one based on the peak locations in the LSF regions, are used in detection of spectrum non-stationarity.

In the beginning of the algorithm, the LSF regions containing peaks are detected and peak locations are found accurately with the methods described in Section 2.1. Since the new speech variation measure based on peak locations requires a peak location for all regions, *virtual* peak locations are computed even for the regions which contain no peak as if they have a peak using the same method. For an m^{th} order LPC model, $m-1$ peak locations are calculated. In order to use in speech variation measure, a vector for the k^{th} frame, $\mathbf{p}^k$, whose entries are the weighted difference between the peak locations of the k^{th} and $(k-1)^{th}$ frames, is defined as follows:

$$\mathbf{p}^k = \left[p_1^k, p_2^k, \cdots, p_{m-1}^k\right] \tag{2.5}$$

where $p_i^k = w_i(\rho_i^k - \rho_i^{k-1})$. The weights, w_i, are obtained experimentally and set according to the state of the regions for the consecutive frames to emphasize the change in state of the LSF region as follows:

1. If the i^{th} region in both frames do not contain peak, w_i is selected to be 0.1.
2. If the i^{th} region in both frames contain peak, w_i is selected to be 0.8.
3. If the i^{th} region in only one of the frames contains peak, w_i is selected to be 1.0.

The speech variation measure based on peak locations, Λ_k, for the k^{th} frame is defined as follows:

$$\Lambda_k = \bar{\Lambda}_k^k + \bar{\Lambda}_k^{k-1} \tag{2.6}$$

where $\bar{\Lambda}_k^k = \mathbf{p}^k \mathbf{W_p}^k \mathbf{p}^{k^T}$ and $\bar{\Lambda}_k^{k-1} = \mathbf{p}^k \mathbf{W_p}^{k-1} \mathbf{p}^{k^T}$ and $\mathbf{W_p}^l$ is the weighting matrix whose entries are determined according to the perceptual sensitivity of the peak locations estimated for the l^{th} frame. By this method, change in the different part of the spectrum is emphasized in a perceptual manner.

Entries of $\mathbf{W_p}$ are determined according to the relationship between the peak locations in the spectrum. We define the weights as the correlation between the peak locations. As discussed in Chapter 1, LSFs are reported to be uncorrelated [16]. If LSFs are used in our system directly, calculation of the main diagonal entries of the weighting matrix would be sufficient and rest of the entries will be set to zero. As peaks are derived from two consecutive LSFs, consecutive peaks are also correlated. Therefore, the diagonal entries next to the main diagonal of the weighting matrix must also be calculated. Although usage of peaks instead of LSFs may seem redundant, it must be noted that the LSF regions which contain a formant can be emphasized easily by this method while forming the vector $\mathbf{p}^k$.

Entries of $\mathbf{W_p}$ are computed according to (2.10). Note that the entries other than main diagonal entries and diagonal entries next to the main diagonal are set to zero, as there is no correlation between those peak locations.

Let us rewrite (2.4):

$$\begin{aligned} \rho_i &= f_{i-1} + (f_i - f_{i-1})\frac{P_i^T}{P_i^T + P_{i-1}^T} + \mu_i \\ &= \omega_{i1} f_i + \omega_{i2} f_{i-1} + \mu_i \end{aligned} \tag{2.7}$$

where

$$\omega_{i1} = \frac{P_i^T}{P_i^T + P_{i-1}^T} \tag{2.8}$$

and

$$\omega_{i2} = 1 - \frac{P_i^T}{P_i^T + P_{i-1}^T} \tag{2.9}$$

Entries of the weighting matrix are defined as follows:

$$\begin{aligned} \mathbf{W}_{\mathbf{p}_{ij}} &= \mathcal{E}\{\rho_i \rho_j\} \\ &= \mathcal{E}\{(\omega_{i1} f_i + \omega_{i2} f_{i-1} + \mu_i)(\omega_{j1} f_j + \omega_{j2} f_{j-1} + \mu_j)\} \end{aligned} \tag{2.10}$$

The main diagonal entries are given as follows:

$$\begin{aligned} \mathcal{E}\{\rho_i \rho_i\} &= \mathcal{E}\{(\omega_{i1} f_i + \omega_{i2} f_{i-1})(\omega_{i1} f_i + \omega_{i2} f_{i-1})\} \\ &= \omega_{i1}^2 \mathcal{E}\{f_i^2\} + 2\omega_{i1}\omega_{i2}\mathcal{E}\{f_i f_{i-1}\} + \omega_{i2}^2 \mathcal{E}\{f_{i-1}^2\} \\ &= \omega_{i1}^2 \mathcal{E}\{f_i^2\} + \omega_{i2}^2 \mathcal{E}\{f_{i-1}^2\} \end{aligned} \tag{2.11}$$

and the diagonal entries next to the main diagonal are given by:

$$\begin{aligned} \mathcal{E}\{\rho_i \rho_{i+1}\} &= \mathcal{E}\{(\omega_{i1} f_i + \omega_{i2} f_{i-1})(\omega_{(i+1)1} f_{i+1} + \omega_{(i+1)2} f_i)\} \\ &= \omega_{i1}\omega_{(i+1)2}\mathcal{E}\{f_i^2\} \end{aligned} \tag{2.12}$$

and

$$\begin{aligned}\mathcal{E}\{\rho_i\rho_{i-1}\} &= \mathcal{E}\{(\omega_{i1}f_i + \omega_{i2}f_{i-1})(\omega_{(i-1)1}f_{i-1} + \omega_{(i-1)2}f_{i-2})\} \\ &= \omega_{i2}\omega_{(i-1)1}\mathcal{E}\{f_{i-1}^2\} \end{aligned} \qquad (2.13)$$

The other entries turn out to be zero as shown below:

$$\begin{aligned}\mathcal{E}\{\rho_i\rho_{i+k}\} &= \mathcal{E}\{(\omega_{i1}f_i + \omega_{i2}f_{i-1})(\omega_{(i+k)1}f_{i+k} + \omega_{(i+k)2}f_{i+k-1})\} \\ &= 0 \qquad k \geq 2 \text{ or } k \leq -2 \end{aligned} \qquad (2.14)$$

because $\mathcal{E}\{f_i f_j\} = 0$ for $i \neq j$ and μ_i's are neglected in calculations because these values are negligibly small compared to the other parameters in (2.10).

The only missing part in this formulation is the variance of the LSFs. For the variances, $\mathcal{E}\{f_i^2\} = P_i^{0.15}$ is used similar to [30,46].

Speech variation measure, Γ, based on the bandwidths of the LSF regions is also similarly determined. First of all, the area between 0^{th} and 1^{st} LSFs and the area between last LSF and π are also considered as LSF regions for this speech variation measure. Therefore, $m+1$ parameters are extracted for the m^{th} order LPC filter. A vector for the k^{th} frame, $\mathbf{b}^k$, whose entries are the weighted difference between the bandwidths of the LSF regions in the k^{th} and $(k-1)^{th}$ frame is defined as:

$$\mathbf{b}^k = \left[b_0^k, b_1^k, \cdots, b_m^k\right] \qquad (2.15)$$

where $b_i^k = w_i(\psi_i^k - \psi_i^{k-1})$ for $i = 1$ to m-1. Note that ψ_0 and ψ_m is equal to f_0 and $(\pi - f_{m-1})$, respectively. The weights, w_i, are assigned according to the following criteria and are determined experimentally:

1. If the i^{th} region in both frames do not contain peak, w_i is selected to be 0.25.

2. If the i^{th} region in both frames contain peak, w_i is selected to be 0.9.

3. If the i^{th} region in only one of the frames contains peak, w_i is selected to be 1.0.

Since the first and the last regions can never contain a peak, their weights are automatically set to 0.25.

The speech variation measure, Γ_k, for the k^{th} frame is defined as follows:

$$\Gamma_k = \bar{\Gamma}_k^k + \bar{\Gamma}_k^{k-1} \tag{2.16}$$

where $\bar{\Gamma}_k^k = \mathbf{b}^k \mathbf{W_b}^k \mathbf{b}^{k^T}$ and $\bar{\Gamma}_k^{k-1} = \mathbf{b}^k \mathbf{W_b}^{k-1} \mathbf{b}^{k^T}$ and $\mathbf{W_b}^l$ is the weighting matrix whose entries are determined according to the perceptual sensitivity of the bandwidth of the peaks for the l^{th} frame.

The entries of the weighting matrix for bandwidths are defined as follows:

$$\begin{aligned} \mathbf{W_b}_{ij} &= \mathcal{E}\{\psi_i \psi_j\} \\ &= \mathcal{E}\{(f_i - f_{i-1})(f_j - f_{j-1})\} \end{aligned} \tag{2.17}$$

The main diagonal entries are given as follows:

$$\begin{aligned} \mathcal{E}\{\psi_i \psi_i\} &= \mathcal{E}\{(f_i - f_{i-1})(f_i - f_{i-1})\} \\ &= \mathcal{E}\{f_i^2\} + \mathcal{E}\{f_{i-1}^2\} \end{aligned} \tag{2.18}$$

The diagonal entries next to the main diagonal are given by:

$$\begin{aligned} \mathcal{E}\{\psi_i \psi_{i+1}\} &= \mathcal{E}\{(f_i - f_{i-1})(f_{i+1} - f_i)\} \\ &= -1 \cdot \mathcal{E}\{f_i^2\} \end{aligned} \tag{2.19}$$

and

$$\begin{aligned} \mathcal{E}\{\psi_i \psi_{i-1}\} &= \mathcal{E}\{(f_i - f_{i-1})(f_{i-1} - f_{i-2})\} \\ &= -1 \cdot \mathcal{E}\{f_{i-1}^2\} \end{aligned} \tag{2.20}$$

The other entries turn out to be zero as shown below:

$$\begin{aligned} \mathcal{E}\{\psi_i \psi_{i+k}\} &= \mathcal{E}\{(f_i - f_{i-1})(f_{i+k} - f_{i+k-1})\} \\ &= 0 \qquad k \geq 2 \text{ or } k \leq -2 \end{aligned} \tag{2.21}$$

Finally, the only remaining task is to compare the calculated speech variation measures with the experimentally determined thresholds, λ and η, for Λ

and Γ, respectively. However, to eliminate small fluctuations, one frame delay is introduced to compare the calculated measures with the ones calculated in previous and next frames: If the variation calculated for any of the previous or next frame is larger than the current one, the calculated variation measure is not compared with the thresholds and frame is assumed to be stationary. Otherwise, variation values are compared with thresholds and if one of the speech variation measures exceed these thresholds, frame is flagged to be non-stationary. This method extracts the frames which has the maximum variation with respect to its neighboring frames.

Values of λ and η can be adjusted for different purposes: It may be lowered to catch small changes, or it may be set to high values to catch only abrupt changes.

2.3 Simulation Studies

In this section, two simulations are performed to test the performance of proposed algorithms. We make the first simulation to test the performance of peak estimation algorithm whose results are given in Section 2.3.1. The other simulation is carried out to see the performance of non-stationarity detector. The results for this test are presented in Section 2.3.2.

2.3.1 Performance Test for Peak Estimation Algorithm

For this simulation, a database, owned by TÜBİTAK-BİLTEN, containing voice of 104 people including child voice is used. This database is formed by digitizing speech from close microphone talk. The database includes 10 words from each speaker - the numbers from zero to nine in Turkish. Total number of frames for the voiced and the entire speech is approximately 27000 and 40000, respectively. Sampling rate is 8000 sample/sec and number of bits per sample is 16.

Performance Test for Peak Detection Algorithm

Performance of peak detection algorithm for the voiced and the entire speech is given in Table 2.8 and Table 2.9, respectively. Overall results are also illustrated in Table 2.10. In this test set, it is observed that the performance of the algorithm is similar compared to the one obtained in training set, even better results for some LSF regions.

Table 2.8: Statistics about classification of LSF regions for proposed algorithm for the voiced speech for the test set.

	Regions								
	1	2	3	4	5	6	7	8	9
P_C	90.79	95.89	85.05	89.35	89.77	96.85	91.05	94.62	96.19
P_M	9.21	4.11	14.95	10.65	10.23	3.15	8.95	5.38	3.81
P_{FP}	4.15	4.19	15.67	16.08	8.78	13.04	8.42	14.40	16.96
P_{FNP}	2.49	1.17	0.96	1.17	3.08	3.93	3.24	4.08	12.26

Table 2.9: Statistics about classification of LSF regions for proposed algorithm for the entire speech for the test set.

	Regions								
	1	2	3	4	5	6	7	8	9
P_C	90.01	93.73	75.93	90.26	89.20	96.04	90.01	94.18	96.34
P_M	9.99	6.27	24.07	9.74	10.80	3.96	9.99	5.82	3.66
P_{FP}	7.10	6.59	15.27	18.40	10.51	15.53	12.91	18.19	22.20
P_{FNP}	3.14	0.94	1.59	2.54	3.75	4.78	4.20	5.17	13.68

Table 2.10: Overall performance of the proposed method for the test set. Percentage of correct classification of the LSF region state is tabulated.

	Regions								
	1	2	3	4	5	6	7	8	9
Entire speech	94.71	98.40	96.07	96.66	94.40	95.40	94.42	94.70	89.65
Voiced speech	94.90	98.19	97.58	97.49	95.03	96.24	95.17	95.66	90.99

Performance Test for Accurate Peak Location Estimation Algorithm

In this test, percentage of peak estimation error smaller than 25 Hz and 50 Hz are extracted from the same database for selected τ_i and μ_i values and the results are tabulated in Table 2.11 and Table 2.12, for the voiced and the entire speech, respectively.

Table 2.11: Percentage of the error smaller than 25 Hz and 50 Hz between actual peak location and estimated peak location for test set for the voiced speech.

	LSF Regions				
	1	2	3	4	5
$\leq$ 25 Hz	97.79	95.85	93.40	94.51	96.60
$\leq$ 50 Hz	100.00	99.80	99.53	99.42	99.58
	6	7	8	9	All Regions
$\leq$ 25 Hz	91.31	92.55	87.17	88.09	92.90
$\leq$ 50 Hz	98.79	98.73	97.62	98.66	99.11

Table 2.12: Percentage of the error smaller than 25 Hz and 50 Hz between actual peak location and estimated peak location for test set for the entire speech.

	LSF Regions				
	1	2	3	4	5
$\leq$ 25 Hz	94.75	92.65	91.88	90.42	94.60
$\leq$ 50 Hz	99.42	99.73	99.33	97.96	99.24
	6	7	8	9	All Regions
$\leq$ 25 Hz	88.21	91.11	85.45	84.97	90.19
$\leq$ 50 Hz	98.29	98.49	96.85	97.08	98.46

From these results, it is observed that nearly 93% and 90% of the peaks are estimated within 25 Hz error range for voiced and the entire speech, respectively. This states a small performance degradation over the training set. As percentage for the error smaller than 50 Hz are approximately 99% for both voiced and entire speech, it can be said that this estimation algorithm can still obtain location of the peaks in an acceptable range.

2.3.2 Performance Test for Non-Stationarity Detector

Simulations for this section is performed on three artificially created speech signals. All sequences are created by the MELP synthesizer [17] and they contain two synthesized speech phonemes with a transitional region between them. Parameters used in the synthesizer are extracted from a real speech signal. Proposed algorithm uses a frame length of 22.5 ms and a step size of 11.25 ms. LSFs are extracted with the same method described in Section 2.1.1 and λ and η are set to 0.1 and 0.2, respectively. These thresholds are determined experimentally from part of the same database described in Section 2.1 and found sufficient to detect small changes.

Three signals are used in this experiment:

1. The first signal contains two noise excited parts: First half contains utterance */sh/*, where as the second half contains utterance */s/*. This signal is used to test performance of non-stationarity detector for stochastic signals.

2. The second signal contains two periodic impulse excited parts: First half contains utterance */a/*, where as the second half contains utterance */o/*. This signal is used to test performance of non-stationarity detector for deterministic signals.

3. The third signal contains both periodic impulse and noise excited parts: First half contains utterance */sh/*, where as the second half contains utterance */o/*. This signal is used to test performance of non-stationarity detector in regions where both excitation and spectral shape change.

The result of the application of the algorithm on these signals can be seen in Figure 2.5.

Although these three experiments do not give much information about the real performance of the algorithm, the simulation section of the next chapter, which describes an implicit speech segmentation algorithm, gives detailed success rates of this algorithm for different parameter settings.

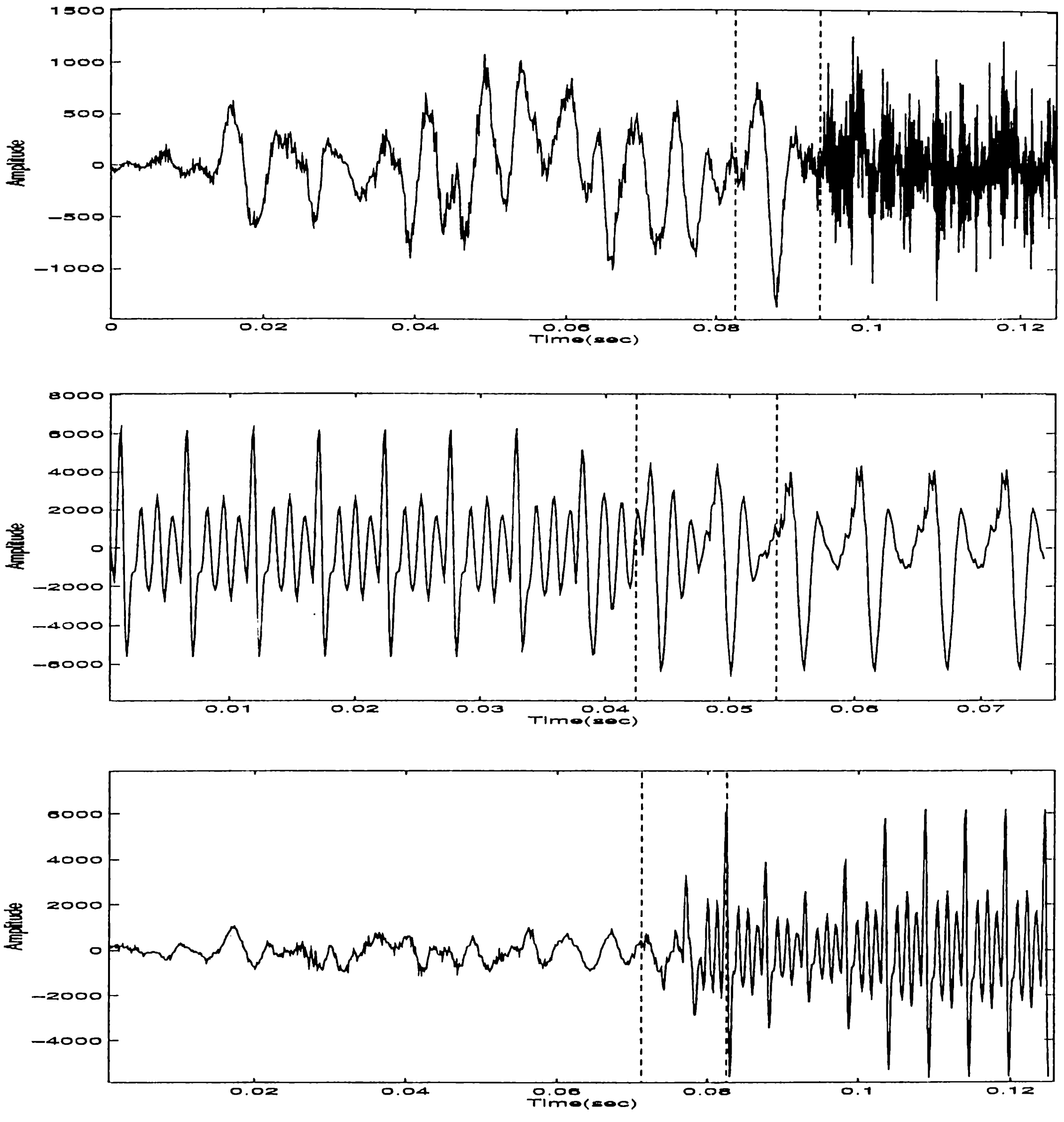


Figure 2.5: The top, middle and bottom figures belong to the signals which contain two noise excited regions, two pulse excited regions and both excitation and spectral change regions, respectively. The region between the dashed lines are the ones which are flagged as non-stationary regions by the proposed algorithm.

2.4 Summary

In this chapter, a new spectral peak location estimation and spectrum non-stationarity detection algorithm based on LSFs are presented. Proposed algorithm consists of three parts:

1. Detection of the peak presence in the current LSF region.
2. Accurate estimation of the location of the peaks.
3. Detection of the spectrum non-stationarity based on two speech variation measures using the peak presence in LSF regions, and the location of the peaks and bandwidth of these regions.

First part combines two different methods to detect the presence of a peak in the LSF region by comparing these values with experimentally found thresholds. By using this algorithm, correct state of an LSF region can be found with 95% accuracy. Second part of the system obtains location of peaks with weighted mean of LSFs. The weights are calculated by applying exponent values, which are different for each LSF region, to the power spectrum values evaluated at location of LSFs. If 25 Hz error is considered to be an acceptable error, success rate of this algorithm is around 95% for voiced speech. These two parts also prove the strong correlation between displacement of LSFs and speech spectrum. Final part of the system uses the parameters extracted in previous parts to detect spectrum non-stationarity. It is experimentally observed that the system can be tuned by adjusting thresholds for the different requirements of the applications. The thresholds can be set lower values to detect smaller changes to be used in speech segmentation systems, while they can be set higher to detect only abrupt changes.

Finally, it must be stated that the computational complexity of the proposed algorithm is low: LSFs can be extracted in an efficient way described in [41] and the computation of the rest of the parameters can be done without making high complexity arithmetic operations.

In the next two chapters, two applications, using this non-stationarity detection algorithm, is presented. In Chapter 3, an implicit speech segmentation algorithm is discussed. The novel non-stationarity detector is used to find the transient regions of the speech signal in this segmentation system. As the second application, a variable bit-rate speech coder based on mixed excitation linear prediction model is described. This vocoder utilizes a voice activity detector to detect the silence parts of the conversation. The new non-stationarity detector is used in the voice activity detector to provide non-stationary noise immunity to the vocoder.

Chapter 3

Speech Segmentation

Speech segmentation algorithms are necessary for several speech processing applications to overcome various problems like increasing computational complexity according to the nature of the algorithm. They can solve these problems by providing stationary regions in the speech signal.

In a speech recognition system, especially designed for continuous speech recognition, memory and computational complexity of the algorithms increase dramatically with the increasing number of vocabulary size. In order to solve this problem, recognition of sub-word units, like diphone or triphones, followed by a segmentation system may be used. In these systems, segmentation algorithms are used to segment speech signal into desired sub-word units and recognition algorithms are applied to these segmented parts to extract content of the speech signal.

Similarly, speech segmentation can be used in very low bit-rate speech coding systems. In these systems, speech is segmented into stationary regions and these regions are coded separately, sometimes with different algorithms according to the type of the phonological units [47]. Very low bit rates, as low as 150 b/s, can be achieved with segmentation based vocoders [48].

Segmentation methods can be roughly classified into two groups: Implicit segmentation methods and explicit segmentation methods. Implicit segmentation methods split up the utterance into segments without the use of phonetic transcription. These systems define segments as spectrally stable part of the signal. References [5–8] are typical examples for this kind of segmentation algorithms. Explicit segmentation methods split up the incoming utterance into segments that are explicitly defined by phonetic transcription. In general, explicit segmentation methods have the disadvantage that the reference patterns have to be generated before the method can be used. Since implicit methods do not use any reference patterns, explicit methods are expected to perform better. However, such patterns may not fit well to the utterance and may not account all variability occurring in the natural speech. References [9–13] are good examples for the explicit segmentation methods. Furthermore, in [2] both implicit and explicit methods are used to obtain better results. Disadvantage and advantage of implicit and explicit methods are tabulated in Table 3.1.

Table 3.1: Some characteristics of the explicit and implicit segmentation methods.

Implicit Segmentation	Explicit Segmentation
The method does not always give correct number of segments.	The method produces the number of segments given by the phonetic transcription.
The segments are unlabelled.	The segments are labeled in accordance with the phonetic transcription.
The segment boundaries are determined accurately enough for diphone segmentation.	The segment boundaries may be inaccurate due to a possible poor resemblance between reference and test spectrum.

In this chapter, a speech segmentation system based on the proposed spectrum non-stationarity detection algorithm in Chapter 2 is presented. Outline of this chapter is as follows: In Section 3.1, brief description of phonological units is given. In the following section, Section 3.2, the new speech segmentation algorithm is described. Section 3.3 presents experimental results of this algorithm. Conclusion remarks for this chapter are given in Section 3.4.

3.1 Phonological Units

When humans communicate with each other, they use meaningful words which are constructed with concatenation of some basic sound units. These basic sound units are called *phonemes.* However, speech production system of humans does not work only to produce these sounds and hence spoken words are not generally composed of these ideal sounds. Therefore, segments of pronounced words are not usually interchangeable. As an example, if the 'b' used in '*beş*' is concatenated to '*ol*' to produce '*bol*', it sounds disjointed or weird. This phenomenon can be explained as follows: If a phoneme is spoken in isolation, acoustic waveform of that phoneme can be distinguished without any difficulty. However, if they are used in context, boundaries between phonemes are hardly be detected according to the speech articulators. Since vocal tract articulators are formed by human tissue, transition from one phoneme to other one is controlled by muscle movements. Hence, movement of tissues generally slightly modify the production of phonemes. Therefore, associated with each phoneme, there is a collection of *allophones* (variation of phones) that represents acoustic variations on the basic unit. Allophones contain the degree of freedom in production of phonemes, which is not only related with the structure of the unit, but also the position of the unit within the word. As a result, despite the phonemes are defined to be the basic units in speech production, speaker has some degree of freedom in producing of these sounds.

3.1.1 Phonemic and Phonetic Classification

Phonemes can be classified according to the following criteria:

1. Time waveform.
2. Spectral characteristics.
3. Manner of articulation.
4. Type of excitation.
5. The stationarity of phoneme:

- *Continuant*: Vocal tract configuration is fixed during production of phoneme. Examples for this type of phonemes are vowels, fricatives and nasals.
- *Noncontinuant* : Change in vocal tract configuration is occurred during production of phoneme. Examples for this type of phonemes are diphthongs, liquids, glides and stops.

Furthermore, phonemes are generally classified according to the articulatory movement and their acoustical features:

Vowels and Vowel Like Phonemes

1. *Vowels*:
   - Vowels have highest energy among all other phonemes.
   - Duration of vowels can vary form 40 to 400 ms.
   - The variations in cross-sectional areas in the vocal tract determines spectral shape of vowel.
   - Vowel formant characteristics have great variation across different speakers.
   - The length of vocal tract affects location of formants in spectrum.
   - The bandwidths of formants can also characterize vowels.
2. *Diphthongs*: They contain two target vowel formation. Hence, it can also be defined as transition between two vowels, especially transition of their formant structures.
3. *Glides*: They contain transient part of one vowel and their duration are short.
4. *Liquids*: They have similar spectral characteristics with vowels but since vocal tract is more constricted, they are weaker than vowels.

Consonants

1. *Fricatives*: They are produced by excitation of vocal tract with a steady air stream. Turbulence may occur at some points of constructions.

2. *Nasals*: They are produced by glottal waveform exciting an open nasal cavity and closed oral cavity. Due to the nature of excitation source, they resemble vowel, but their energies are weaker due to the limited ability of nasal cavity to radiate sound.

3. *Stops*: They are transient sounds that are produced by building up a pressure behind a total closure somewhere along vocal tract, and suddenly releasing this pressure. After the air pressure is released, there is a brief period of noise like frication occurred due to the sudden turbulence from escaping air. Unvoiced plosives usually possess longer periods of frication than voiced stops. The frication and aspiration is called stop-release. The interval of time leading up to the release during which pressure is built up is called stop-gap.

3.1.2 Characterization of Segments and Boundaries

Characterization of Boundaries

Boundaries are formed due to the change in articulatory movement or spectral formation. These changes may be occurred in different forms:

1. An abrupt change such as termination of or start of voice.

2. Some degree of spectral changes:
   - A variation of energy inside a frequency band.
   - A fluctuating variation of formant locations.
   - A loss of formant structure.

Segment Categories

1. Stationary segments.
2. Short segments.
3. Transient segments:
    - Between two voiced phonemes: Monotonous changes occur among formants.
    - Between a voiced phoneme and unvoiced phoneme : The formant structure and noise are superposed.
    - Between a phoneme and silence : These regions occur at the end of word or before a plosive.

3.2 Speech Segmentation System

In this section, a speech segmentation system based on the spectrum non-stationarity detector proposed in Chapter 2 is described. The proposed segmentation system is of implicit type, i.e. it does not require any phonetical transcription, and success of the algorithm directly depends on the pronunciation of the words which are tried to be segmented. System does not incorporate an end-point detector, but the non-stationarity detector in conjunction with a silence detector may also be used to make accurate detection of the end-points of the utterances. In our system, the non-stationarity detector is used to detect end-points as well.

The system consists of two main parts. The first part is used to detect transient regions by the same method described in Chapter 2. In the second part, a modified version of the same algorithm is applied sample-by-sample to find the exact location of boundaries.

3.2.1 Pre-Processing System

In pre-processing system, speech signal is processed frame-by-frame and the non-stationarity detector is applied to each frame to detect any change in spectrum with one frame delay. To detect minor changes, the thresholds, λ and η, are set to 0.1 and 0.2, respectively. As discussed in Chapter 2, these values are determined experimentally from a speech signals consisting phonetically balanced words. Output of this block provides the transient regions explained in Section 3.1.

An example to the output of this block is given in Figure 3.1. In this experiment, the window length, the frame length and the overlapping time amount is set to 25 ms, 22.5 ms and 11.25 ms, respectively.

Proposed algorithm almost detect all of the non-stationary regions, including voiced-voiced transitions, voiced-unvoiced transitions and silence-voiced transitions. Further test results on the performance of the algorithm is given in Section 3.3.

3.2.2 Boundary Location Estimation

After stationary parts are detected, another measure which is based on the speech variation measures described in Section 2.2 is used to find the location of boundaries. This algorithm is applied to transient regions of the signal in a sample-by-sample basis. For each sample, (2.6) and (2.16) are used to calculate Λ and Γ by taking the center of the frame as the current sample, the l^{th} sample. Furthermore, next Λ and Γ values are also calculated by taking their center as $(l+L)^{th}$ sample where L is the frame length. A new measure, Ω_l, is defined for the l^{th} sample is defined as follows:

$$\Omega_l = \Lambda_l + \Lambda_{l+L} + \Gamma_l + \Gamma_{l+L}. \tag{3.1}$$

The sample, l^*, which maximize Ω_l is selected as the boundary location:

$$l^* = \underset{(k \cdot L-(L-O)<l<(k+1) \cdot L)}{argmax} \{\Omega_l\} \tag{3.2}$$

where O represents the overlapping time amount and k is processed frame number in which transient behaviour exists.

An example of this algorithm applied to the transient regions extracted by the pre-processing system is shown in Figure 3.2.

This algorithm can also be analyzed intuitively: For each sample, a frame, whose center is the current sample, is constructed and the vocal tract filter, corresponding LSFs and required parameters are extracted. This process is also repeated for the samples which are one frame length apart from the current sample in both directions in the time domain. After calculation of the required parameters, Λ and Γ are calculated for both directions. Since value of these parameters reflects amount of change in spectrum, the sample which makes maximum change with respect to both previous and next frames is accepted as the boundary point. As length of transient regions may vary from 8.75 ms to 40 ms, success of the algorithm depends on the selected frame length. If the selected frame length is close to the length of the transient regions, exact point of the boundary location is obtained. Otherwise, large deviations from the exact boundary location may be encountered.

3.3 Simulation Studies

Simulation studies are performed on two signals consisting 50 words which contain balanced phonological units, one for male and one for female speech. In all simulations, the window length and the frame update duration is set to 25 ms and 11.25 ms, respectively. The duration between analyzed frame and comparison frame is varied to test the performance of the system for different tests. Success rates are extracted from the transient regions extracted by the pre-processor system described in Section 3.2, since transient region estimation is more important in segmentation systems. Moreover, finding accurate locations of boundaries are sometimes impossible even by hand.

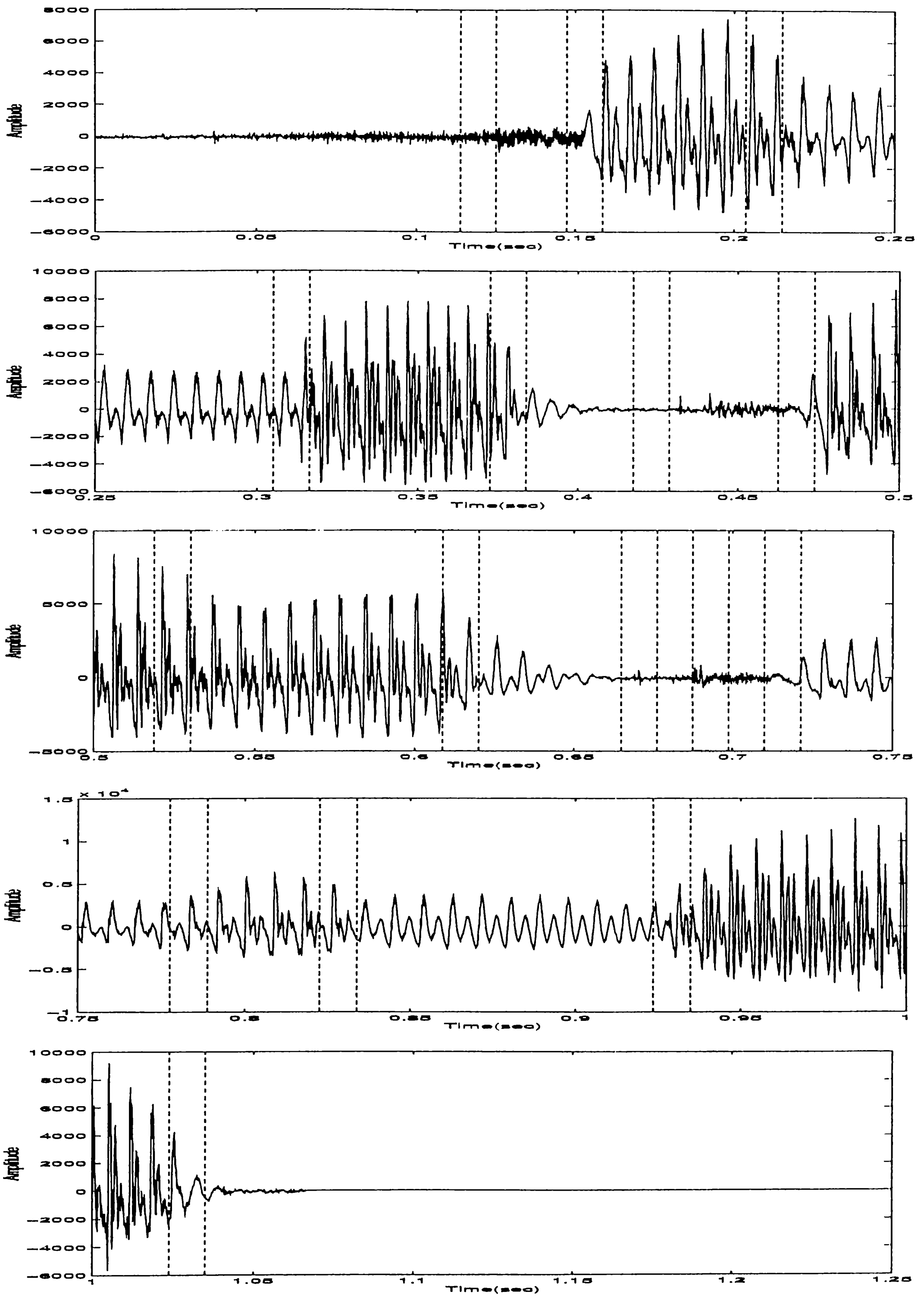


Figure 3.1: Pre-processing system applied to 1.05 second male speech containing words "*Firma tanıtımında*". The regions between two dashed lines are transient regions detected by proposed algorithm.

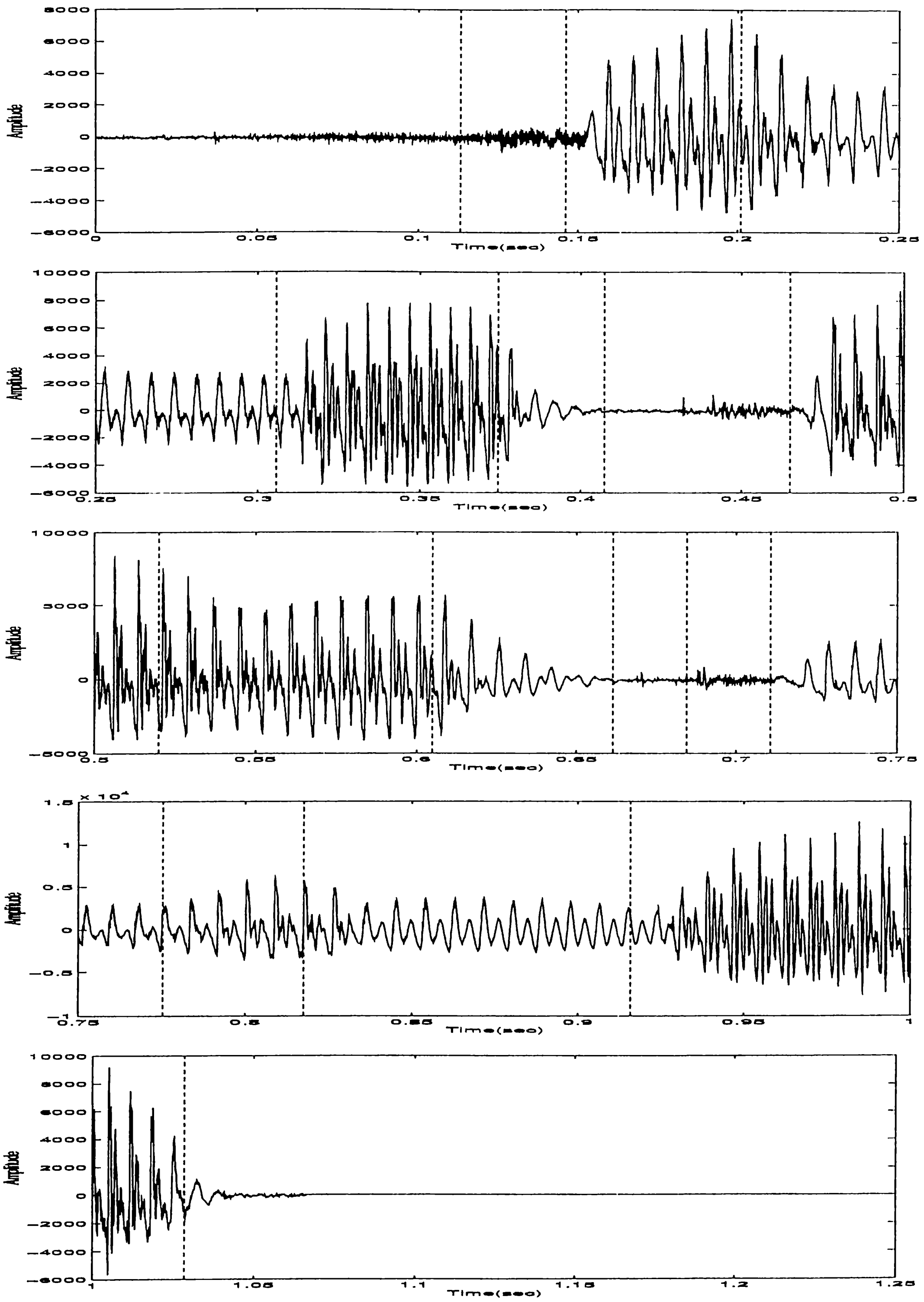


Figure 3.2: Boundary location estimator applied to 1.05 second male speech containing words "*Firma tanıtımında*" after detection of non-stationary regions. The dashed lines show detected boundaries.

Experiments are performed for three different cases:

1. The duration between analyzed frames is set to 22.5 ms.
2. The duration between analyzed frames is set to 11.25 ms.
3. The duration between analyzed frames is set to 16.875 ms.

Results for these three configurations for male and female speech is presented in Table 3.2:

Table 3.2: Success rate about the estimation of the transient regions in the continuous speech signal. Both end-point detection and segmentation within word is performed by pre-processing system of the new algorithm. P_E stands for the percentage of the correct estimated end-points. P_B stands for the percentage of the correct estimated segment boundaries. P_I stands for the percentage of insertions with respect to the whole non-stationary detected regions.

	P_E	P_B	P_I
1^{st} case for male speech	95.00	94.44	24.23
1^{st} case for female speech	90.00	93.51	18.72
2^{nd} case for male speech	76.00	62.03	8.33
2^{nd} case for female speech	59.00	64.81	9.79
3^{rd} case for male speech	85.00	85.18	14.61
3^{rd} case for female speech	80.00	85.18	15.68

The proposed algorithm obtains highest scores when duration between analyzed frames is set to 22.5 ms. This is logical, since any increase in the difference between analyzed frames enables algorithm to catch more transient regions whose transient duration is lower than the difference between analyzed frames. Unfortunately, since thresholds λ and η is selected so that the algorithm detects even small spectrum changes, number of accidentally detected regions are also increased. It can be observed that in the 2^{nd} case, number of insertions are decreased dramatically, since difference between analyzed frames is decreased, but this also yields to large number of misses.

The best solution for this problem is to set the duration between the analyzed frames larger than the largest possible length for a transient region and use another approach like Bradt's Generalized Likelihood Ratio (GLR) to eliminate the insertions and to find accurate locations of segment boundaries [6].

3.4 Summary

In this chapter, a simple speech segmentation algorithm is proposed. The algorithm consists of two main blocks:

1. Pre-processing system operating on the speech signal in a frame-by-frame basis and extracts transient regions.
2. Boundary location estimator operating on the detected transient regions to extract exact boundaries of the phonemes by sample-by-sample processing.

In simulation studies, it is observed that an increase in the duration between analyzed frames provides the proposed algorithm the capability of catching more transient regions. Unfortunately, since the thresholds for speech variation measures are set to small values to catch even minor spectral changes, the algorithm also classifies non-transient regions as transient regions. In order to overcome this problem, more sophisticated algorithms can be used to eliminate these wrong classified regions. Since most of these sophisticated algorithms requires high computational complexity [5,6], these algorithms may be used only in the regions which is flagged as transient with the pre-processing part of our algorithm, whose computational complexity is proved to be low. Furthermore, more parameters like energy of signal and pitch contour can be used to improve the performance of this system with additional computational cost.

Chapter 4

Variable Bit-Rate Mixed Excitation Linear Prediction Vocoder

Compression of telephone-bandwidth speech has been an ongoing area of research for several decades [49, 50]. Especially, in the last several years, with the improving speed and decreasing price of DSP microprocessors, many of the algorithms and coding methods formerly impossible to implement can be realized in real-time. Most of these efforts are focused on the usual telephone-bandwidth which is between 200 Hz and 3.4 kHz.

Speech coding algorithms are classified according to their bit-rates:

- Large bit-rate : Bit-rates larger than 16 kb/s.
- Medium bit-rate : Bit-rates between 8 kb/s and 16 kb/s.
- Low bit-rate : Bit-rates between 8 kb/s and 2.4 kb/s.
- Very low bit-rate : Bit-rates below 2400 b/s.

Analysis-by-synthesis models can work well down to 4.8 kb/s but the quality deteriorates rapidly below this rate. Therefore, usually low bit-rates can be achieved only by parametric representation of speech. The most well-known and studied speech coding algorithm is the LPC-10 vocoder based on the human speech production system, described in Chapter 1.

Although LPC-10 algorithm can encode speech intelligibly at 2.4 kb/s, its quality is usually unacceptable for many applications. After 1993, a competition for a new 2400 b/s vocoder standard to replace old LPC10-E vocoder was started in the USA by U.S. Department of Defense, Digital Voice Processing Consortium (DoD-DDVPC). After long time of testing stages and refinement of the candidate algorithms, Mixed Excitation Linear Predictive coding (MELP) is selected as the new 2400 b/s federal standard in 1997. MELP algorithm is originally developed by Alan McCree and T.P. Barnwell at Georgia Institute of Technology [51].

Variable Bit-Rate (VBR) coding is a special type of multi-mode coding scheme, where each acoustic phonetic class is encoded by different coding algorithm and represented by different amount of bits. VBR coders are particularly useful for voice storage, code-division multiple access (CDMA) wireless networks, and packetized communication systems.

One of the most important part of VBR coders is the voice activity detector (VAD), which is used to detect the presence of speech signal in the channel. In a typical two-way telephone conversation, the voicing activity is generally around 40 percent. Therefore, the average coding rate can be reduced with efficient coding of these silent regions.

In the past decades, several researchers presented different variable rate coding techniques. Although various speech representation techniques are used, most of them are based on CELP coders [52]: The first VR-CELP coder is proposed by Vaseghi [53]. In his work, several versions of different bit-rate CELP coders, allocating different numbers of bits to quantized LPC parameters and excitation signal, are used to code different speech types. Besides this coder, several VR-CELP algorithms are reported in literature which addresses different problems: In Cellario *et al.* work [54], a VR-CELP coder for CDMA application is demonstrated. Gomez *et al.* presents real-time implementation of Federal Standard FS1016 CELP coder which can also switch to higher and lower bitrates according to the distortion in the synthesized speech [55]. In Iacovo and Serena work [56], an embedded variable rate CELP coding technique is presented. In their algorithm, embedded bit-stream is used to provide robustness to packet loss in packetized transmission systems. Lupini *et al.* proposed a VR-CELP coder which selects its coding mode due to the conditions of both input signal and network conditions [57]. In recent studies, McClellan and Gibson presents a VR-CELP algorithm based on subband measures of spectral flatness using entropy functional [58] and Kroon and Recchione implements a low-complexity toll-quality VR-CELP coder for CDMA cellular systems [59]. Majority of these coders use bit-rates varying from 16 kb/s down to 0.8 kb/s.

In addition to VR-CELP coders, several other variable rate coding algorithms based on other speech coding methods are reported: Peng and Cuperman presents a variable rate coder based on lattice low-delay vector excitation coding technique [60]. It is reported that it is possible to obtain good quality speech between 8 and 16 kb/s with this method. Francesco *et al.* presents an algorithm which makes speech segmentation and coding at the same time with reasonable complexity with fast algebraic codes [61]. Their algorithm is reported to achieve an average bit-rate between 5 and 6 kb/s. Wang and Gersho propose a phonetically segmented vector excitation based vocoder which works at 3.0 kb/s average bit-rate whose subjective performance closely matches 4.8 kb/s DoD CELP coder [62]. Another variable-rate subband coder is presented by Shen *et al.* [63]. At 12 kb/s average bit-rate, this coder is reported to produce better quality speech than QCELP coder and better quality music than both QCELP and full-rate GSM coder. In Paksoy *et al.* work [64], a variable rate multimodal speech coder which is based on analysis-by-synthesis method is presented. This coder is reported to achieve 3 kb/s average bit-rate with a quality comparable to full-rate GSM coder. In Yu and Chan work [65], a variable rate coder based on multiband excitation coding algorithm is presented. This coder utilizes different bit allocation for spectral quantization for different frame types. This coder is capable of transmission of good quality speech at average bit-rate of 1.24 kb/s. In addition to these coders, Villette *et al.* presents a high quality split-band LPC vocoder in both fixed rate and variable rate versions [66]. Variable rate coder is reported to achieve an average bit-rate of 1.4 kb/s.

In this chapter, a variable bit-rate MELP vocoder, based on federal standard, is presented. Parameter extraction scheme is the same as the original MELP vocoder, and after detecting the frame type of the speech signal, different number of bits is assigned to different frame types. This encoding scheme reduces bit-rate to approximately 1000 b/s from 2400 b/s without considerable loss of quality. Silence and background noise sections are detected by a novel voice activity detector, in which the non-stationarity detector, proposed in Chapter 2, is used to extract the stationary segments in the signal. Our non-stationarity detector provides non-stationary background noise immunity to the voice activity detector in this vocoder.

Outline of this chapter is as follows: A brief description of MELP vocoder and voice activity detectors are given in Section 4.1 and Section 4.2, respectively. Design of VBR-MELP vocoder and the novel VAD is presented in Section 4.3. Section 4.4 presents performance of the new VBR vocoder.

4.1 Mixed Excitation Linear Prediction Vocoder

Although it is possible to obtain synthetic speech at 2400 b/s with traditional LPC vocoder, its performance is unacceptable for many applications due to low quality. The problems for the traditional LPC vocoder can be summarized as follows:

- Lack of naturalness,
- tonal noise especially for female speakers,
- buzziness especially for male speakers,
- mechanical and tense sound,
- thumps,
- lack of transition control within frames, and
- noise robustness.

The first problems are according to the spectral envelope mismatching. Since the LPC vocoder is based on an oversimplified model of the human voice generation system, this system can not produce more complex sounds, humans generate.

Another problem in the LPC vocoder is the wrong decisions by the voiced/unvoiced detectors. Since neither of the voiced/unvoiced decision algorithms can always find the true state of this switch, wrong decisions leads to deterioration of the quality of the synthesized speech. Unvoiced classification for the uncertain situations increases thumps, where voiced classification of these parts makes synthesized speech sound buzzy.

In order to correct the above problems, more complex models must be used to simulate other properties of human voice generation system. Some of these problems can be eliminated as follows:

- Some phonemes have both voiced and unvoiced regions in the spectrum such as voiced fricatives (e.g. */z/* and */v/*). Therefore, mixture of voiced and unvoiced excitation can produce them more naturally [67]. Fourier transform of a mixed excited utterance can be seen in Figure 4.1. In this figure, spectrum contains both harmonic structures and noise in distinct bands. In addition to this property, mixed excitation removes the voiced/unvoiced switch which decreases the buzziness and thumps in the synthesized speech. Furthermore, mixed excitation improves background noise immunity of the system [68].

- Since tonal noise occurs according to the strong periodicity of the impulse train, this problem can be eliminated by destroying the periodicity of the impulse train [69].

- Pulse shape of the periodic excitation waveform may be changed to glottal excitation shape by spreading the energy of the excitation signal between consecutive pitch periods [68].

- Noise due to the mismatch of the LPC filter spectrum may be compensated by reducing the noise in the valleys of LPC spectrum by a time varying filter [70].

- Vocoders may also try to match the waveform of original sequence in transitional areas by other kind of methods similar to [71].

MELP vocoder implements all of these new features except the last one. Although MELP vocoder has no special solution for transitional regions of speech, it can still produce good quality speech at those regions by using its new features.

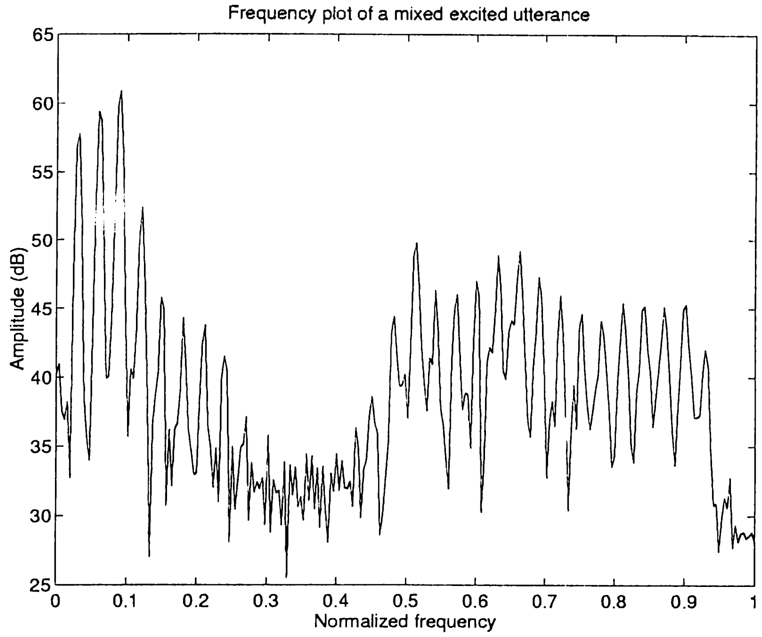


Figure 4.1: Fourier transform of mixed excited phoneme.

4.1.1 Basic Synthesizer

The basic synthesizer of MELP vocoder is shown in Figure 4.2. Both voiced and unvoiced excitation is created separately. The only difference in the generation of the voiced excitation is that periodicity is destroyed if the position jitter is activated and pulse shape changes from frame to frame due to the different Fourier magnitudes. These two excitations are fed into spectral shaping band-pass filters. The band-pass filters are linear phase FIR filters and are complement of each other. Therefore, addition of these two band-pass filters gives unity magnitude in all bands. After this step, two excitations are summed together to form the mixed excitation. It is still possible to simulate voiced/unvoiced switch in the LPC-10 encoder with this excitation generation method.

After this step, the excitation is filtered with the adaptive spectrum enhancement filter. The filtered excitation signal is spectrally shaped by the vocal tract filter and then the resulting signal is filtered with the pulse dispersion filter after energies of the original and synthesized signal are matched. All of these new blocks are described in detail in following subsections.

This system is an enhanced version of the traditional LPC vocoder described in Chapter 1. To increase the performance of this old vocoder, these new blocks are added and the voiced/unvoiced switch is replaced with the mixed excitation model. Note that mixed excitation model is successfully used in multiband excitation coding in frequency domain as well [72].

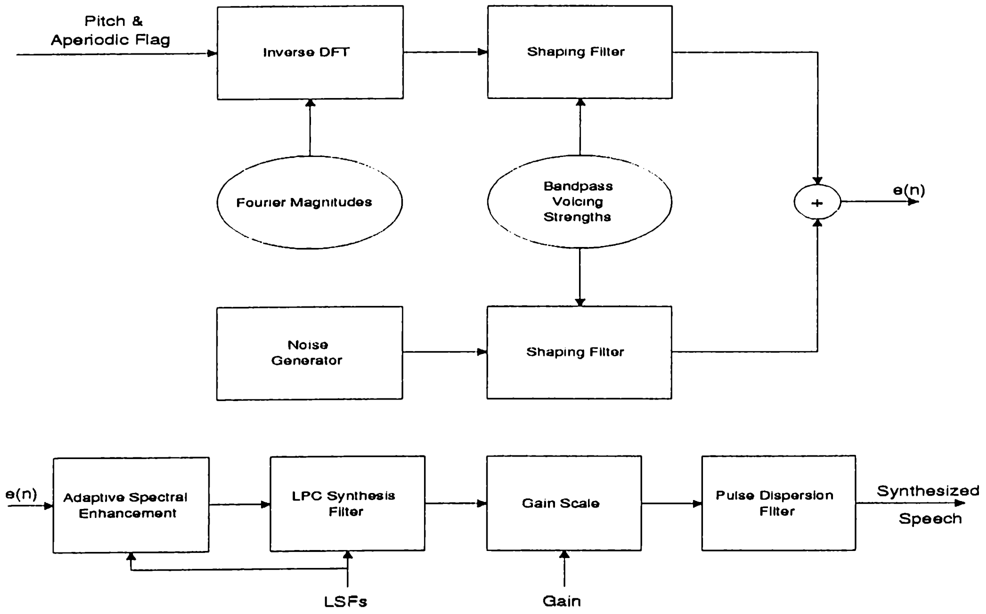


Figure 4.2: Synthesizer of MELP Vocoder.

The new properties of MELP vocoder can be summarized as follows:

- *Mixed Excitation* : Reduces buzz, more accurate in spectrum matching and more robust when operating in noisy environments.
- *Aperiodic Pulses* : Reduces tonal noise.

- *Adaptive Spectral Enhancement Filter* : Emphasize formants and decrease the noise in valleys of the LPC spectrum.
- *Pulse Dispersion Filter* : Modifies pulse shape and spread the energy of excitation signal between consecutive pitch periods. Since it is spectrally flat, it has no influence on unvoiced excitation signal.
- *Fourier Magnitudes* : Reduces mismatch between excitation of original and synthesized speech.

4.1.2 Mixed Excitation

Mixed excitation model is mainly used for synthesizing natural sounding speech which can be possible with more accurate matching of the original speech signal's spectrum with that of the synthesized speech signal's spectrum. In addition to this property, this type of excitation system eliminates the need for an voiced/unvoiced switch and reduces the thumps and buzziness in the synthesized speech due to the wrong voicing decision. Because of the nature of the excitation generation, the new vocoder's performance is also superior to the traditional vocoder in noisy environments.

There are several versions of the mixed excitation generation in literature: In Makhoul *et al.* work [73], periodic impulse train is low-pass filtered and the white noise sequence is filtered with a high-pass filter where the cut-off frequency of these filters are equal. The encoder tries to find the cut-off frequency for these filters to match the original speech with the synthesized one. Since the cut-off frequency is varied in the multiples of 500 Hz, the vocoder does not work well. It is not robust to background noise, as vocoder has only two bands. In Kwon and Goldberg work [67], some degree of voiced and unvoiced excitation signal is mixed for all bands. This approach tries to match the residual signal with the excitation signal of the synthesizer. The problem of this approach is that because the entire spectrum has the same degree of voicing, there is no band specific mixing. First version of the MELP vocoder uses two filters like in [73]. However in this implementation, the zeros of these filter can be varied so that the cut-off frequency is varied continuously and the degree of mixing for the frequency bands can be adjusted. But still this system is not robust to background noise [69].

4.1.3 Aperiodic Pulses

There is short, isolated tones in the synthesized speech especially for female speakers in LPC10 vocoder. This can be removed by adding noise to low frequency region, but the artificial noise results in harsh speech. Another solution is to destroy the periodicity of the impulse train. When the periodicity is destroyed for all voiced frames, strong voiced frames sounds distorted. Therefore, an additional voicing state is added to the MELP vocoder [69]. With this jittery voiced flag, speech may be modeled with aperiodic pulses for voiced excitation at the expense of 1 bit/frame. Aperiodic pulses are generated by varying each pitch period length by a pulse position jitter which is uniformly distributed by $\pm 25\%$ of its original pitch period value. Jittery voicing corresponds to erratic glottal pulses, so it can be detected from either marginal correlation or peakiness in the input speech. Peakiness is defined to be the ratio of RMS power to the average value of the full-wave rectified LPC residual signal. If the value of peakiness is lower than a predetermined value, the sequence is assigned to be aperiodic. This carefully controlled use of aperiodic pulses effectively removes the occasional tones from the synthetic speech without introducing any distortion.

4.1.4 Adaptive Spectral Enhancement

There are various reasons for the usage of adaptive spectral enhancement filter:

1. This block compensates the quantization error of the LPC coefficients. This adaptive filter widens the bandwidth and reduces the peakiness of the formant frequencies in the LPC spectrum. This would increase speech quality because the mismatch in the spectrum is decreased.

2. The noise injected into the system is more audible in the frequency bands which corresponds to the valleys in the spectrum of LPC filter due to the masking property of the human auditory system [74]. Since this spectral enhancement filter de-emphasize those regions, the noise is less audible and this increase the subjective quality of the synthesized speech.

3. This filter helps the band-pass filtered speech to match the natural speech waveform in formant regions. Typical formant resonances do not usually completely decay in the time between pitch pulses in either natural or synthesized speech, but synthetic speech waveform reach a lower valley between peaks than natural waveform do [75]. Adaptive spectral enhancement filter corrects this undesired behavior.

These problems can be eliminated by varying the synthesis pole bandwidth within each pitch period. This can be done by replacing 'z^{-1}' term of LPC filter by '$\alpha \cdot z^{-1}$' where α is smaller than 1. This operation moves the poles of the LPC filter away from the unit circle and weakens the pole resonances. Unfortunately, this filter usually has a low-pass characteristic and therefore makes the speech sound hoarse. Therefore, an all-zero filter which has the same phase angles with the all-pole section but farther away from the unit circle than poles are also added. This would remove the low-pass characteristic of the filter while preserving the emphasis on formant frequencies. The resulting filter becomes:

$$H_s(z) = \frac{1 - P(z/\beta)}{1 - P(z/\alpha)} \qquad 0 < \beta < \alpha < 1 \tag{4.1}$$

Generally α is selected to be 0.8 and β is selected to be 0.5. To reduce the low-pass effect further, a first order FIR filter is added in cascade to the system:

$$H_{tilt} = 1 + \mu \cdot z^{-1} \tag{4.2}$$

where μ is generally selected to be $0.5 \cdot k_1$ where k_1 is the first reflection coefficient. Since this filter produces slightly high-pass spectral tilt, it helps to reduce the low-pass effect.

Excellent review of spectral enhancement filters can be found in [70] and also a frequency domain version of this filter is presented in [76].

4.1.5 Pulse Dispersion Filter

One of the main reasons for the buzziness of the LPC vocoder is the pure impulsive excitation of voiced frames. With this type of excitation, the synthesized speech has higher peaks for the first few samples and it decreases so rapidly that it can not match the original speech especially for the frequency bands which does not have formant resonances [68]. This phenomena can also be explained by the duration of the opening and closure of the vocal cords which can not be in pure impulsive form. In order to eliminate this problem, the shape of the impulsive excitation must be changed. Various researchers tried to find the best pulse shape which gives the highest quality in the synthesized speech [77–79]. The best pulse shape is found to be the one which has no discontinuity and varies approximately 50% of the pitch period [79].

In MELP, this spreading of the pulse through samples are performed by a FIR filter [68], continuously applied to the synthesized speech. Coefficients of this filter is synthesized as follows:

First the DFT of a triangular pulse is computed and the magnitude of the DFT is set to unity. Then, its inverse DFT is evaluated by preserving phase. This filter produces less peaked synthesized speech. Since the synthesized pulse does not decay rapidly, it matches the original speech in a higher degree.

4.1.6 Fourier Series Magnitudes

This extension is first used to enhance the quality of the vocoder in higher bitrate version (4800 b/s). First original speech is filtered with the inverse linear prediction filter. Then, DFT of the resultant residual signal is computed. Although DFT of this signal is expected to be spectrally flat, this is not true in most of the cases. As a remark, if the sequence is considered to be periodic with the period of the size of DFT, the DFT coefficients are exactly the same as the Fourier series expansion coefficients. To increase the quality of speech, magnitudes of the peaks of the harmonics are also transmitted to the decoder. In decoder, these quantized magnitude values are decoded and the impulse train is synthesized by computing the inverse DFT of these coefficients. The magnitudes of harmonics which are not transmitted are assumed to be unity in the decoder.

In MELP vocoder, magnitudes of the only first 10 harmonics are quantized with vector quantization and transmitted to the decoder [80].

4.1.7 Flowchart of the MELP Decoder

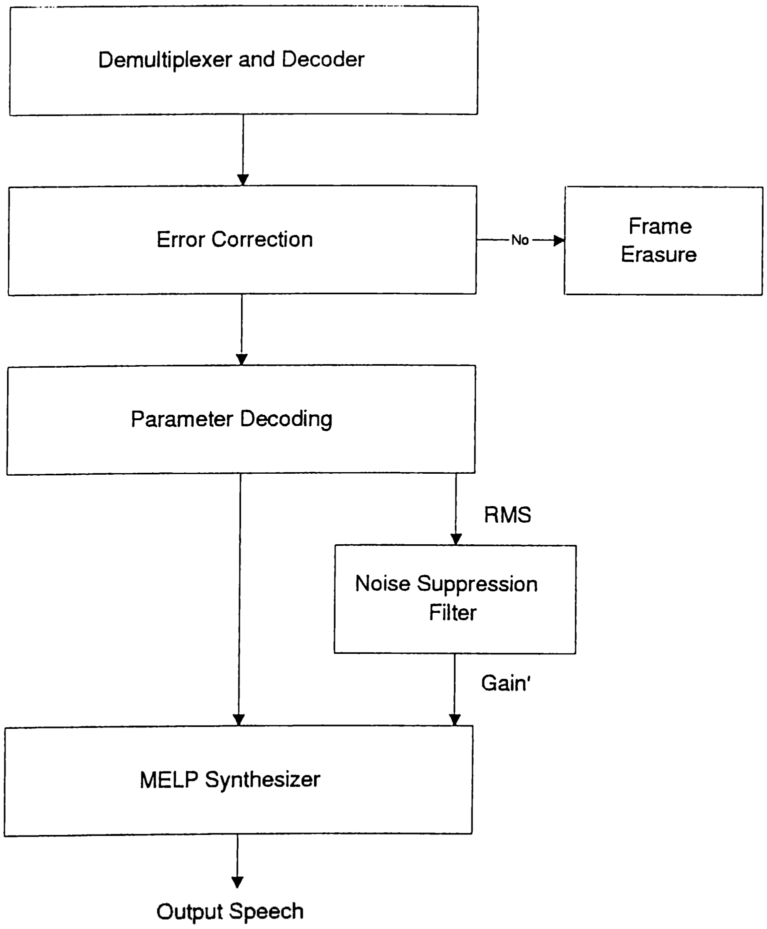


Figure 4.3: Flowchart of the MELP decoder.

4.1.8 Flowchart of the MELP Encoder

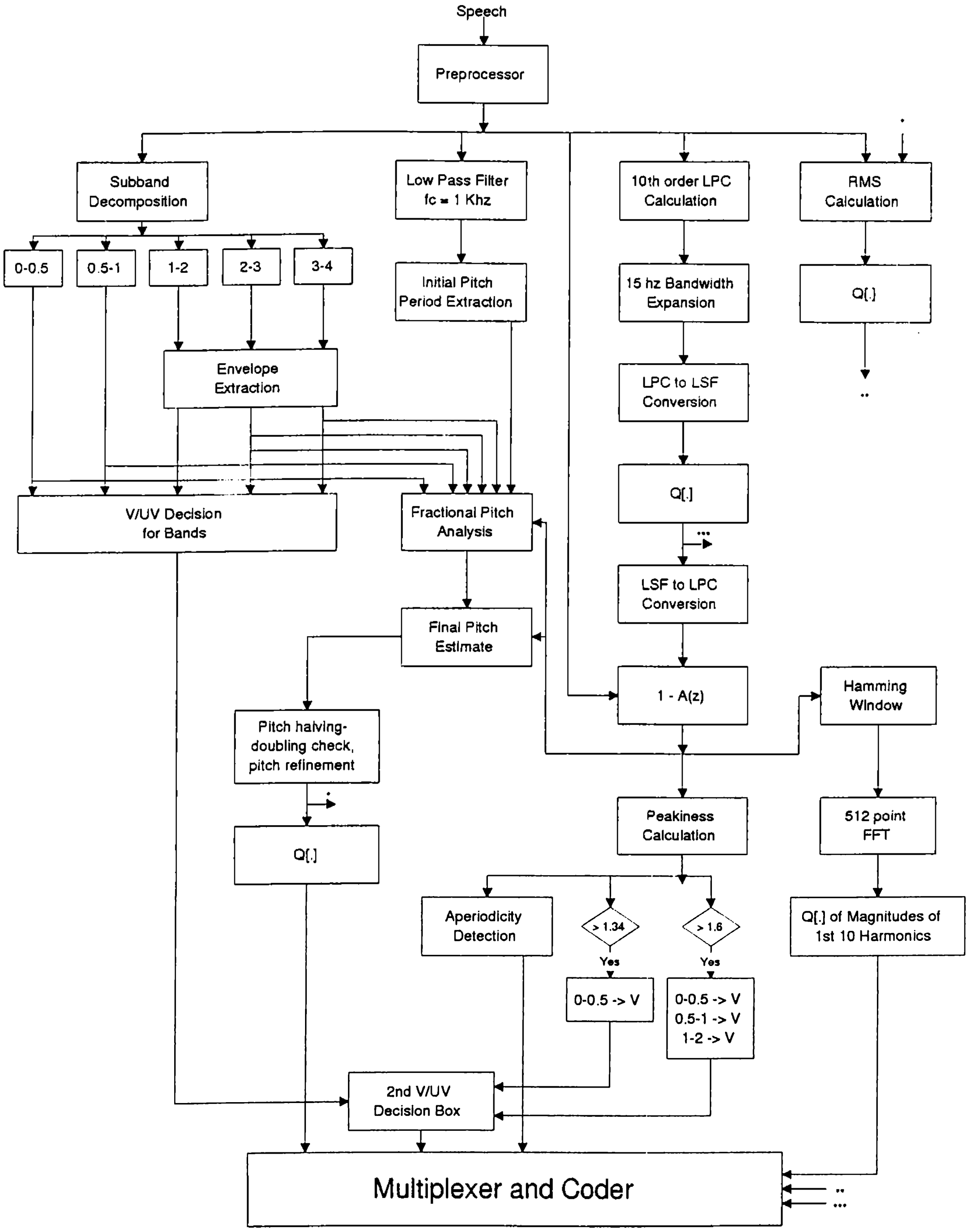


Figure 4.4: Flowchart of the MELP encoder.

4.1.9 Performance Evaluation

Detailed performance analysis for the MELP vocoder was carried out in [81]. In most of these tests, MELP outperforms LPC10 and CSVD and nearly has the same performance with the higher bitrate, 4.8 kb/s, FS1016 CELP standard. Besides in noisy environments, MELP also outperforms federal standard CELP algorithm. More subjective test results are given in Section 4.3 from the results of the subjective tests conducted by TÜBİTAK-BİLTEN / Speech Processing Group.

4.2 Voice Activity Detectors

The Voice Activity Detector (VAD) is the most important part of VBR coders and DTX transmission systems. Basic assumptions behind the design of VAD algorithms can be summarized as follows:

- Speech is a non-stationary signal which changes in 20-30 ms periods.
- Background noise is stationary during much longer periods, i.e. silences and pauses between "*talk-spurts*" in two-way conversations.
- Energy of the speech signal is usually higher than that of background noise.

Based on the assumptions, it is possible to design VAD algorithms to detect silence gaps as well as background noise without speech. If energy of background noise in the environment is very low, a simple algorithm based on signal energy level can be used to distinguish between silence and active speech periods. However, for high energy non-stationary background noise, which is usually encountered in mobile communication systems, more sophisticated algorithms must be developed.

Another problem for a VAD algorithm is the discrimination of low-energy unvoiced sounds like fricatives in high background noise energy level. Since it is hard to detect these parts, a "hang-over time" is used during which VAD algorithm delays its decision to declare silence and continues to observe the signal until it decides that a transition has occurred from active speech to silence. This approach extremely reduces misclassification of weak fricatives.

The accuracy and robustness of VAD determines the quality and capacity of the vocoders. Reliable silence detection is essential. If silence is detected as speech, the capacity is reduced; On the other hand, if speech is detected as silence, "clipping" and other degradations in the synthesized speech are introduced.

Another important consideration for the VBR systems is the synthesis of non-speech segments. In order to preserve naturalness, background noise must be reproduced in some fashion. For stationary noise, it is sufficient to transmit noise characteristic once in the beginning and decoder can produce "comfort noise" during non-speech intervals based on this information. For non-stationary background noise, spectral parameters and gain of the noise must also be transmitted continuously at a very low bit rate.

The following parameters were used to detect the speech activity and silence discrimination in the literature:

- Energy of the first derivative of the signal [82],
- ratio between energies of consecutive frames [82],
- modeling error of the LPC filter [83],
- periodicity and the pitch period [83],
- optimum modeling order of the LPC filter [83],
- LPC cepstrum distance between analyzed frames [83],
- energy and log-energy of the signal [4,84],
- zero-crossing rate [84],
- Teager's energy measure [85], and
- a distance measure based on wavelet analysis [86].

It is experimentally observed that the distance measure given in the last entry works best in low SNR signals [86].

The first systems based on VAD, known as time assignment speech interpolation (TASI), was introduced to increase the capacity of submarine telephone system used in analog telephony [87]. TASI was subsequently replaced with a similar digital system, known as digital speech interpolation (DSI) system.

Several VAD algorithms were presented in the literature: The early works are mostly concentrated on the on-off pattern extraction of speech. These algorithms do not provide any noise robustness. In Brady's work [88], statistical distribution of spurts and gaps were obtained experimentally and a VAD algorithm was designed based on these statistical data. In Yatsuzuka work [87], a speech and voice-band data discriminator for DSI-ADPCM systems was presented. This algorithm provides highly sensitive speech detector with a decision system consisting of a finite state machine, but the system does not have any noise immunity. Another well-known VAD algorithm is used in Pan-European Digital Cellular Mobile Telephone Service [89]. This algorithm is selected by CEPT-GSM to be used in the DTX systems. This VAD is explained in detail in Section 4.3.1. In literature, large number of variations of this algorithm is presented. Most of these variations are reported to increase noise immunity of the system. In Paksoy *et al.* work [90], energy levels in four distinct subbands are compared with corresponding thresholds and if thresholds of any of these subbands are exceeded, frame is declared as speech. Furthermore, spectral flatness at the output of the noise suppression filter is also measured to detect speech activity. Finally, a variable hang-over time is utilized for different noise levels to improve the efficiency of the VAD. In Cellario *et al.* work [54], shorter frames are reported to have better performance in noise. In recent studies, more complicated algorithms are proposed: In Sohn and Sung work, a VAD algorithm, employing a soft decision based noise spectrum adaptation, is presented [91]. This algorithm is reported to have a good noise spectrum tracking property. In Cavallaro *et al.* work [92], a fuzzy logic based VAD algorithm is presented. This algorithm is reported to have better efficiency than the one presented in [89].

4.3 Variable Bit-rate MELP Vocoder

Design of a variable bit-rate vocoder mainly concerns the selection of encoding modes for the optimization of the system in both computational complexity and bit-rate reduction. Besides, the sequence of the mode selection and processing of the encoding algorithms must be decided.

Since our aim is to develop a variable bit-rate vocoder based on MELP without changing the overall system, a simple bit-rate reduction scheme is developed. MELP vocoder separates bit-stream format into two sections as shown in Table 4.1:

Table 4.1: Bit allocation table for fixed bit-rate MELP vocoder.

Parameters	Voiced	Unvoiced
LSFs	25	25
Fourier Magnitudes	8	–
Gain (2 per frame)	8	8
Pitch, Overall Voicing	7	7
Bandpass Voicing	4	–
Aperiodic Flag	1	–
Error Protection	–	13
Sync Bit	1	1
Total Bits / 22.5 ms frames	54	54

Synthesis of unvoiced sections do not require *Fourier magnitudes*, *pitch period*, *bandpass voicing decisions* and *aperiodic flag*. Hence, these bits are used for error protection (13 bits/frame) and transmission of sequence type (7 bits/frame in the place of *pitch period*). Since a header is always transmitted in a variable bit-rate vocoder, the requirement for transmission of pitch period information is eliminated for unvoiced frames. Furthermore, error protection is also removed in our vocoder. In addition, gain is generally stable in unvoiced sections, therefore transmission of the first gain parameter is also redundant. As a result, elimination of transmission of these parameters decreases the required bits from 54 bits/frame to 30 bits/frame for unvoiced frames.

In addition to the bit-rate reduction in unvoiced sections, further bit-rate reduction can be achieved by efficient coding of silence and background noise sections of the input sequence, which covers nearly 40 percent of a typical conversation. For this purpose, the new VBR-MELP coder utilizes a voice activity detector to detect silent regions and makes efficient coding to reduce the average bit-rate.

In following subsections, a novel VAD algorithm for MELP vocoder and final bit allocation table for variable MELP vocoder is presented.

4.3.1 VAD for VBR-MELP Vocoder

The design of proposed voice activity detector is based on the design of two voice activity detectors: First one is designed for DSI-ADPCM systems [87], and utilizes a finite state machine for the detection of silence parts. This detector uses energy of the frame, ratio of energies in consecutive frames and zero-crossing number as the parameter set. Due to the selection of the parameter set, this detector does not have any noise robustness.

The second VAD, on which our algorithm is based, is the VAD used in second phase of GSM 6.10 standard [89]. This detector, utilizes two separate detectors, one for giving voice activity decision by comparing some parameters with predetermined thresholds and one for adapting inverse prediction filter of background noise. In the first detector, the incoming signal is filtered with the inverse prediction filter of the background noise and energy of the resulting signal is computed. Frame is assumed to contain speech signal, when value of this energy is larger than a threshold. Otherwise, that frame is classified to be noise. Furthermore, an 60-100 ms hang-over time is also introduced to avoid clipping of the final parts of the utterances in the speech signal. In the second detector, the distance between the LPC filter computed from averaged autocorrelation values of noise sequence and the LPC filter computed for the current frame is calculated by Itakura likelihood ratio. Frame is assumed to be stationary, while this result is smaller than a pre-determined threshold. Furthermore, periodicity of the signal is also controlled. If input frame is classified as non-stationary and does not contain any periodic structure, frame is assumed to be initial point of noise sequence. Besides, before updating the coefficients of the prediction filter of the noise, this system waits for 8 frames, to ensure the continuity of the noise sequence. If all these conditions are met, coefficients of the prediction filter of the noise used in the first detector is updated. The illustration of this VAD can be seen in Figure 4.5. The computational complexity of this VAD is low. The parameters required for the algorithm is already extracted within RPE-LTP vocoder and the only computation performed within this system is the computations for the decision system.

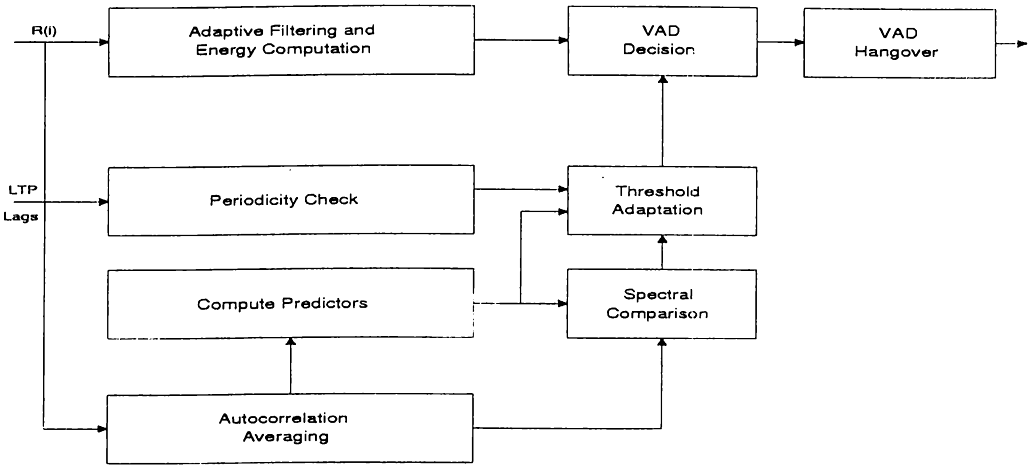


Figure 4.5: The voice activity detector for the pan-european digital cellular mobile telephone service.

MELP algorithm extracts following parameters within the encoder:

- Pitch period,
- bandpass voicing strengths,
- LSFs, and
- two gain calculation for the first and the second half of the frame.

In order to decrease computational complexity, these parameters are used directly within our VAD.

The new proposed VAD uses the distance measure, D_k, defined as follows:

$$D_k = 10 \cdot \log \left[\frac{1}{L} \sum_{l=1}^{L} \frac{E_l^k}{\sigma_l^2} \right] \tag{4.3}$$

This measure is based on the logarithm of the ratio of the signal energies in subbands to the variance of noise in the corresponding bands, which is the modified version of the one used in [86]. σ_l is the estimated variance of the background noise in the l^{th} subband, and E_l^k is the energy parameter of the k^{th} frame for the l^{th} subband over a time window and computed as follows:

$$E_l^k = \frac{1}{N_l}\sum_{n=1}^{N_l}(s_l(n))^2 \qquad\qquad l = 1, 2, \cdots, L \tag{4.4}$$

where N_l is the number of samples in the l^{th} subband. In our system, there is no decimation in subband decomposition. N_l is equal to 90 samples corresponding to half of an 22.5 ms MELP frame. As signal is decomposed into 5 subbands, L is selected to be 5.

The original method is reported to be successful in end-point detection in noisy environments. However, end-point detection system described in [86] assumes that the initial few frames are always noise and the noise variances in the subbands are extracted from these frames. Unfortunately, in a typical telephone conversation, the first few frames may contain speech information. Therefore, it is impossible to make this kind of assumption in our system. To overcome this problem, the required variances of the noise in the subbands are extracted by a similar method described in GSM-VAD.

The diagram of the new VAD is shown in Figure 4.6. The system has two main parts labeled as *VAD1* and *VAD2*:

VAD1 is used to detect the presence of the speech signal. First, energy in 5 subbands are extracted for the first and the second half of the frame and D_k is computed twice for the two sub-frames to obtain D_{k_1} and D_{k_2}. Besides, an initial silence detector is used to detect inaudible signal by comparing the gain values with an experimentally derived threshold. Furthermore, this detector also provides echo-suppression in some degree when no background noise is present: Voiced sections and information tones always have high energies. If a strong periodic structure is detected with a gain smaller than a second threshold, that frame is assumed to belong to a part of an echo signal. The output of this initial silence detector and D_k values are used by the decision system to make final decision about the voicing state. The decision box contains a finite state machine, which consists of four different states about silence detection including 'hang-over' state. States are updated twice in a 22.5 ms frame, one for the value of D_{k_1} and one for the value of D_{k_2}.

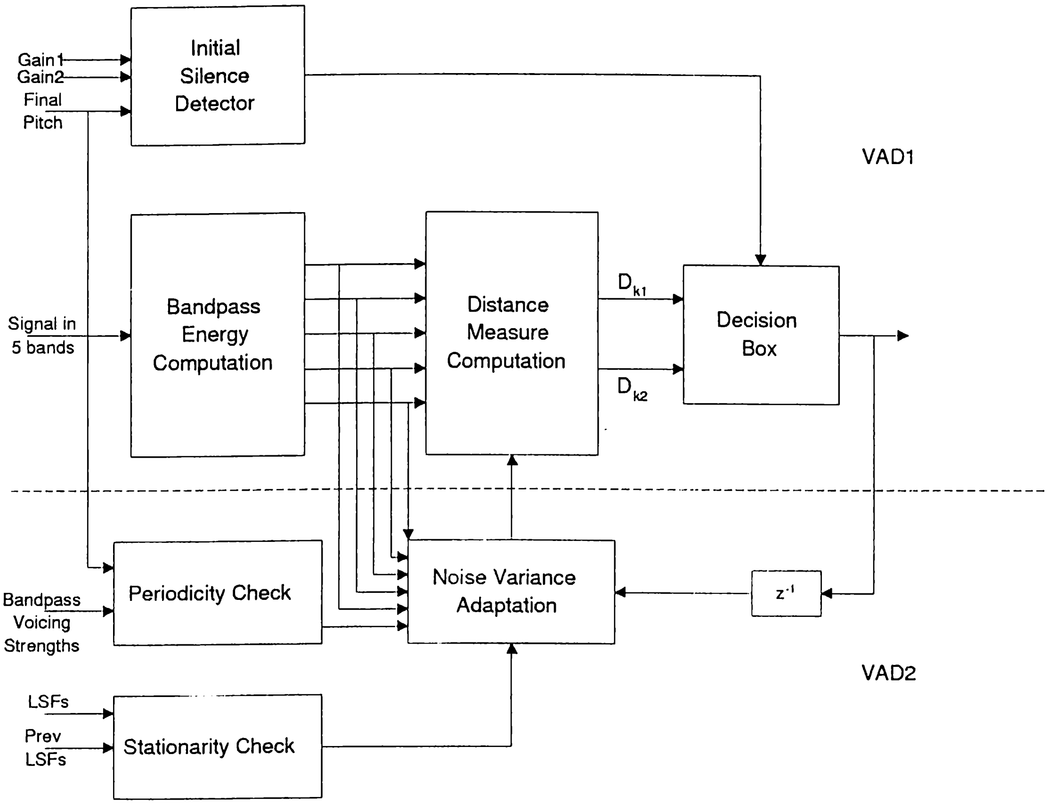


Figure 4.6: Voice activity detector for VBR-MELP Vocoder.

VAD2 is used in parallel with *VAD1* to update the variances of noise in 5 subbands. The adaptation is performed by *Noise Variance Adaptation* block, which takes the required parameters from *Stationarity Check* block, *Periodicity Check* block and the decision of *VAD1* for the previous frame. Since proposed non-stationarity detector has one frame delay, *VAD2* also has one frame delay. The frame length of this block is 180 sample similar to the MELP vocoder.

Details of the sub-blocks are described in following subsections:

Initial Silence Detector

This block is used to detect inaudible frames prior to the application of the rest of the voice activity detector to the current frame. Furthermore, it uses a simple echo suppression logic in the detection of the frames containing quasi-periodic signal with very low energy level, never encountered in voiced speech.

This block requires the gain values and the final estimated pitch period calculated in the MELP encoder. The gain values are compared with two gain thresholds, T_{GL} and T_{GH}. If both gain values are lower than T_{GL}, the analyzed frame is declared as silence. If any of them exceeds T_{GH}, classification of the state of the frame is left to *Speech/Silence Decision Block*. If neither conditions are met, voicing strength measure of extracted pitch period is controlled [17]. If periodicity is detected, frame is classified as echo of the other channel. Otherwise, classification of the state of the frame is left to *Speech/Silence Decision Block*. Flowchart of the algorithm is plotted in Figure 4.7.

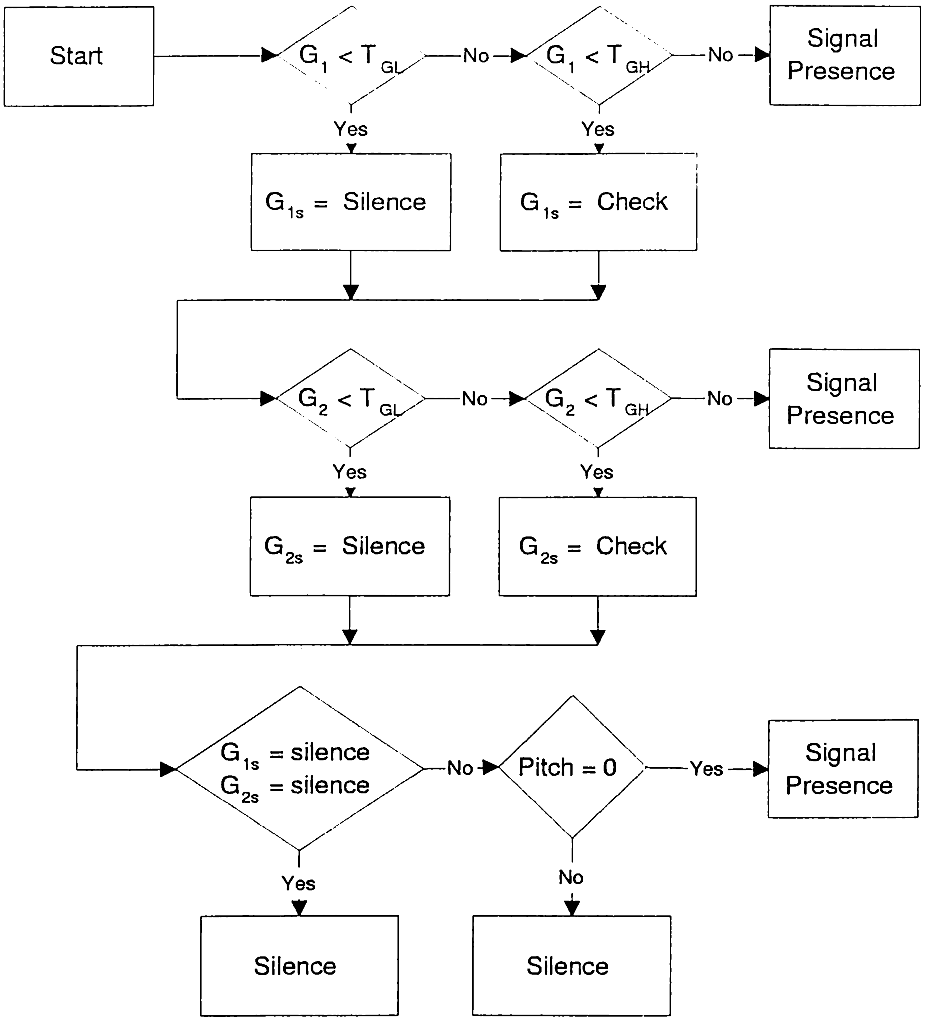


Figure 4.7: Flowchart of initial silence detector.

The thresholds, T_{GL} and T_{GH}, are found experimentally from a 2 minute conversation[1], including echoes from other channel. G_{t1} and G_{t2} are set to 38.0 dB and 48.0 dB, respectively.

Bandpass Energy Computation

This block computes the energies of bandpass signals, E_k^l, for the k^{th} frame which are filtered in encoder with sixth order Butterworth filters, whose passband regions are between following regions: [0 - 500 Hz], [500 - 1000 Hz], [1000 - 2000 Hz], [2000 - 3000 Hz] and [3000 - 4000 Hz].

Energies are computed twice per frame, one for the first half of frame, $E_{k_1}^l$, and one for the second half, $E_{k_2}^l$. These values are used in distance measure calculation, D_k, and *Noise Variance Adaptation* block.

Distance Measure Calculation

In this block, D_k defined in (4.3) is computed twice for both of the first and the second half of the frame. The variances, σ_l, are set to 1.0 for all bands prior to first update in order to classify all frames as speech detected frames. Computed values, D_{k_1} and D_{k_2}, are used in the decision block.

Speech/Silence Decision Block

In this block, D_{k_n} and the decision of the initial silence detector is used to discriminate silence frames in the conversation. For this purpose, a four states finite state machine is developed:

[1] This conversation is taken from `'Switchboard Corpus - Recorded Telephone Conversations'` database collected by Texas Instruments

1. *Silence* (SI): These regions contain only background noise.
2. *Primary Detection of Signal* (PD): These regions are primary detection for signal which may contain information. If system stays in this stage for longer than a pre-determined duration, '*Speech Enable*' state is activated.
3. *Speech Enable* (SE): These regions contain speech signal.
4. *Hang-over Period* (HO): Silence is detected in these regions, however, to eliminate misclassification of the weak fricatives and the final nasals at the end of the speech, a hang-over period is inserted in the system.

State transitions are performed due to the values of two coefficients:

- PDF_k : Primary detection of speech for the k^{th} frame. Its value is set to 0, if silence is declared in initial silence detector or D_{k_n} is below D_{t1}. Otherwise, it is set to 1.
- SDF_k : Definite presence of speech for the k^{th} frame. Its value is set to 1, if D_{k_n} exceeds D_{t2}. Otherwise, it is set to 0.

Setting of these coefficients can also be summarized as follows:

Table 4.2: Setting of the coefficients.

$PDF_k = 0$	Silence is declared in the initial silence detector or $D_{k_n} < D_{t1}$
$PDF_k = 1$	$D_{t1} < D_{k_n} < D_{t2}$
$PDF_k = 1$, $SDF_k = 1$	$D_{t2} < D_{k_n}$

Values of D_{t1} and D_{t2} are obtained experimentally. These values are selected such that clipping of speech and misclassification of silence frames are minimized. The system is found to be optimum, when D_{t1} and D_{t2} are set to 5.0 and 10.0, respectively.

The state transition diagram is given in Figure 4.8. Transitions from *PD* to *SE* and *HO* to *SI* requires some past information, i.e. memory. To go from *PD* to *SE*, *PDF* must be set to 1 for 20 half-frames, that corresponds to a wait state of 225 ms. Transfer from *HO* to *SI* in general requires 45 ms in this system. This relatively short hang-over period is due to the robustness of the system to the background noise. Furthermore, if the *talk-spurt* duration is shorter than 56.25 ms, this period is also reduced to 22.5 ms. In simulation studies, it is observed that only *stop-gaps* are missed with this approach.

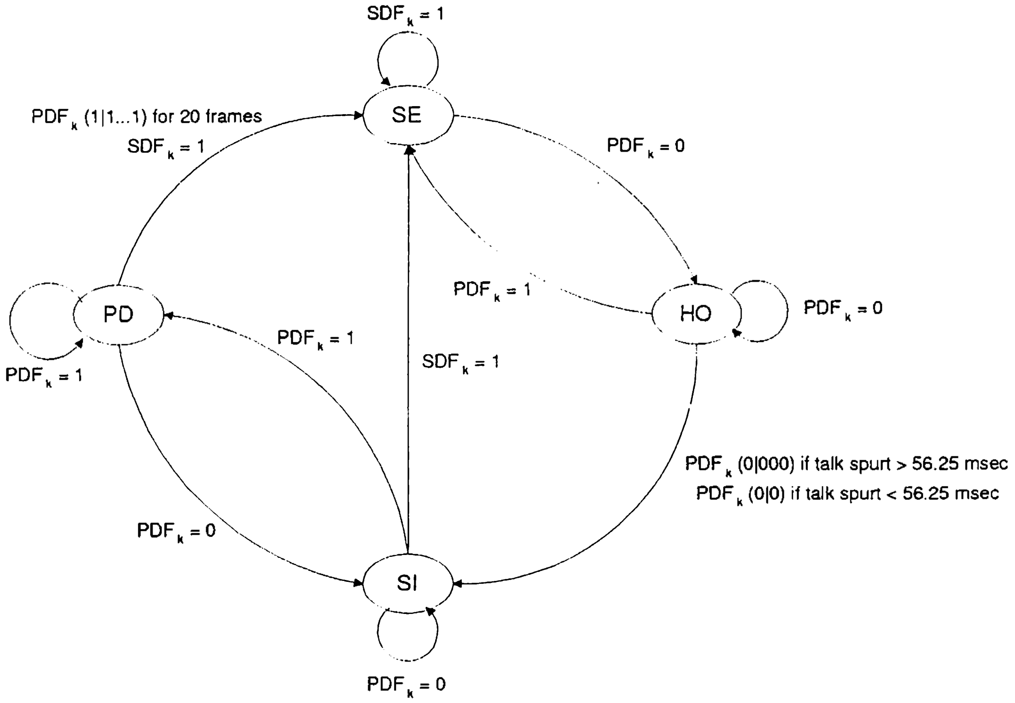


Figure 4.8: State transition diagram of the decision box. SI stands for silence state. PD stands for primary detection state. SE stands for speech detected frames. HO stands for hangover state.

Periodicity Control

This block obtains state of periodicity by analyzing the bandpass voicing strength of bands and voicing strength measure of the estimated pitch period in the encoder [17]. The flowchart of the algorithm is shown in Figure 4.9.

The flags *vbp1*, *vbp2*, *vbp3* and *vbp4* are bandpass voicing strengths for the first 4 bands. Only the first bandpass voicing strength has a fractional value,

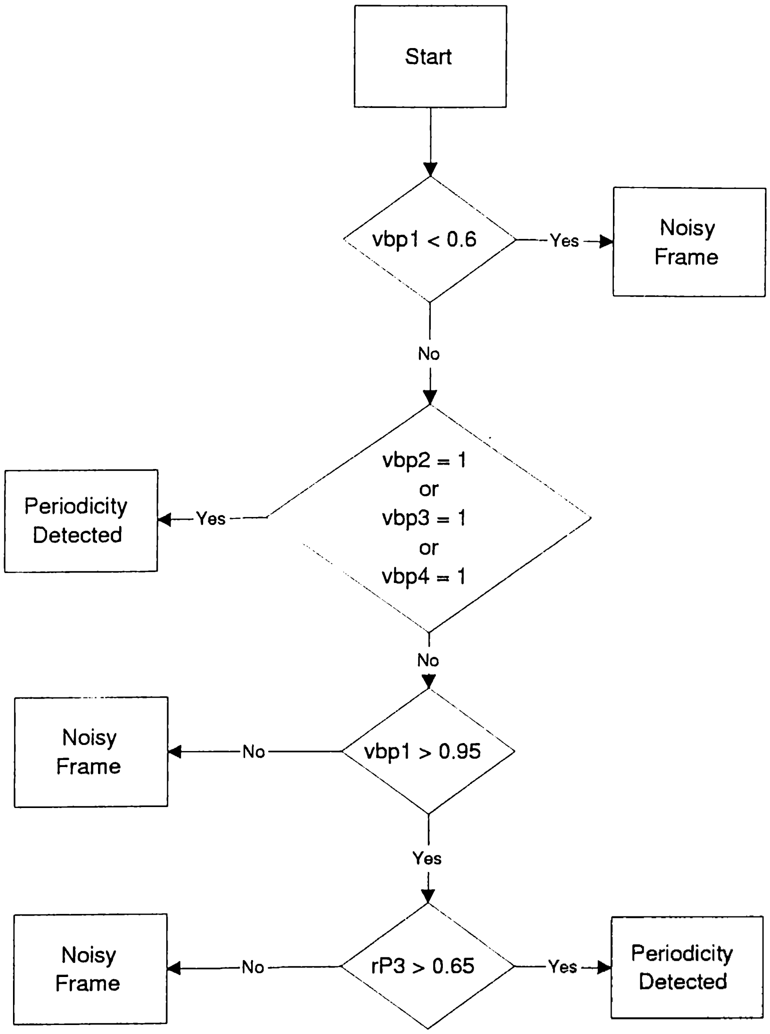


Figure 4.9: Flowchart of periodicity detector.

between 0 and 1. Other ones are set to either 0 or 1. The coefficient, *rP3*, stands for voicing strength measure of final estimated pitch. For strong voiced frames, the value of this parameter exceeds 0.65.

Stationarity Control

This block uses the non-stationarity detector algorithm described in Chapter 2. Thresholds, λ and η, are both set to 1.0, in order to detect only abrupt changes. Output of this block has one frame delay.

Noise Variance Adaptation

This block makes adaptation for the variances of noise in subbands, when the statistics of the background noise is changed. Before adaptation takes place, following conditions must be met:

1. Signal must be stationary for a period of time longer than $S_x \cdot 22.5$ ms. S_x is the number of frames for the system to wait before adaptation takes place in which signal is stationary.
2. Signal must not be periodic. Since information tones has long duration, they may be classified as long stationary regions. These regions must not be included in noise adaptation.
3. Decision state of final decision box in *VAD1* must be same for previous two frames.

If these conditions are met, variances of noise in five subbands are calculated and then continuously averaged in every frame as follows:

$$\sigma_l^2 = \frac{\sigma_l^{2'} \cdot N_a + E_l^k}{N_a + 1} \tag{4.5}$$

where N_a are the number of stationary frames after the first frame of adaptation and $\sigma_l^{2'}$ is the variance of noise in the previous frame. This averaging is repeated until at least one of the three conditions written above is violated.

Value of S_x changes from 6 to 20. It is observed that 8 is a reasonable value, because practically none of the unvoiced phonemes last longer than 180 ms.

Flowchart of this block is illustrated in Figure 4.10

4.3.2 Bit Allocation for VBR-MELP Vocoder

In this novel variable rate MELP coder, a two-bit header is used to classify frame type: Voiced, unvoiced and silence/noise. For the silence/noise frames,

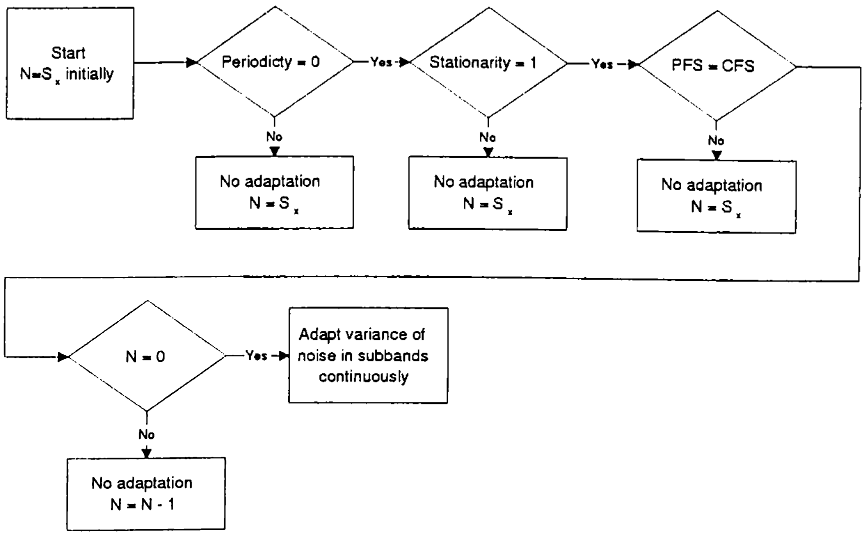


Figure 4.10: Flowchart of noise variance adaptation block. PFS and CFS stand for the state of the previous frame and current frame, respectively.

the parameters of the first frame of the silence regions are transmitted as if it is an unvoiced frame and these parameters are repeatedly used until a header showing different frame type other than silence/noise is encountered. Table 4.3 shows final bit allocation.

Table 4.3: Bit allocation table for variable bit-rate MELP vocoder.

Parameters	**Voiced**	**Unvoiced**	**Silence/Noise**
Header	2	2	2
LSFs	25	25	–
Fourier Magnitudes	8	–	–
Gain (2 per frame)	8	5	–
Pitch, Overall Voicing	7	–	–
Bandpass Voicing	4	–	–
Aperiodic Flag	1	–	–
Total Bits / 22.5 ms frames	55	32	2

4.4 Simulation Studies

As a part of a sponsored project, a MELP vocoder, slightly modified version of the federal standard, is implemented in both floating point and fixed point with

C programming language in TÜBİTAK-BİLTEN / Speech Processing Laboratory by my colleagues and me. Furthermore, fixed point C source codes are ported to TMS320C54x DSP architecture and in order to obtain highest efficiency, the entire code is hand optimized in assembler language. Both of the encoder and decoder of the MELP vocoder is bitstream compatible with the federal standard.

At the end of the project, subjective listening tests are conducted to make comparison between ACELP, a CELP based vocoder whose quality is better than FS1016 federal standard CELP coder, LPC-10 and MELP algorithms. Tests are evaluated using the speech signals taken from two native Turkish speaker, one for male speech and one for female speech. Two types of tests are conducted:

1. Diagnostic Rhyme Test (DRT)
2. Mean Opinion Score (MOS)

The goal of these tests are to evaluate the intelligibility and quality of the output of the synthesizer. Final data are obtained by evaluation of the test results from 20 subjects.

Noise robustness tests are performed as follows: A Volvo340 car noise[2], driven on a rainy asphalt road is mixed with the clean speech. SNR value of the resultant signal is approximately 10 dB. The resultant signals are used as the noisy test data in the following tests.

4.4.1 Diagnostic Rhyme Test

This test was carried out with 50 rhyming word pairs. Subjects are asked to find out which word of the pair is spoken. The classification of the word pairs are as follows:

[2] Institute for Perception, TNO, The Netherlands

1. *Voiced/Unvoiced Difference*: These sounds have only three letters. Word pairs have different utterances only in the beginning of the word. In these tests, subjects are asked to discriminate the letter pairs, b-p, g-k, v-f, j-ş and d-t.

2. *Nasality Difference*: These sounds have again only three letters. Word pairs have different utterances only in the beginning of the word. In these tests, subjects are asked to discriminate the letter pairs, m-b and n-d.

3. *Sustained/Interrupted Difference*: In these tests, words which have similar sounds, but different meanings are used. In this case, one word of the pair lasts longer to read.

4. *Sibilated/Unsibilated Difference*: In these tests, subjects are asked to discriminate the letter pairs, z-t, ç-k and j-g.

In this test, following results are obtained:

Table 4.4: DRT scores of MELP and ACELP vocoders. *WD* stands for wrong decision.

Male Speech	ACELP	MELP
50 word pairs in clean speech	2 WD	3 WD
50 word pairs in noisy speech	29 WD	9 WD
Female Speech	ACELP	MELP
50 word pairs in clean speech	7 WD	8 WD
50 word pairs in noisy speech	57 WD	33 WD

From these results, ACELP vocoder is only 0.2% more intelligible then MELP vocoder in clean speech. However, in noisy speech, MELP vocoder clearly outperforms ACELP: MELP vocoder is 4% and 4.8% more intelligible for male speech and female speech, respectively.

4.4.2 Mean Opinion Score

This test was carried out with ten short *phonetically balanced* sentences. In test, subjects are asked to grade the quality of the synthesized speech by giving marks between 1 and 5. 5 indicates *excellent* and 1 indicates *unacceptable.* Subjects gave grades for the output all three vocoders for male and female speech.

Results of these tests are as follows:

Table 4.5: MOS scores of MELP, ACELP and LPC-10 vocoders.

Male Speech	ACELP	MELP	LPC-10
Clean speech	4.14	4.18	2.36
Noisy speech	1.72	4.33	2.99
Female Speech	ACELP	MELP	LPC-10
Clean speech	3.76	3.64	2.68
Noisy speech	1.90	3.67	2.98

From these tables, it is observed that MELP vocoder has a similar quality with ACELP vocoder for clean speech. However, for noisy speech MELP outperforms ACELP vocoder. It is observed that output of the MELP vocoder is usually preferred over both ACELP and LPC-10 vocoder in noisy environments.

4.4.3 Performance of VBR-MELP Vocoder

Making subjective tests for vocoders requires considerable amount of work and since these tests are already conducted for fixed-rate version, tests are conducted only to obtain the performance of the voice activity detector for various SNR levels. In these tests, SNR values are calculated only from the portions including speech signal. Percentage of clipped regions and misclassified noise regions are tabulated in Table 4.6. Test sequence is obtained from a 50 second telephone conversation[3]. 57 percent of the conversation consists of only background noise. The source of noise is same as the one used in the evaluation of subjective tests of fixed-rate fixed point MELP vocoder.

[3] This conversation is taken from 'Switchboard Corpus - Recorded Telephone Conversations.' database collected by Texas Instruments

Table 4.6: Performance of proposed VAD in various SNR levels for male speech. P_{cl} stands for the percentage of clipped regions with respect to the overall speech sections. P_{ms} stands for the percentage of the missed regions with respect to background noise sections.

SNR level(dB)	P_{cl}	P_{ms}	Avg. bit-rate
∞	3.84	3.76	977.83 bps
30	2.29	3.03	940.05 bps
20	2.35	13.23	1076.00 bps
15	2.41	20.36	1088.50 bps
10	1.83	10.81	1012.00 bps
5	9.92	19.09	998.00 bps

From these experiments, it is observed that our VAD works without any problem when the SNR of the sequence is higher than 10 dB. The clipped regions in these levels are occurred according to a long laugh, in which the detector assumes these regions as noise and equate noise variances to the energy of the speech in that region. Therefore, some utterances are missed due to this wrong adaptation. With the beginning of background noise, system adapts thresholds again and recovers itself. The misclassified silence regions are mostly according to the hang-over period and little energy variations in the background noise. Finally, it is observed that nearly in all cases, average bit-rate is around 1000 b/s which makes further 1 : 2.4 compression over fixed-rate MELP vocoder without a considerable loss of quality.

4.5 Summary

In this chapter, a new variable bit-rate MELP vocoder based on the fixed bit-rate version is presented. In order to decrease the bit-rate, the unvoiced frames and background noise are encoded with fewer bits by eliminating the parameters which are not required in the synthesizing of these frames.

In order to discriminate speech signal from background noise, a novel voice activity detector which is robust to both stationary and non-stationary background noise is introduced. Since the VAD uses the parameters already extracted in MELP encoder, its computational complexity is low. In the simulations, it is observed that the performance of the VAD is high even if the SNR decreases down to 10 dB. Finally, with this new VBR-MELP vocoder, the average bit-rate is decreased from 2400 b/s to 1000 b/s in a typical telephone conversation.

As a final remark, since only three frame types are present in this system, a fourth one can also be defined to encode different type of speech signal, like onsets in the beginning of the utterances to increase the naturalness of the synthesized speech.

Chapter 5

Conclusion

In this thesis, a new frame-based speech spectrum non-stationarity detection algorithm based on line spectrum frequencies is developed. The proposed algorithm is used successfully in a speech segmentation system and a voice activity detector, specifically designed to work in a MELP vocoder.

The proposed algorithm works in three steps: In the first step, the system decides whether a formant is present or not between two consecutive LSFs by analyzing the angular difference between these LSFs and using the logarithmic energies of the spectrum at LSF locations. By applying the decision system described in Section 2.1, the regions which contains formants are detected. The state of the regions are calculated with 95% accuracy. In the second step, exact locations of the peaks are calculated with energy weighted mean of the LSFs, forming that region. To obtain desired accuracy, spectral values at location of the LSFs are calculated by evaluating prediction filter on the unit circle at LSF locations. Then, an exponent value, which is different for each LSF pair and obtained experimentally, is applied to these values. By this method, approximately 95% of the location of the formants are estimated within 25 Hz error range. In the third step, two speech variation measures based on the estimated peak locations and the difference between consecutive LSFs are used to detect speech spectrum non-stationarity. It is observed that the success rate of the algorithm depends on the time difference between consecutive frames and the thresholds, compared with the computed values of the proposed speech variation measures for analyzed frames. These thresholds can be adjusted to change the sensitivity of the algorithm to the spectrum changes.

The proposed algorithm is first used in an implicit-type speech segmentation system. The system does not have any prior information about the input speech signal, hence does not make any assumptions on the input signal. The algorithm tries to locate the boundaries of the phonemes in two steps. In the first step, the non-stationary regions are extracted in a frame-by-frame basis, and in the second step, a modified version of the same measures proposed in Chapter 2 is used to detect exact boundaries of the phonemes. Success rate of this algorithm directly depends on the time difference between consecutive frames, because transient regions in the speech signal may vary form 8.75 ms to 40 ms. Unfortunately, an increase in the time difference between consecutive frames yields to large number of insertions, which must be eliminated by a more complex method. The main advantage of this system, especially of the first step, is its low computational complexity, which makes it possible to use it in the front end of a speech recognition system.

In addition to the speech segmentation system, the proposed non-stationarity detection algorithm is used in a Voice Activity Detector (VAD) of a variable bit-rate MELP vocoder. This new VAD consists of two parts: The first part extracts the energies of the signal in five bands and compute distance measure based on the variance of the noise in the subbands and the energy of the signal in analyzed frame. These values are compared with experimentally derived thresholds to find state of the frame. The decision system is based on a finite state machine, which also includes hang-over period. Prior to the calculations in this first part, a simple silence detector is utilized to detect the inaudible frames and echo signals in some degree. The second part of the system updates the variances of the noise in subbands by using the proposed non-stationarity detector. Long sequences of background noise are searched in the conversation and the variances of noise in subbands are updated from these portions. Nearly all of the parameters used in this VAD is computed in the MELP encoder. Hence, computational complexity of the detector is low. In simulation studies, it is found that proposed VAD has excellent performance until the input signal SNR decreases below 10 dB level. Below this level, clipping becomes a severe problem.

The selection of the thresholds for the proposed speech variation measures and length of the duration between analyzed frames used in the speech segmentation algorithm can be made adaptive to obtain better results as a future work. This will increase computational complexity of the current method however we believe that number of insertions and missed boundaries will be decreased.

APPENDIX A

Threshold Extraction for Elimination of Misclassified LSF Regions

As discussed in Chapter 2, elimination of misclassified regions are performed by comparing bandwidth and ϵ_i defined by (2.3) with some fixed thresholds. If these thresholds are exceeded, state of the regions are changed. Five types of thresholds are extracted from the speech database:

1. T_i : Threshold for the bandwidth for the i^{th} LSF region. If bandwidth of LSF regions which are assigned to contain peaks with both methods exceeds this threshold, the state of these LSF regions are changed.

2. γ_i : Lower threshold for the bandwidth for bandwidth based method for the i^{th} LSF region. If a LSF region is assigned to have peak with only bandwidth based method and if bandwidth of this LSF region is smaller than this threshold, this region is assigned to contain peak in it.

3. α_i : Lower threshold for the bandwidth for energy based method for the i^{th} LSF region. If a LSF region is assigned to have peak with only energy

based method and if bandwidth of this LSF region is smaller than this threshold, this LSF region is assigned to contain peak in it.

4. β_i : Higher threshold for ϵ_i for bandwidth based method for the i^{th} LSF region. If a LSF region is assigned to have peak with only bandwidth based method and if ϵ_i of this LSF region is larger than this threshold, this LSF region is assigned to contain peak in it.

5. ζ_i : Higher threshold for ϵ_i for energy based method for the i^{th} LSF region. If a LSF region is assigned to have peak with only energy based method and if ϵ_i of this LSF region is larger than this threshold, this LSF region is assigned to contain peak in it.

To extract these thresholds, the percentage of correct and false detected regions which are assigned peak in it versus these thresholds are calculated and plotted on the figures. The thresholds which correct maximum number of misclassified regions and introduce minimum number of misclassified regions are selected. These figures are given in the following pages.

Figures in the first column and the second column show percentage of correct classification and wrong classification versus threshold, respectively. Three pages are clustered to give the figures for one type of thresholds. LSF regions are presented sequentially in those pages. Selected thresholds are also given in the captions.

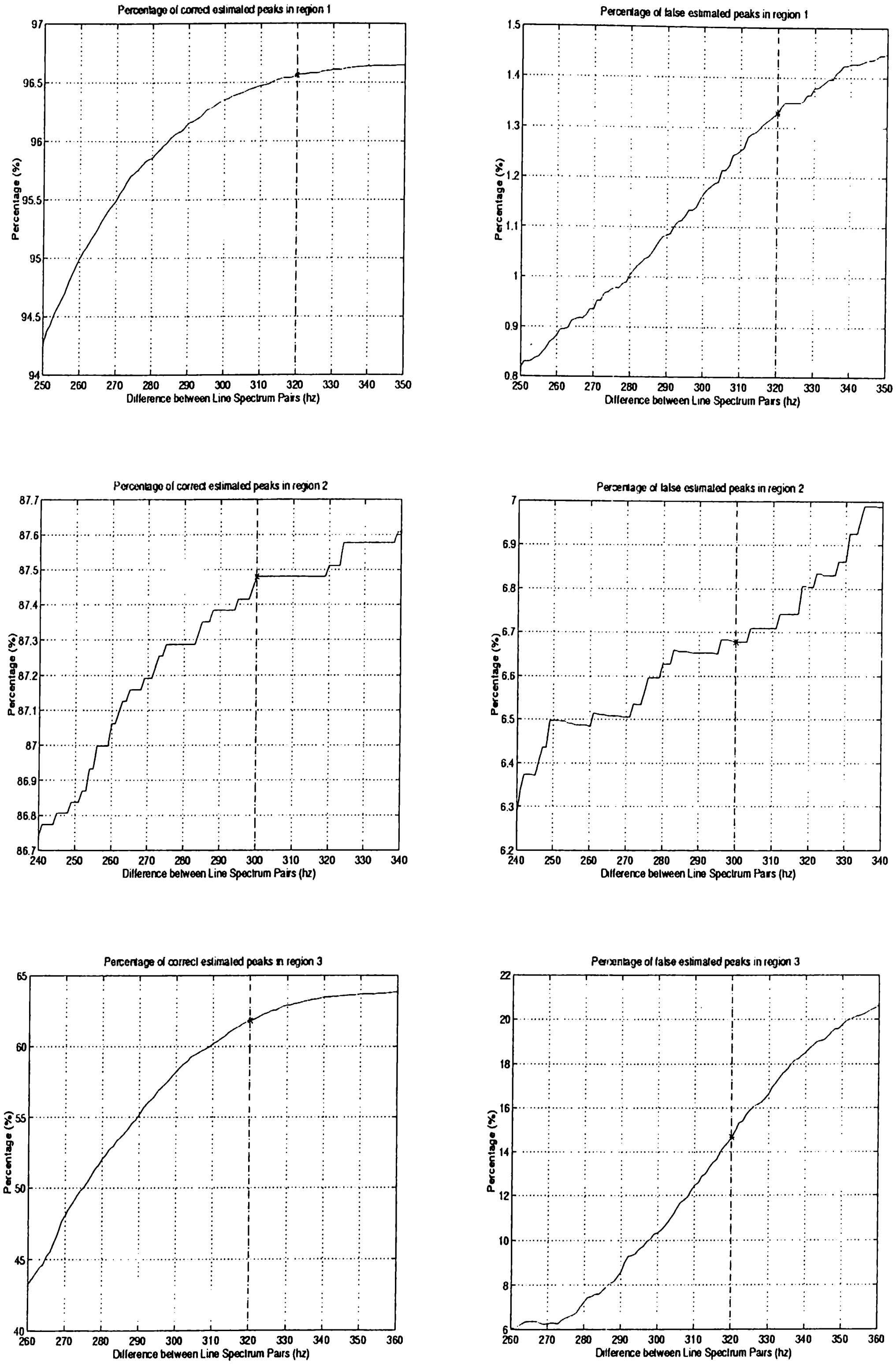


Figure A.1: Percentage versus T_i for the LSF regions 1, 2 and 3. First column shows the percentage of correct estimated of regions which contains peak in it. Second column shows the misclassified regions which has no peak in it. $T_1 = 320$, $T_2 = 300$, $T_3 = 320$.

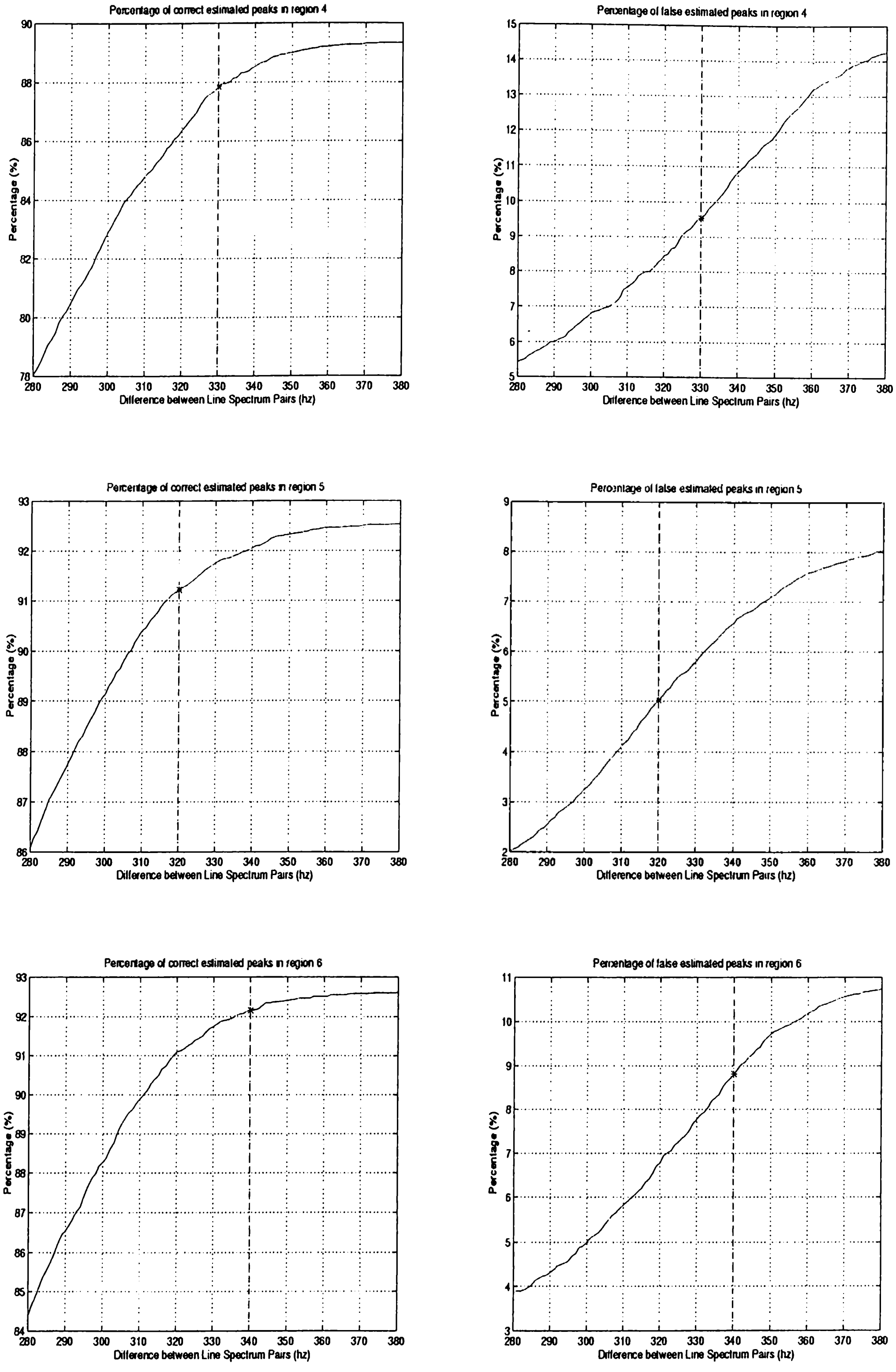


Figure A.2: Percentage versus T_i for the LSF regions 4, 5 and 6. $T_4 = 330$, $T_5 = 320$, $T_6 = 340$.

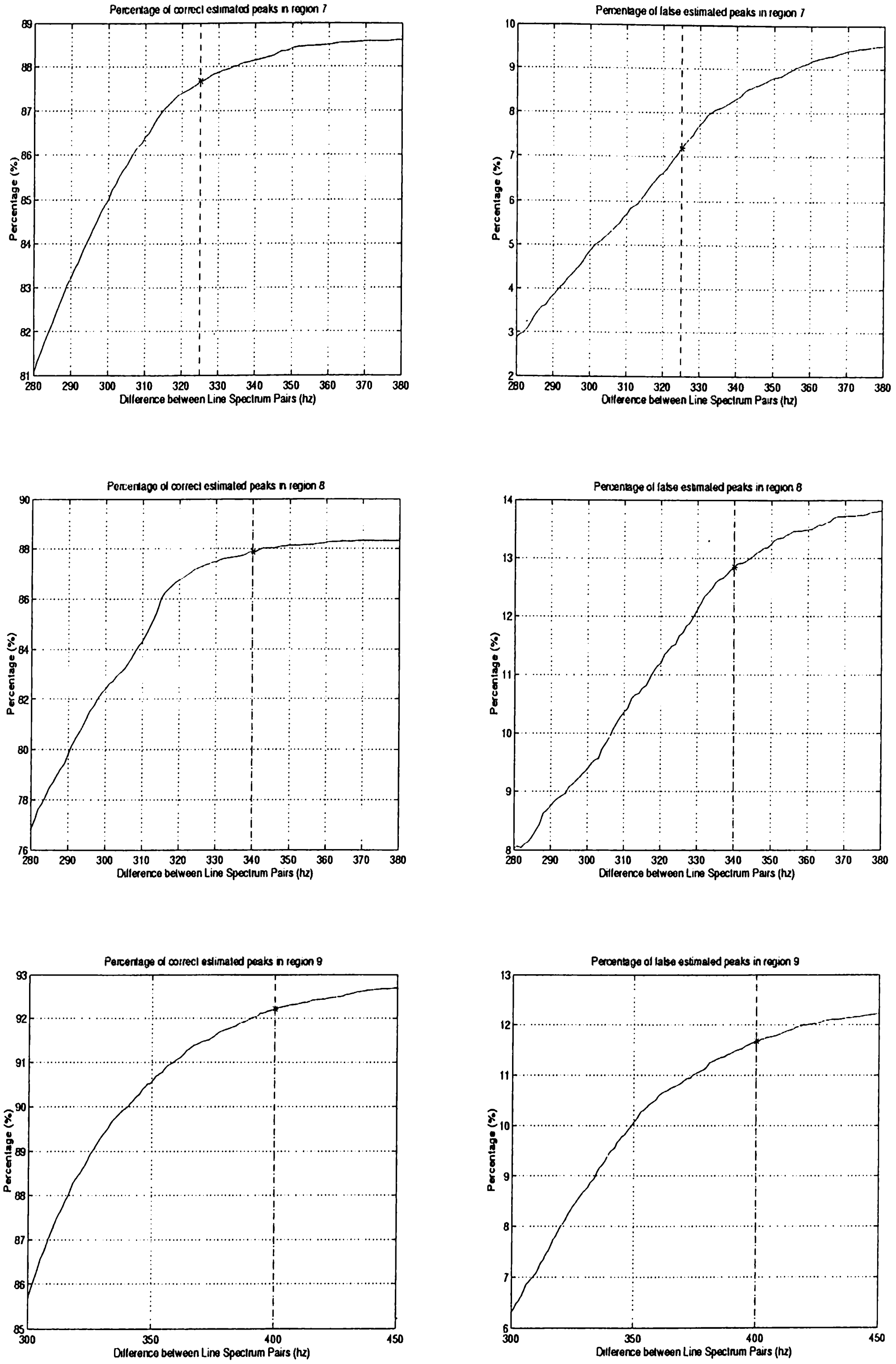


Figure A.3: Percentage versus T_i for the LSF regions 7, 8 and 9. $T_7 = 325$, $T_8 = 340$, $T_9 = 400$.

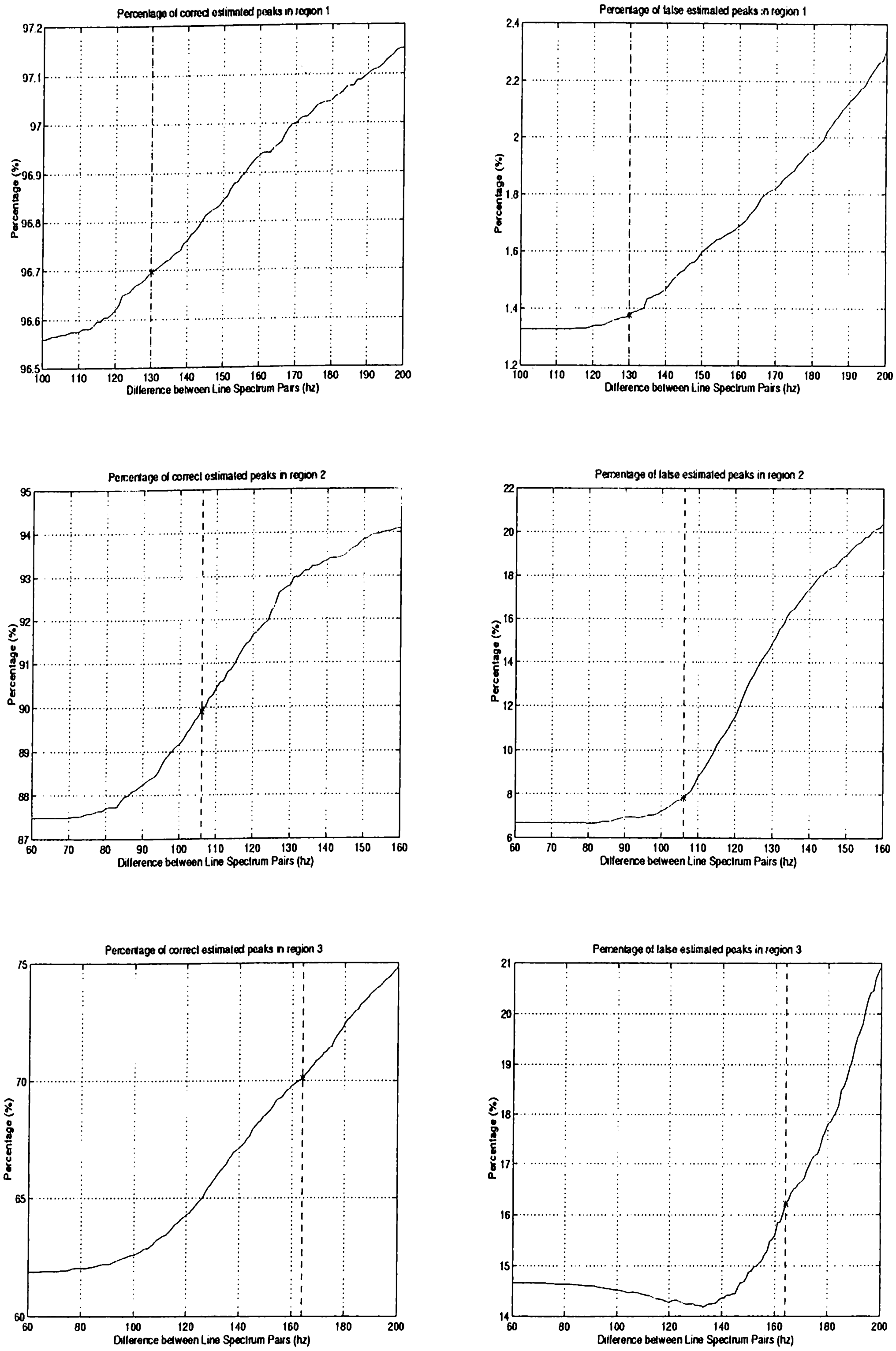


Figure A.4: Percentage versus γ_i for the LSF regions 1, 2 and 3. $\gamma_1 = 130$, $\gamma_2 = 106$, $\gamma_3 = 164$.

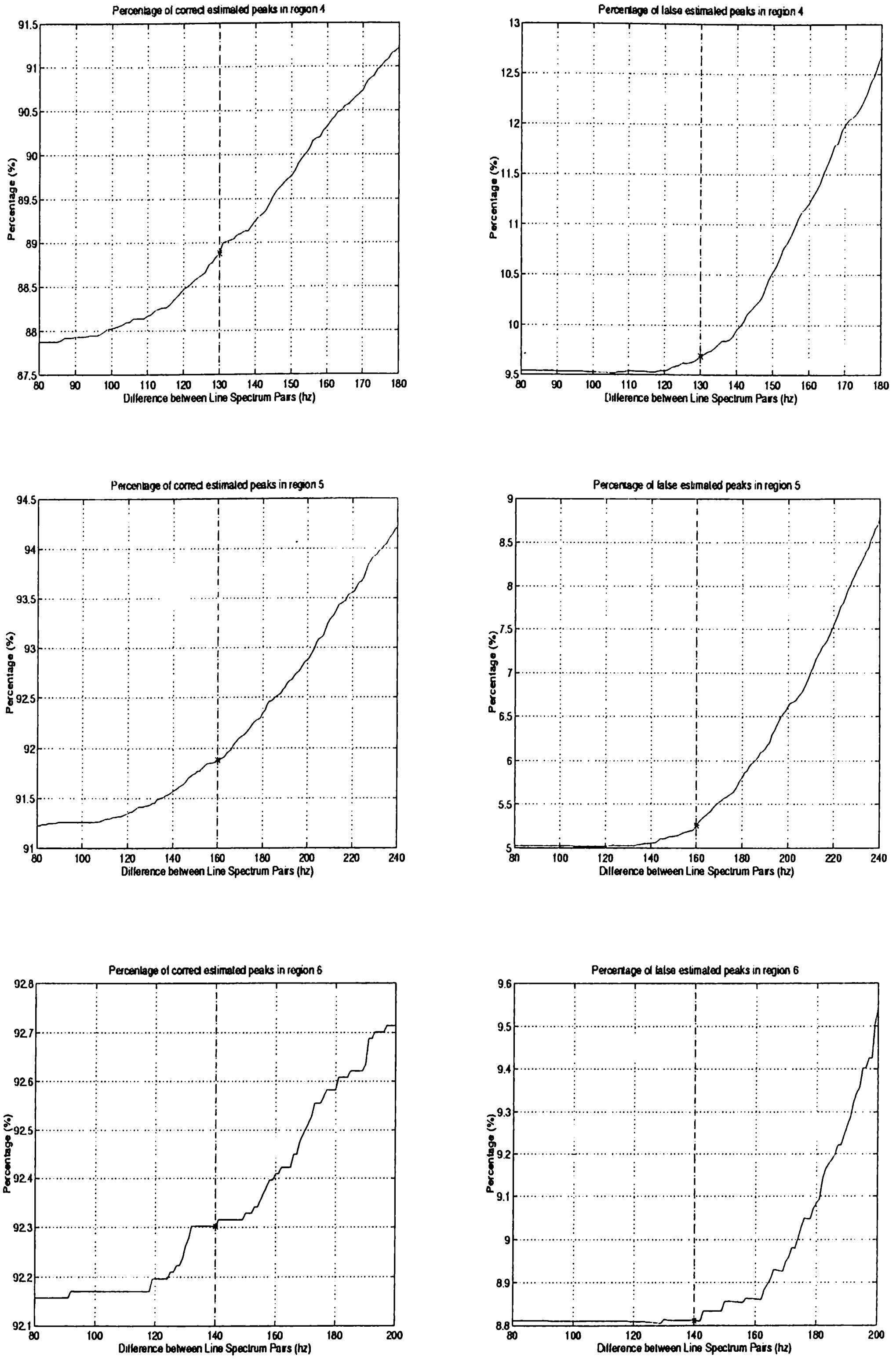


Figure A.5: Percentage versus γ_i for the LSF regions 4, 5 and 6. $\gamma_4 = 130$, $\gamma_5 = 160$, $\gamma_6 = 140$.

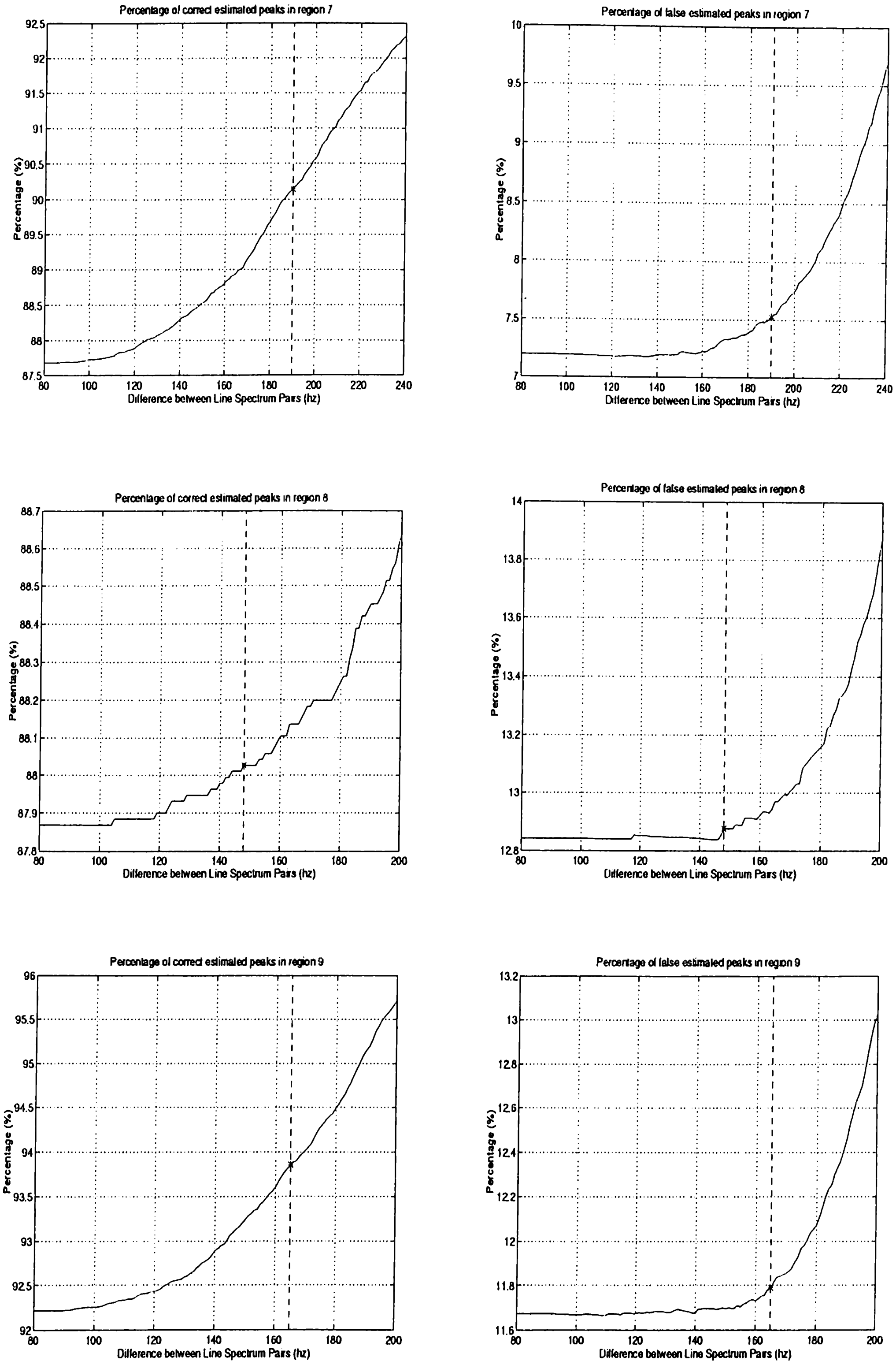


Figure A.6: Percentage versus γ_i for the LSF regions 7, 8 and 9. $\gamma_7 = 190$, $\gamma_8 = 148$, $\gamma_9 = 165$.

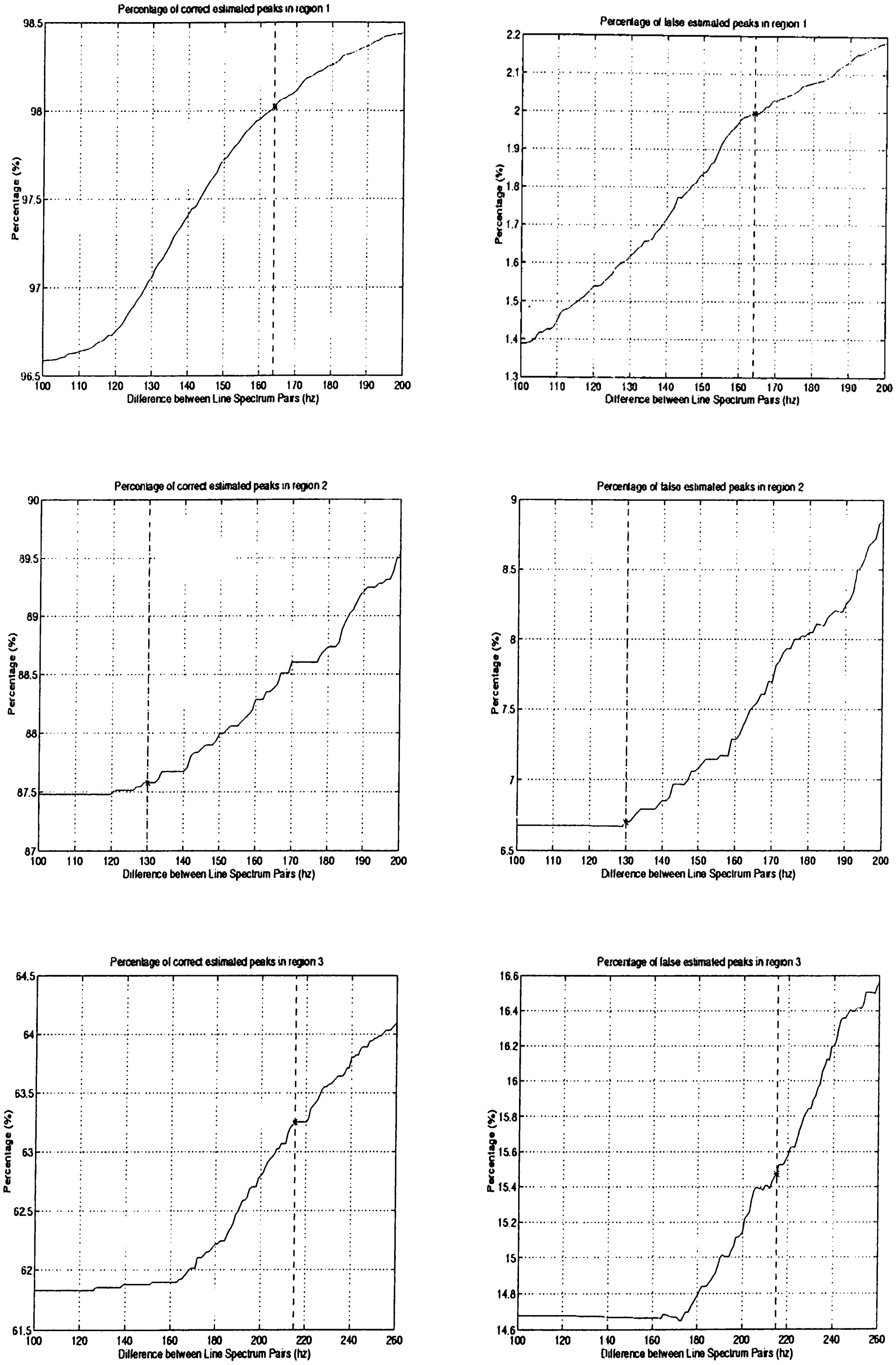


Figure A.7: Percentage versus α_i for the LSF regions 1, 2 and 3. $\alpha_1 = 164$, $\alpha_2 = 130$, $\alpha_3 = 215$.

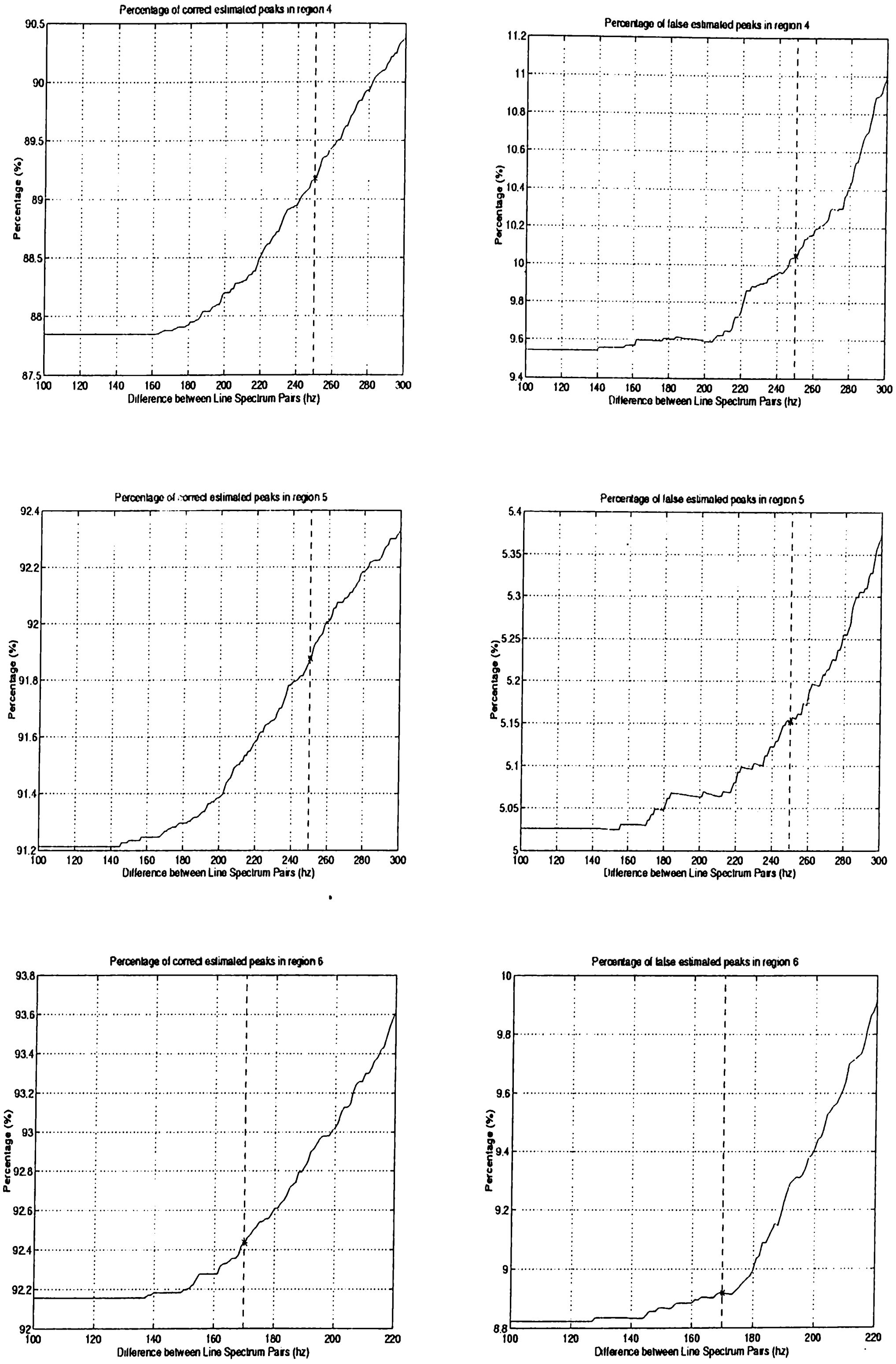


Figure A.8: Percentage versus α_i for the LSF regions 4, 5 and 6. $\alpha_4 = 250$, $\alpha_5 = 250$, $\alpha_6 = 170$.

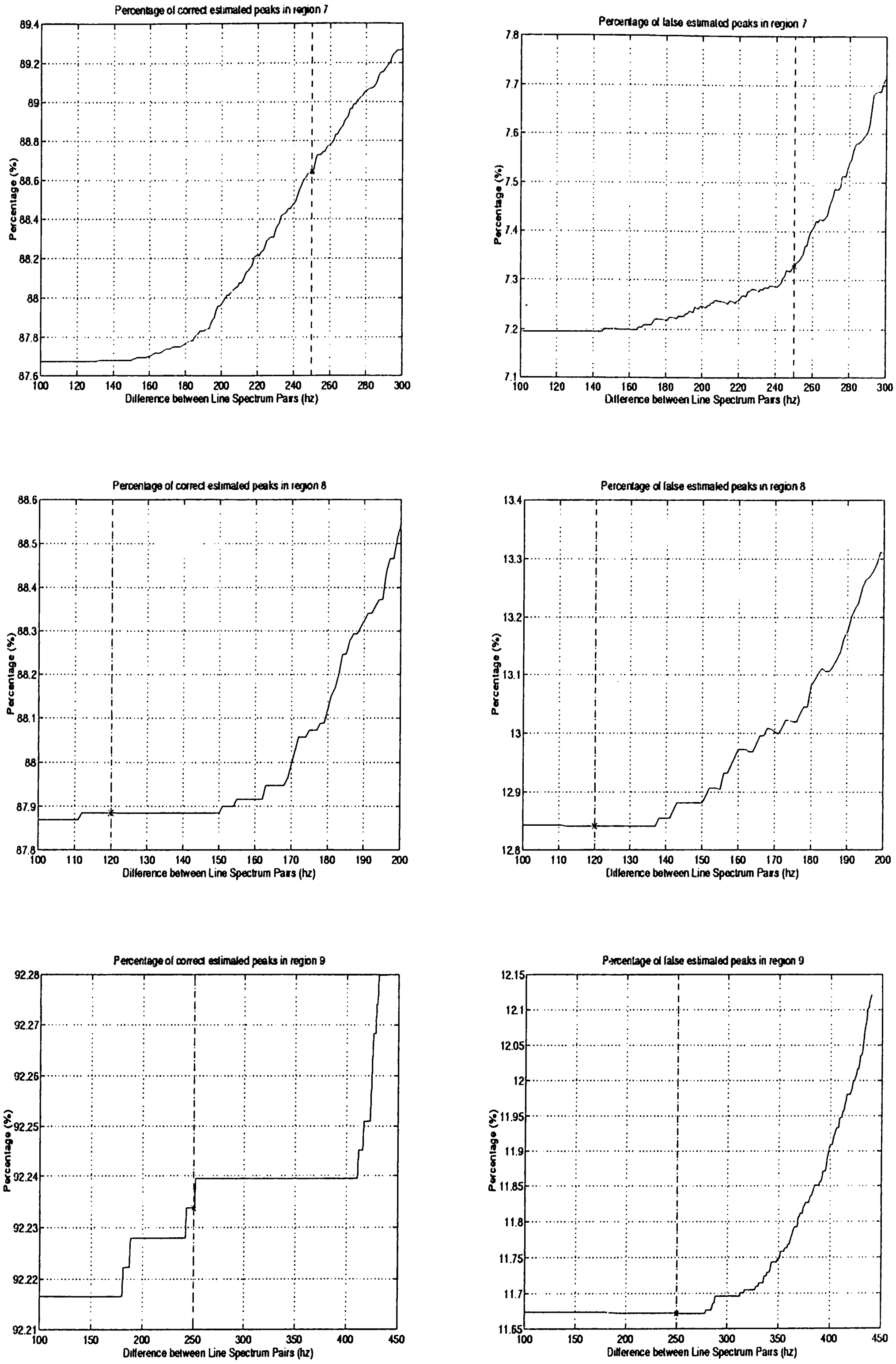


Figure A.9: Percentage versus α_i for the LSF regions 7, 8 and 9. $\alpha_7 = 250$, $\alpha_8 = 120$, $\alpha_9 = 250$.

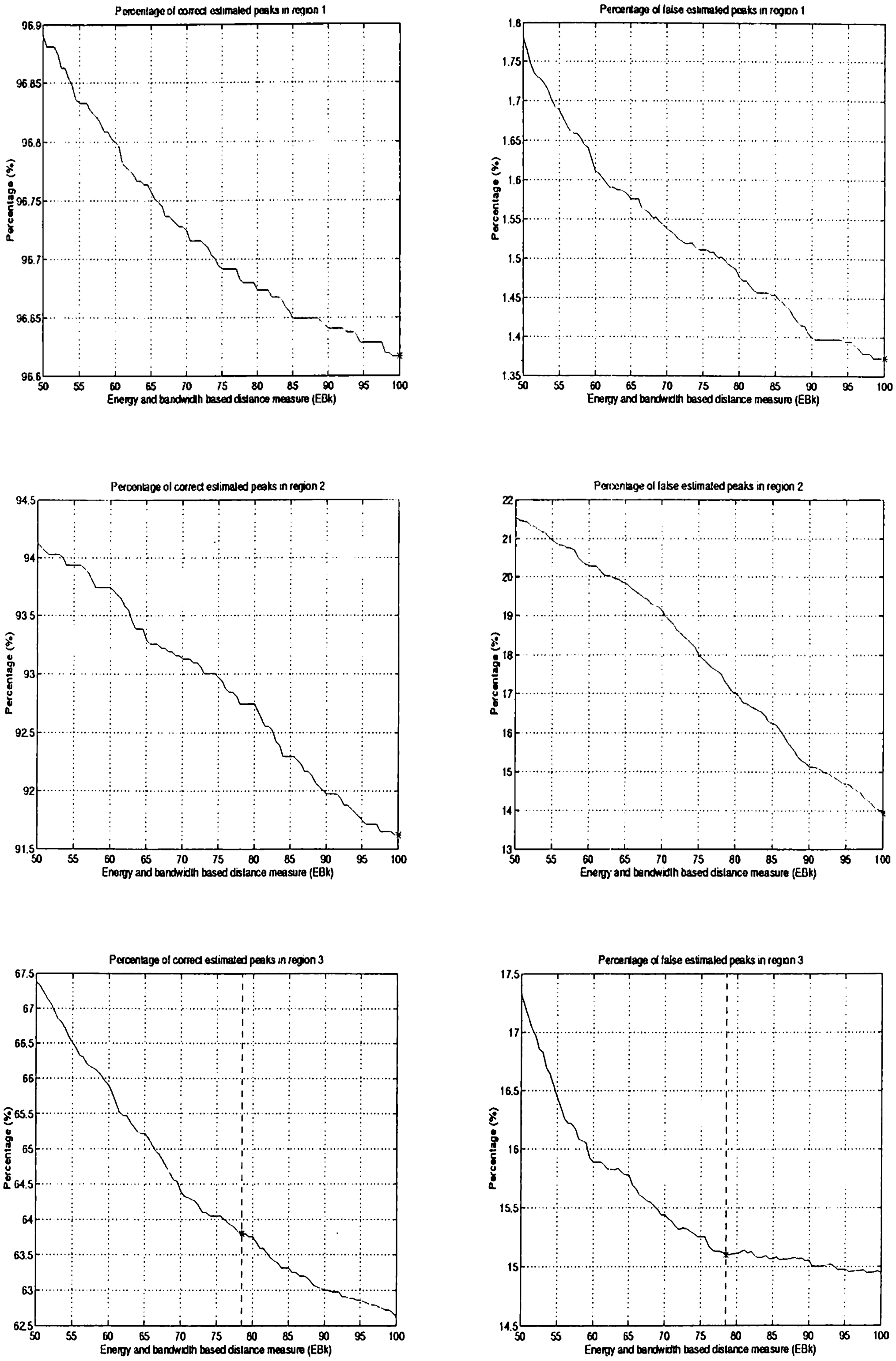


Figure A.10: Percentage versus β_i for the LSF regions 1, 2 and 3. $\beta_1 = N/A$, $\beta_2 = N/A$, $\beta_3 = 78.5$.

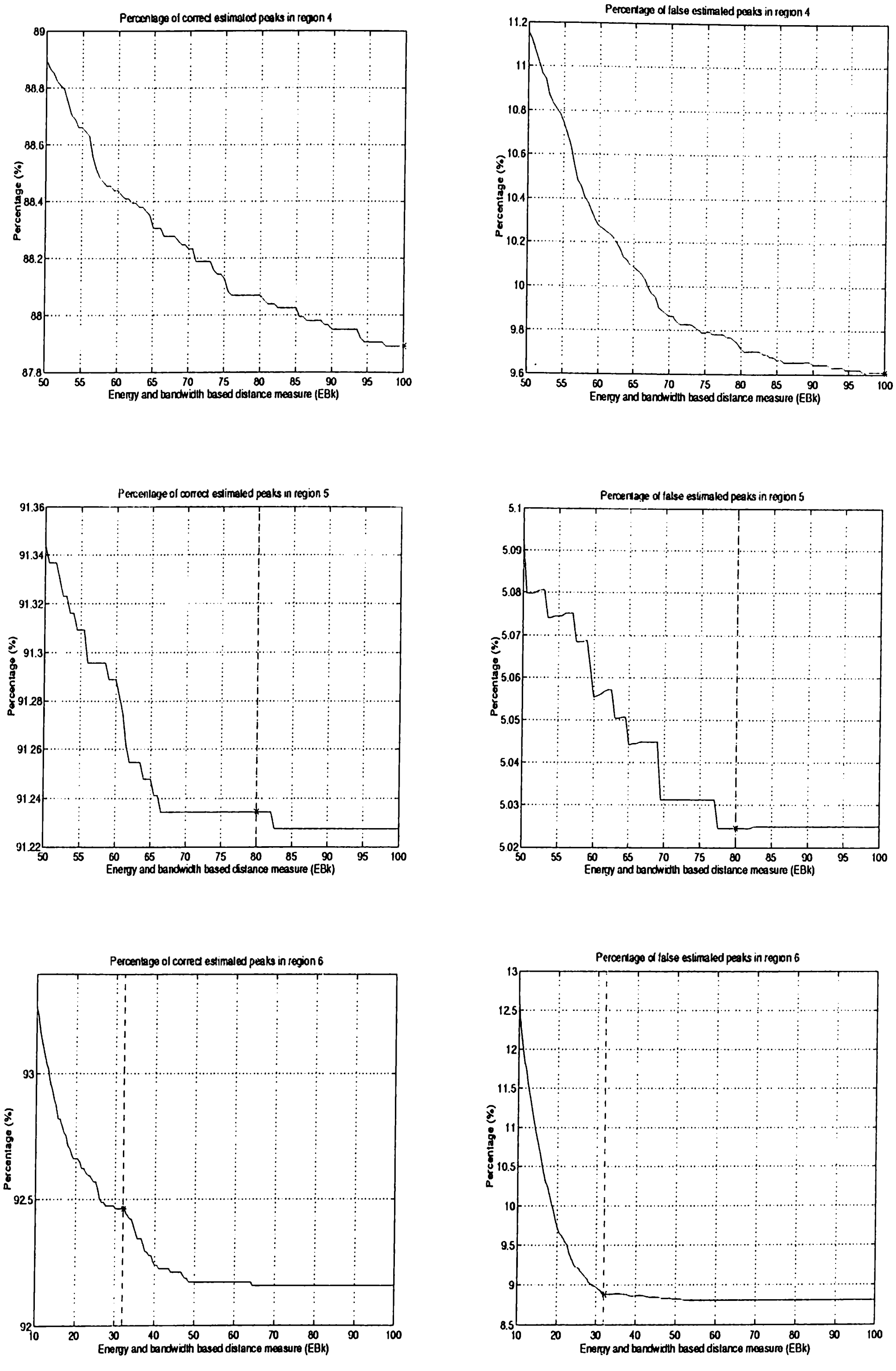


Figure A.11: Percentage versus β_i for the LSF regions 4, 5 and 6. $\beta_4 = N/A$, $\beta_5 = 80$, $\beta_6 = 32$.

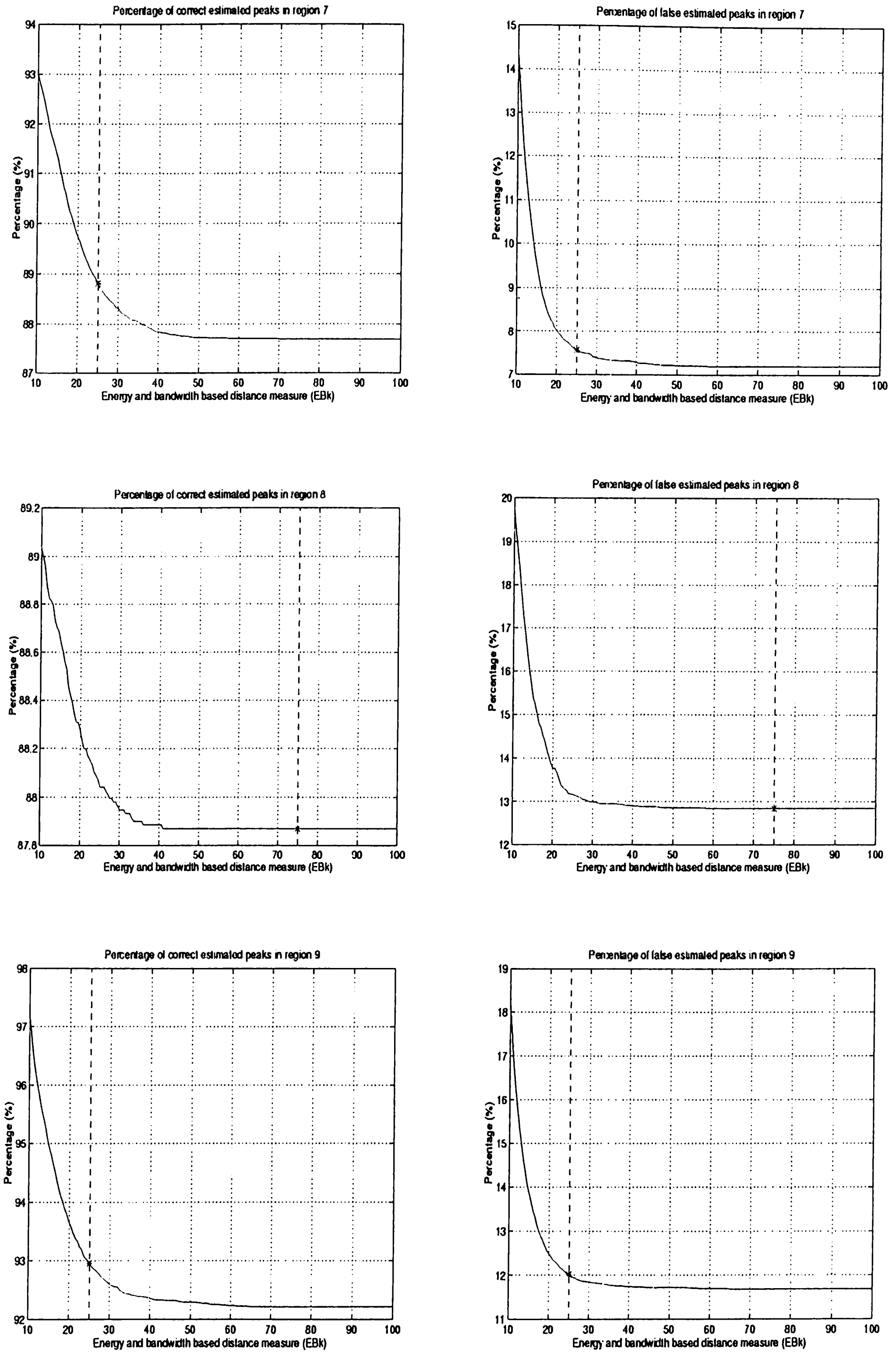


Figure A.12: Percentage versus β_i for the LSF regions 7, 8 and 9. $\beta_7 = 25$, $\beta_8 = 75$, $\beta_9 = 25$.

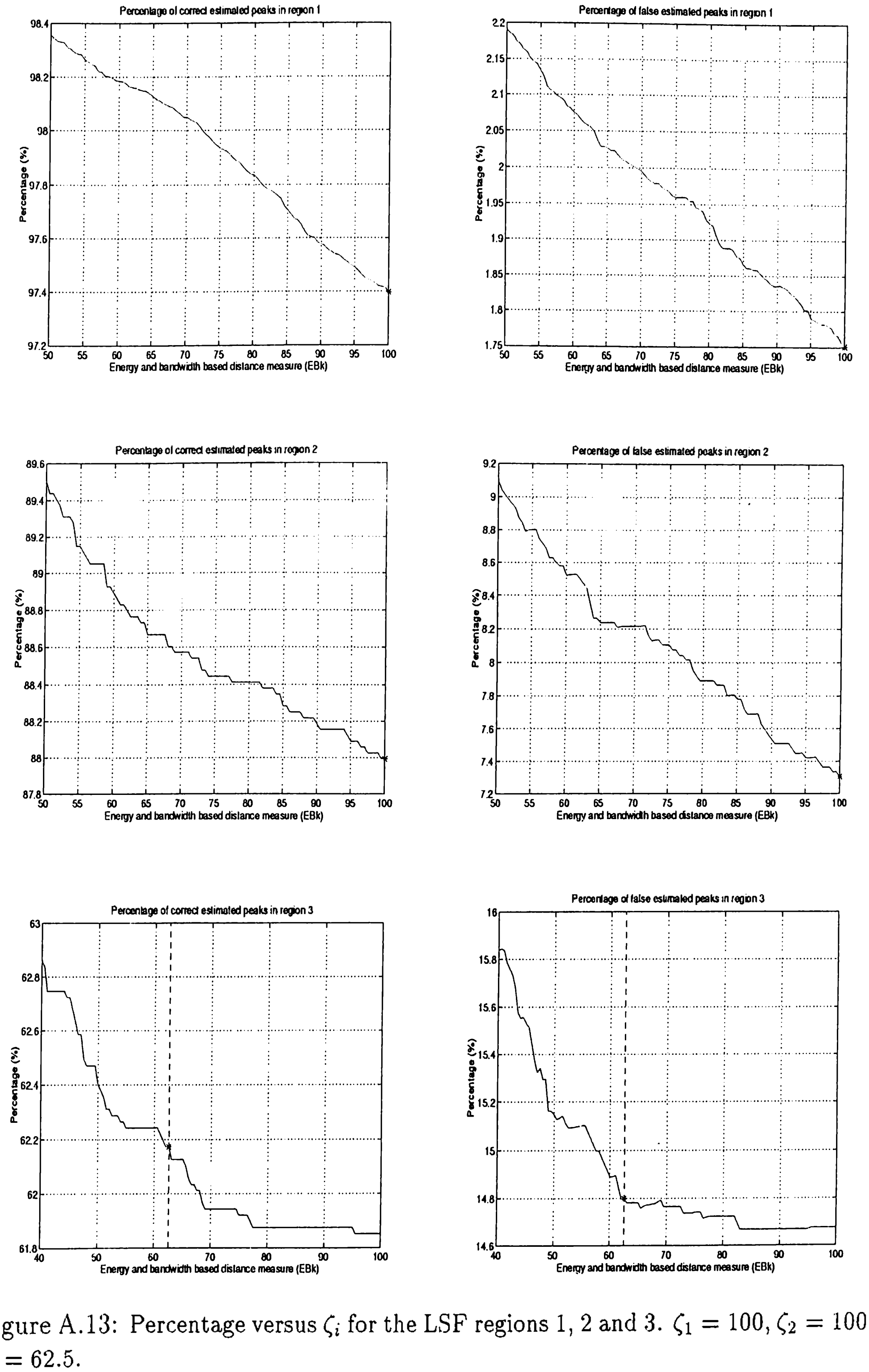


Figure A.13: Percentage versus ζ_i for the LSF regions 1, 2 and 3. $\zeta_1 = 100, \zeta_2 = 100, \zeta_3 = 62.5$.

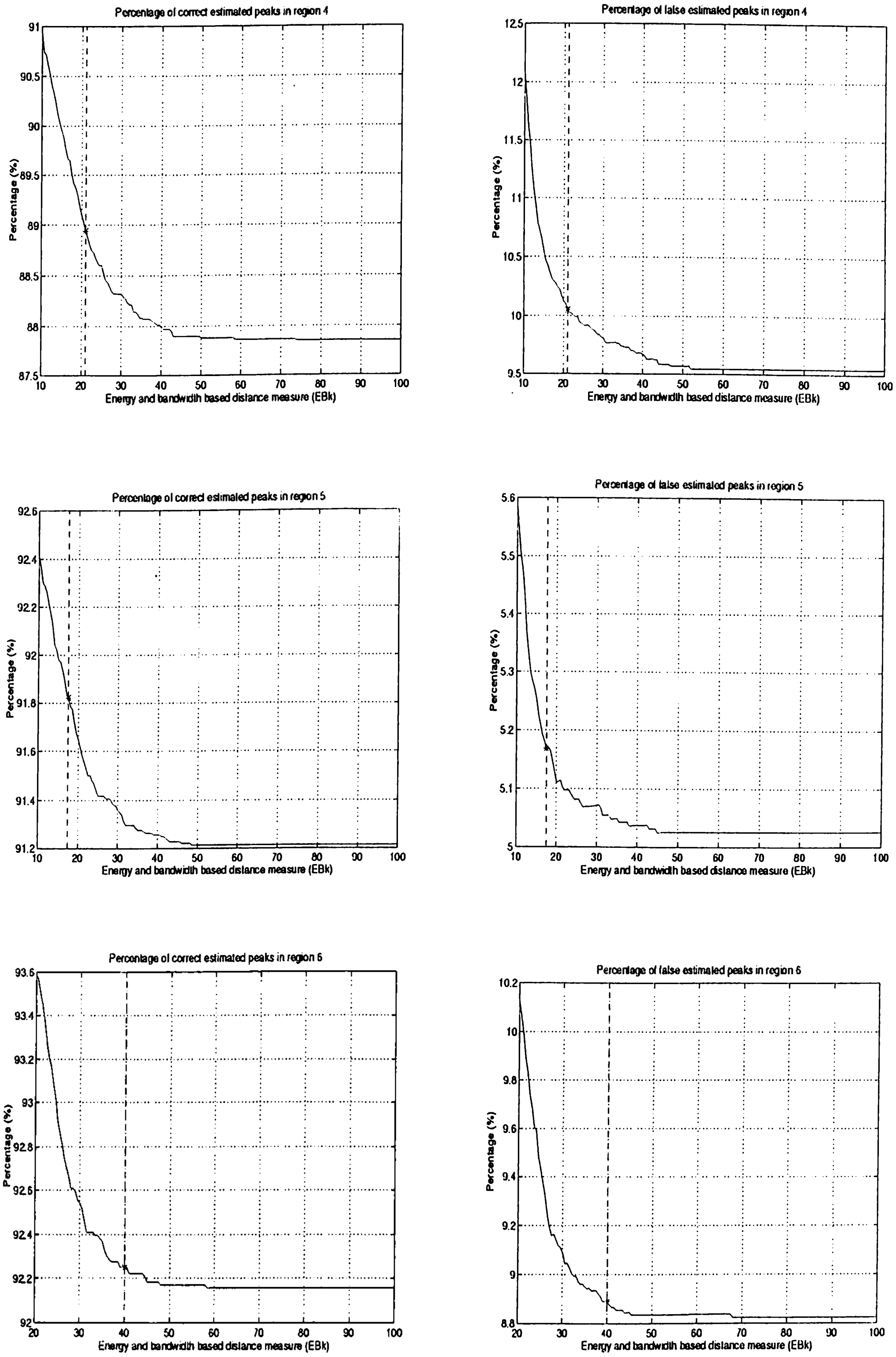
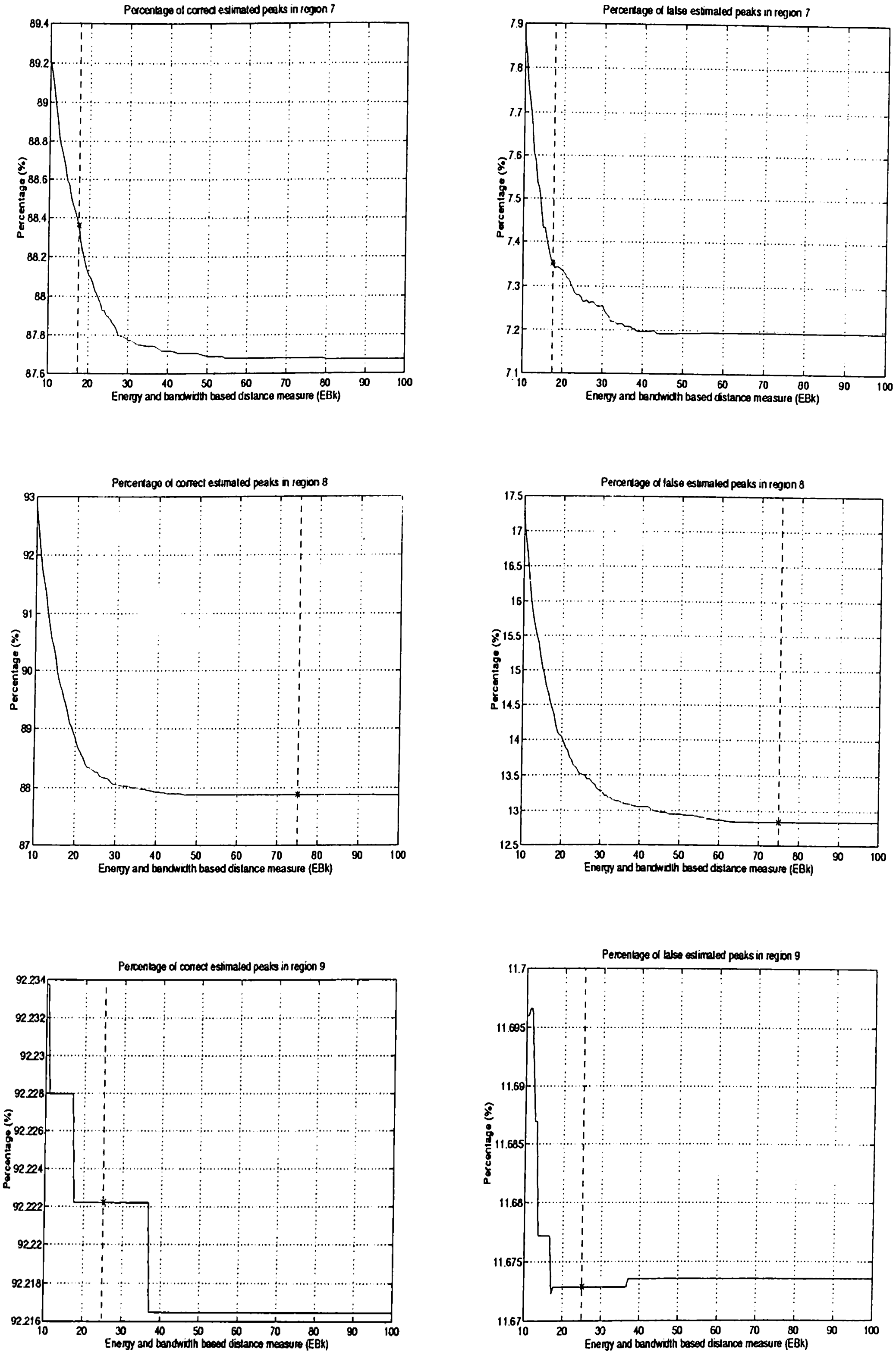


Figure A.14: Percentage versus ζ_i for the LSF regions 4, 5 and 6. $\zeta_4 = 21$, $\zeta_5 = 17.5$, $\zeta_6 = 40$.


Percentage of correct estimated peaks in region 7

Percentage of false estimated peaks in region 7

Percentage of correct estimated peaks in region 8

Percentage of false estimated peaks in region 8

Percentage of correct estimated peaks in region 9

Percentage of false estimated peaks in region 9

Figure A.15: Percentage versus ζ_i for the LSF regions 7, 8 and 9. $\zeta_7 = 17.5$, $\zeta_8 = 75$, $\zeta_9 = 25$.

APPENDIX B

Selection of parameters for minimizing peak location estimation error

In Chapter 2, a novel method to make precise estimation of the peak location is proposed. In this chapter, statistical data used to obtain τ_i and μ_i for $i = 1 \cdots 9$ are presented.

For each LSF region, the number of occurrence of the error between estimated peak and actual peak is calculated for different values of τ_i ranging from 0.15 to 1.75 for both voiced and entire speech by (2.4) from database. Value of τ_i which minimize or almost minimize standard deviation and maximize percentage of error smaller than 25 Hz is selected.

In the following tables, mean, standard deviation, percentage of error smaller than 25 Hz and percentage of error smaller than 50 hz are tabulated for simple averaging method, energy weighted mean method for $\tau = 0.15$ and energy weighted mean method for $\tau = \tau_i$. Selected τ_i values are tabulated in

Chapter 2. Also these tabulated datas versus τ_i is plotted and the figures are given in following pages.

Figures are divided into two sets: In the first set, standard deviation, percentage of error smaller than 25 Hz and percentage of error smaller than 50 Hz are plotted. As page setup, all figures in the first and second column belongs to voiced and entire speech, respectively. First row shows standard deviation versus τ_i and second and third row show percentage of error in peak location of estimation of smaller than 25 Hz and 50 Hz versus τ_i, respectively. LSF regions are sequentially presented in those pages. In the second set, number of occurrence versus error in peak location estimation is given. Column formation is same as the first set. LSF regions are presented sequentially.

Table B.1: Mean of the error between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for voiced speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	−13.44	−25.04	14.05	18.55	−7.681
$\tau = 0.15$	−9.322	−17.03	13.18	14.27	−5.467
$\tau = \tau_i$	5.656	7.956	4.643	−0.33	1.704
	6	7	8	9	
Simple mean	3.326	−0.9552	−6.884	14.08	
$\tau = 0.15$	2.655	0.2175	−4.508	12.91	
$\tau = \tau_i$	−0.569	3.094	3.654	4.535	

Table B.2: Mean of the error between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for entire speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	−34.017	−23.4	17.8	8.194	−7.959
$\tau = 0.15$	−25.67	−16.06	15.34	6.003	−6.062
$\tau = \tau_i$	7.266	7.497	2.394	−1.096	1.179
	6	7	8	9	
Simple mean	4.951	−0.535	−1.565	14.21	
$\tau = 0.15$	4.086	0.266	−0.703	13.03	
$\tau = \tau_i$	−0.177	2.395	1.339	5.132	

Table B.3: Standard deviation of the error between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for voiced speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	37.16	38.42	55.46	62.95	52.22
$\tau = 0.15$	28.9	31.03	44.84	51.52	40.84
$\tau = \tau_i$	8.191	9.235	13.92	18.05	12.98
	6	7	8	9	
Simple mean	64.57	55.41	67.91	55.43	
$\tau = 0.15$	53.46	44.16	56.62	44.84	
$\tau = \tau_i$	16.85	13.44	17.14	13.93	

Table B.4: Standard deviation of the error between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for entire speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	43,74	43.1	58.4	71.45	56.33
$\tau = 0.15$	34.73	35.4	48.49	61.01	45.96
$\tau = \tau_i$	9.047	10.35	13.43	21.39	13.64
	6	7	8	9	
Simple mean	69.4	56.71	71.09	54.44	
$\tau = 0.15$	59.38	46.58	61.09	44.93	
$\tau = \tau_i$	19.31	13.73	19.87	13.33	

Table B.5: Percentage of the error smaller than 25 Hz between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for voiced speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	58.23	56.69	30.29	29.43	39.16
$\tau = 0.15$	67.45	67.04	38.57	38.29	49.17
$\tau = \tau_i$	99.6	98.23	95.28	91.69	97.56
	6	7	8	9	All Regions
Simple mean	25.88	35.41	22.19	38.65	41.07
$\tau = 0.15$	31.62	44.32	27.92	47.77	50.15
$\tau = \tau_i$	92.6	97.32	91.52	98.12	97.07

Table B.6: Percentage of the error smaller than 25 Hz between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for entire speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	43.74	50.71	30.68	24.08	34.98
$\tau = 0.15$	57.29	61.62	37.33	29.99	43.29
$\tau = \tau_i$	99.06	97.44	95.45	86.43	96.2
	6	7	8	9	All Regions
Simple mean	23.74	33.89	20.96	38.48	36.16
$\tau = 0.15$	28.68	41.84	25.5	46.73	45.25
$\tau = \tau_i$	88.83	96.15	86.68	97.71	95.72

Table B.7: Percentage of the error smaller than 50 Hz between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for voiced speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	81.46	83.36	58.16	55.43	65.23
$\tau = 0.15$	90.23	88.33	70.76	68.21	78.03
$\tau = \tau_i$	99.99	99.81	99.36	97.83	99.69
	6	7	8	9	All Regions
Simple mean	49.12	61.9	45.63	66.03	66.46
$\tau = 0.15$	61.74	73.45	56.16	75.43	77.14
$\tau = \tau_i$	98.55	99.76	98.72	99.78	99.55

Table B.8: Percentage of the error smaller than 50 Hz between actual peak location and estimated peak location for simple mean technique, $\tau = 0.15$ and $\tau = \tau_i$ for entire speech.

	LSF Regions				
	1	2	3	4	5
Simple mean	74.58	79.02	56.66	47.18	60.77
$\tau = 0.15$	84.82	84.51	67.24	57.05	72.05
$\tau = \tau_i$	99.97	99.70	99.58	95.82	99.44
	6	7	8	9	All Regions
Simple mean	45.74	60.75	43.36	66.82	63.42
$\tau = 0.15$	56	71.36	52.48	75.79	73.46
$\tau = \tau_i$	97.27	99.54	97.09	99.71	99.17

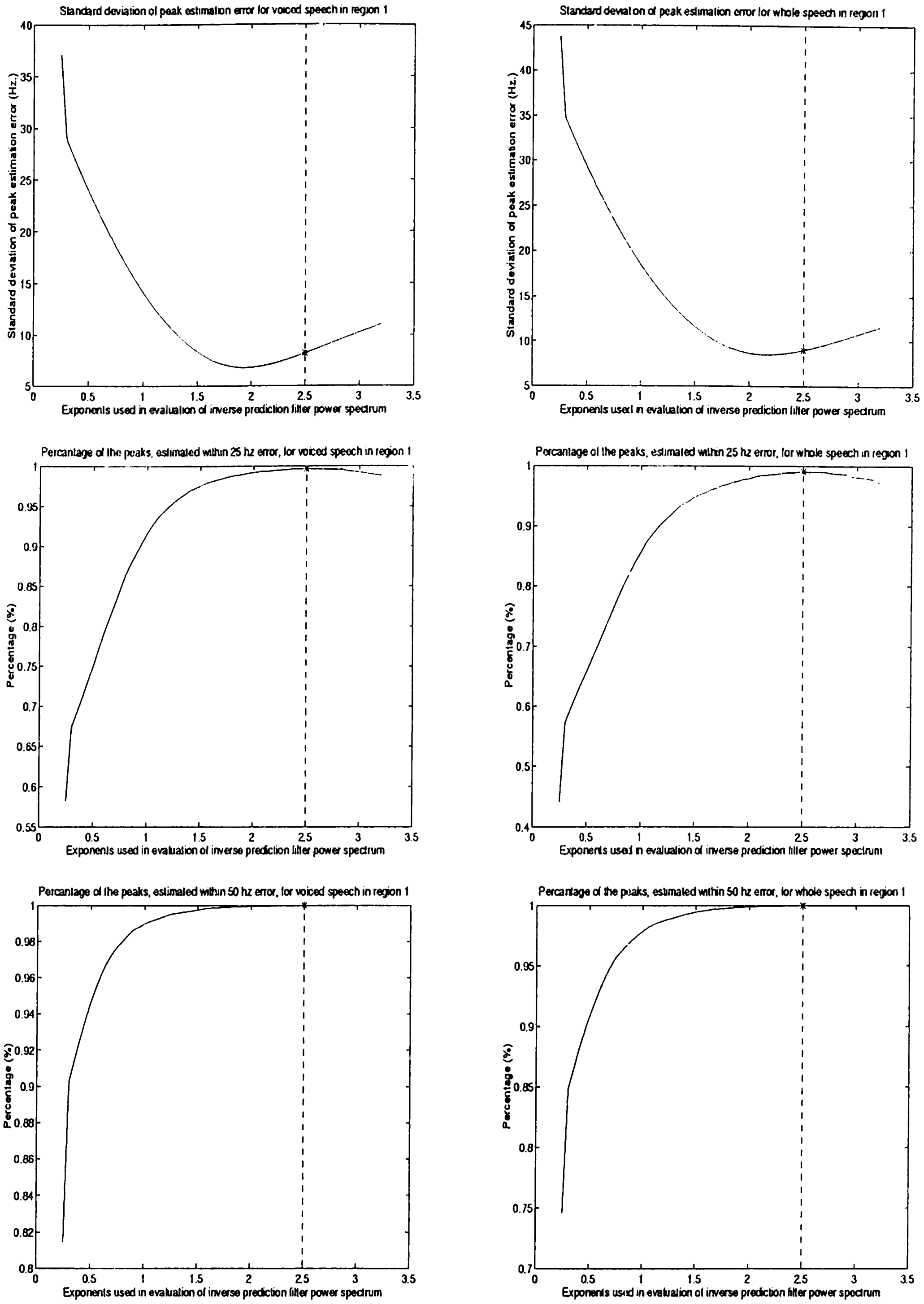


Figure B.1: Statistics of error in peak location estimation error for LSF region 1. First and second column presents statistical data about voiced frames and all frames, respectively. First index in all figures corresponds to the data extracted by simple averaging. First row shows standard deviation of error versus varying τ values. Second and third row shows percentage of error in peak location estimation smaller than 25 Hz and 50 Hz versus varying τ values, respectively. τ_1 is selected as 2.5.

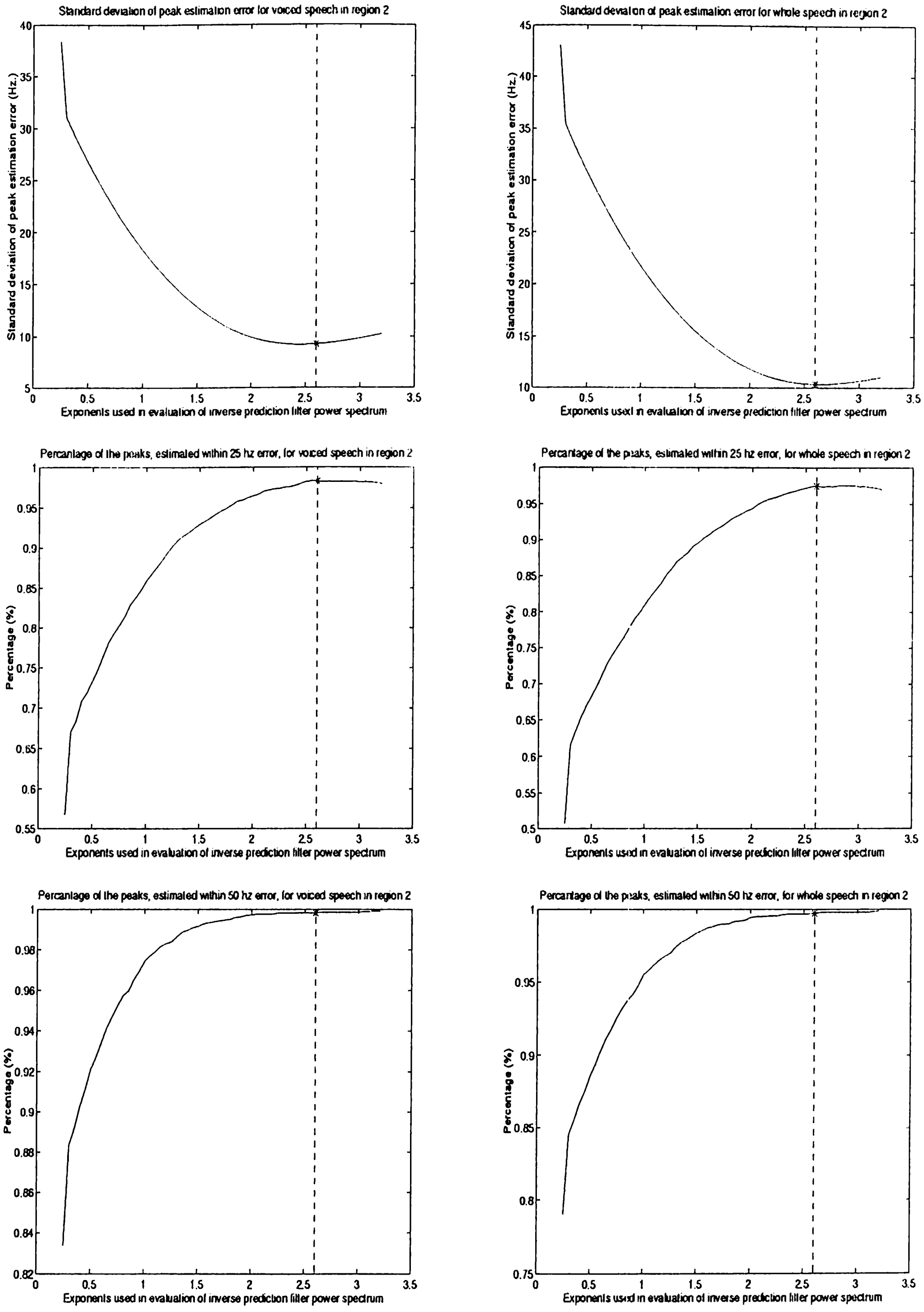
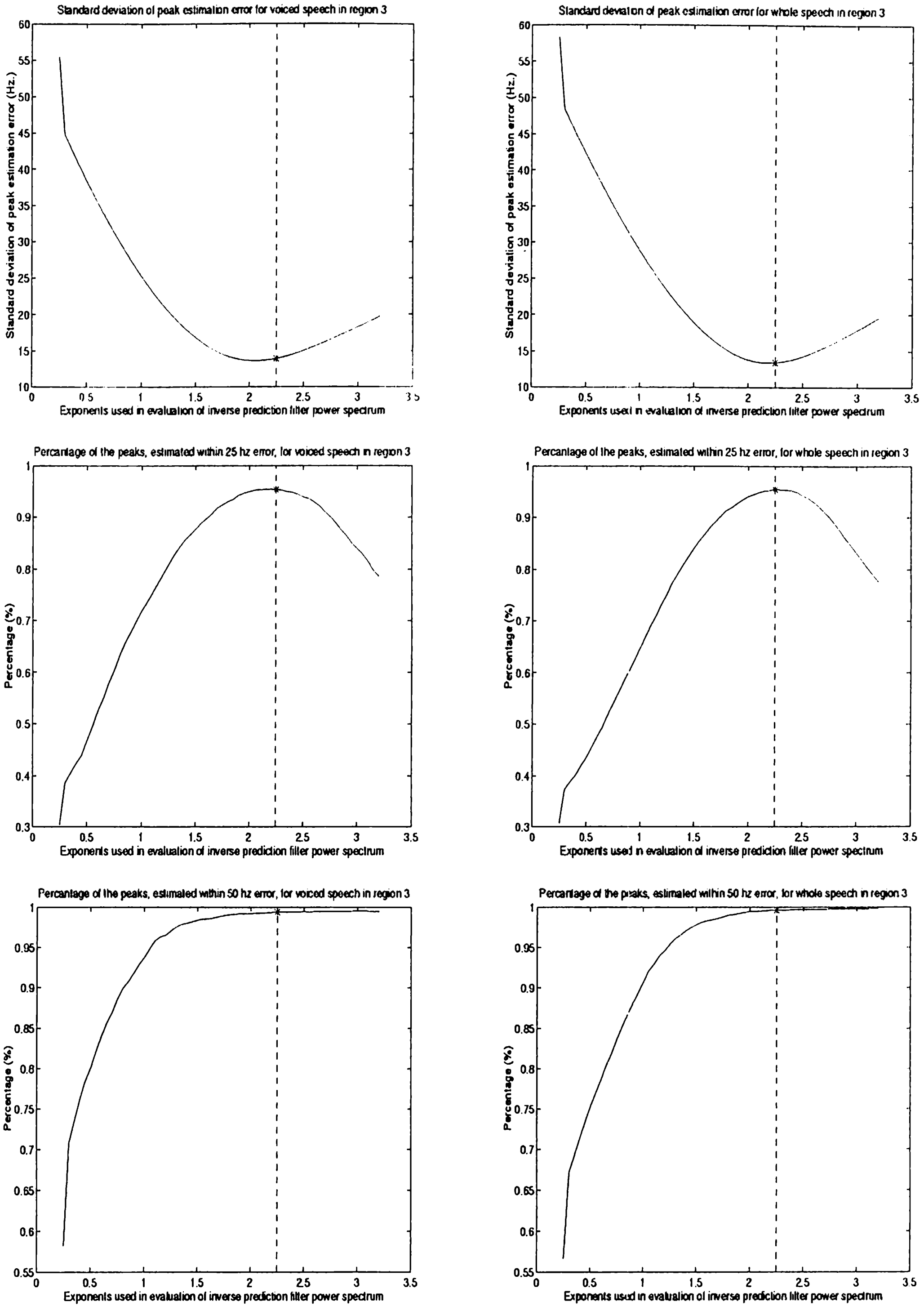


Figure B.2: Statistics of error in peak location estimation error for LSF region 2. τ_2 is selected as 2.6.



Standard deviation of peak estimation error for voiced speech in region 3

Standard deviation of peak estimation error for whole speech in region 3

Percentage of the peaks, estimated within 25 hz error, for voiced speech in region 3

Percentage of the peaks, estimated within 25 hz error, for whole speech in region 3

Percentage of the peaks, estimated within 50 hz error, for voiced speech in region 3

Percentage of the peaks, estimated within 50 hz error, for whole speech in region 3

Standard deviation of peak estimation error (Hz.)
60
55
50
45
40
35
30
25
20
15
10
Percentage (%)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.95
0.85
0.75
0.65
0.55
0
0.5
1
1.5
2
2.5
3
3.5
Exponents used in evaluation of inverse prediction filter power spectrum

Figure B.3: Statistics of error in peak location estimation error for LSF region 3. τ_3 is selected as 2.25.

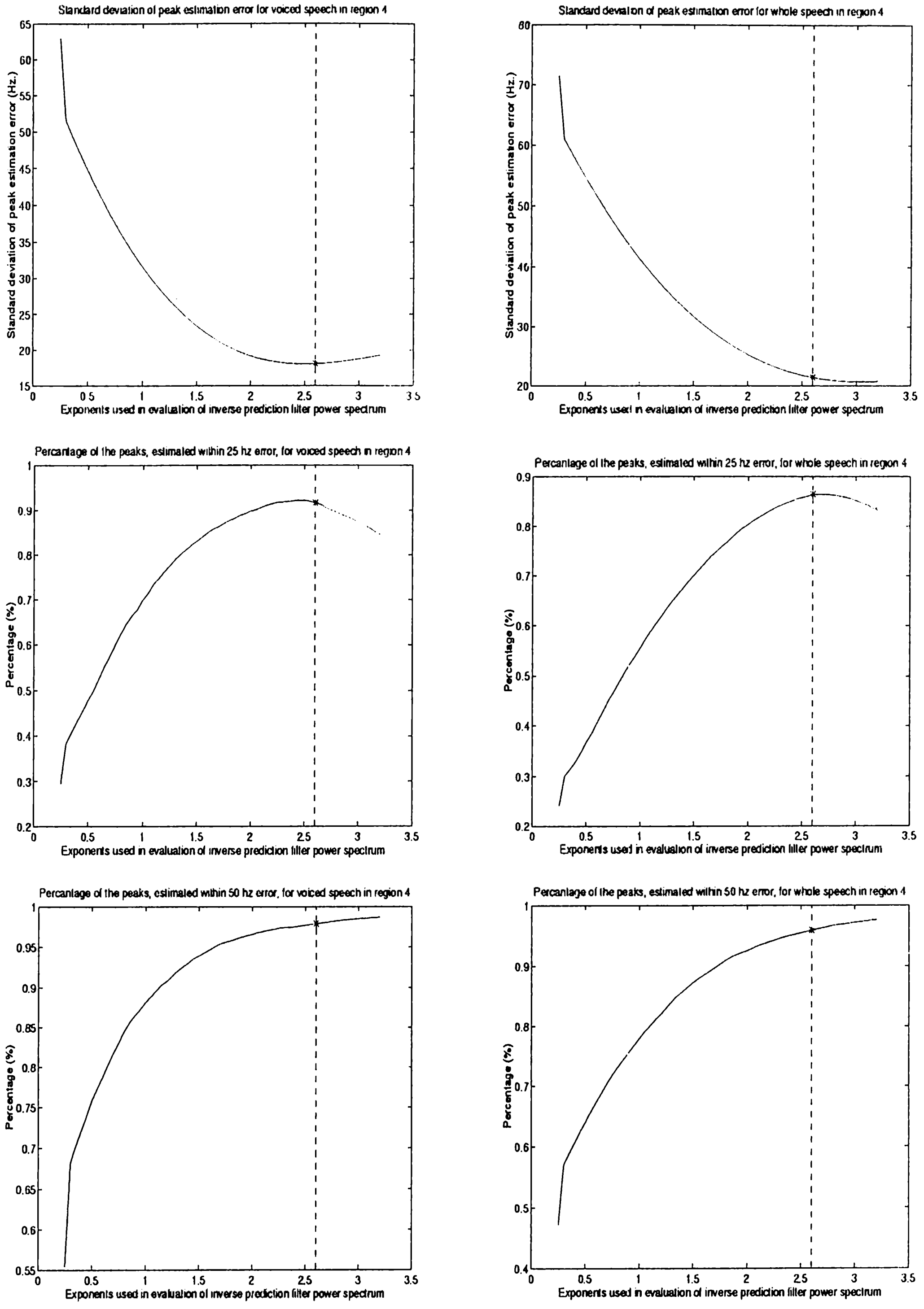


Figure B.4: Statistics of error in peak location estimation error for LSF region 4. τ_4 is selected as 2.6.

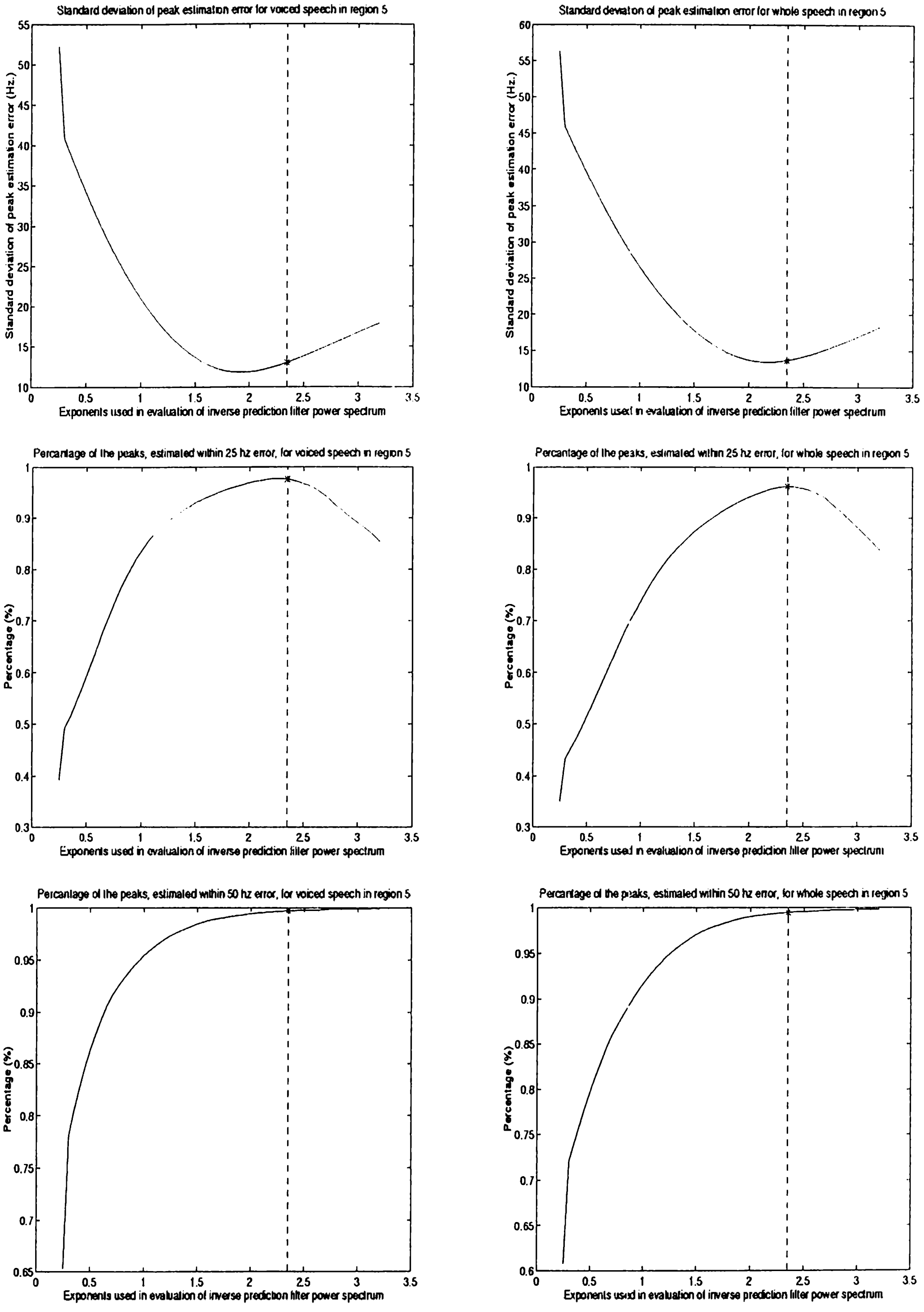


Figure B.5: Statistics of error in peak location estimation error for LSF region 5. τ_5 is selected as 2.35.

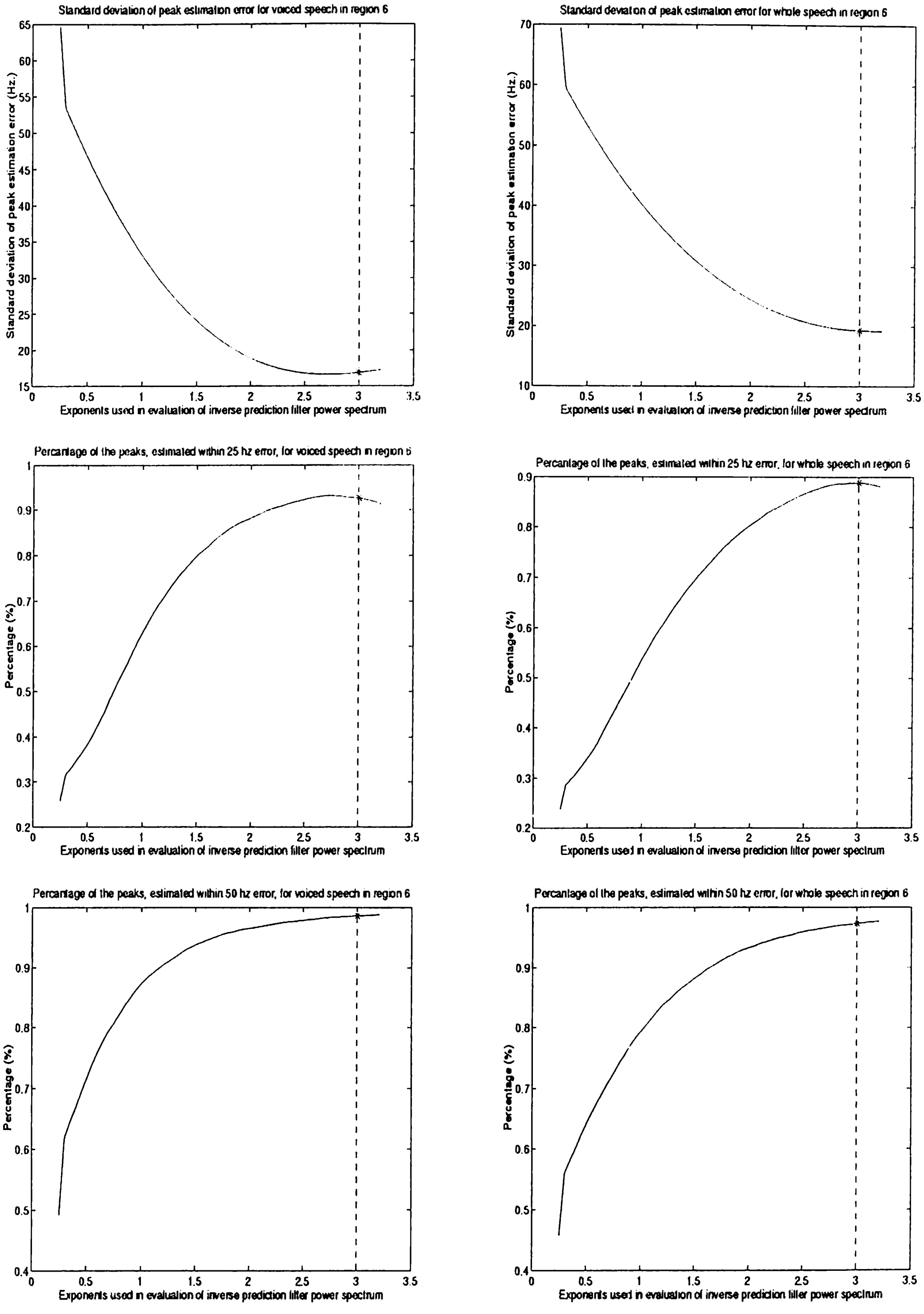


Figure B.6: Statistics of error in peak location estimation error for LSF region 6. τ_6 is selected as 3.0.

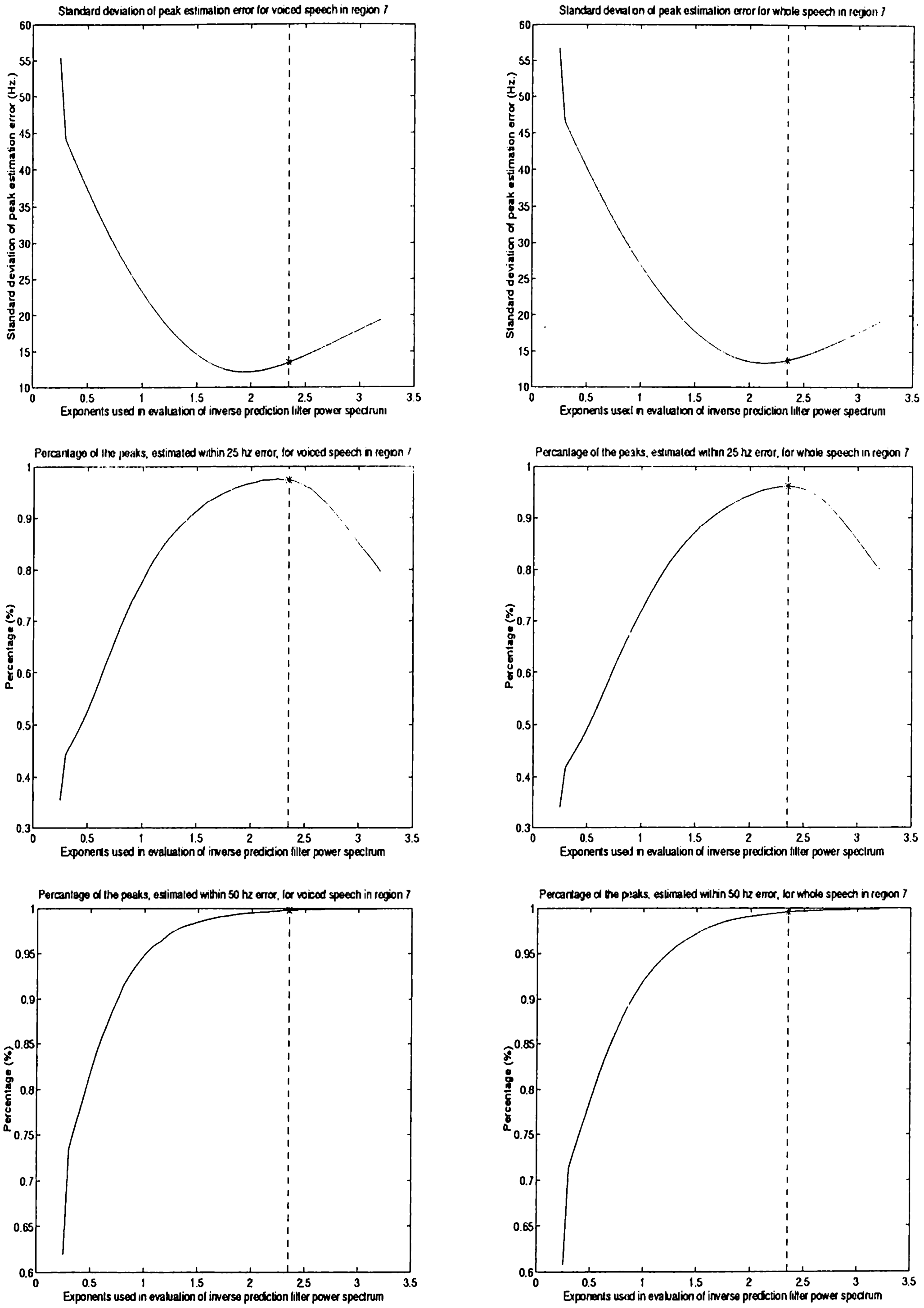


Figure B.7: Statistics of error in peak location estimation error for LSF region 7. τ_7 is selected as 2.35.

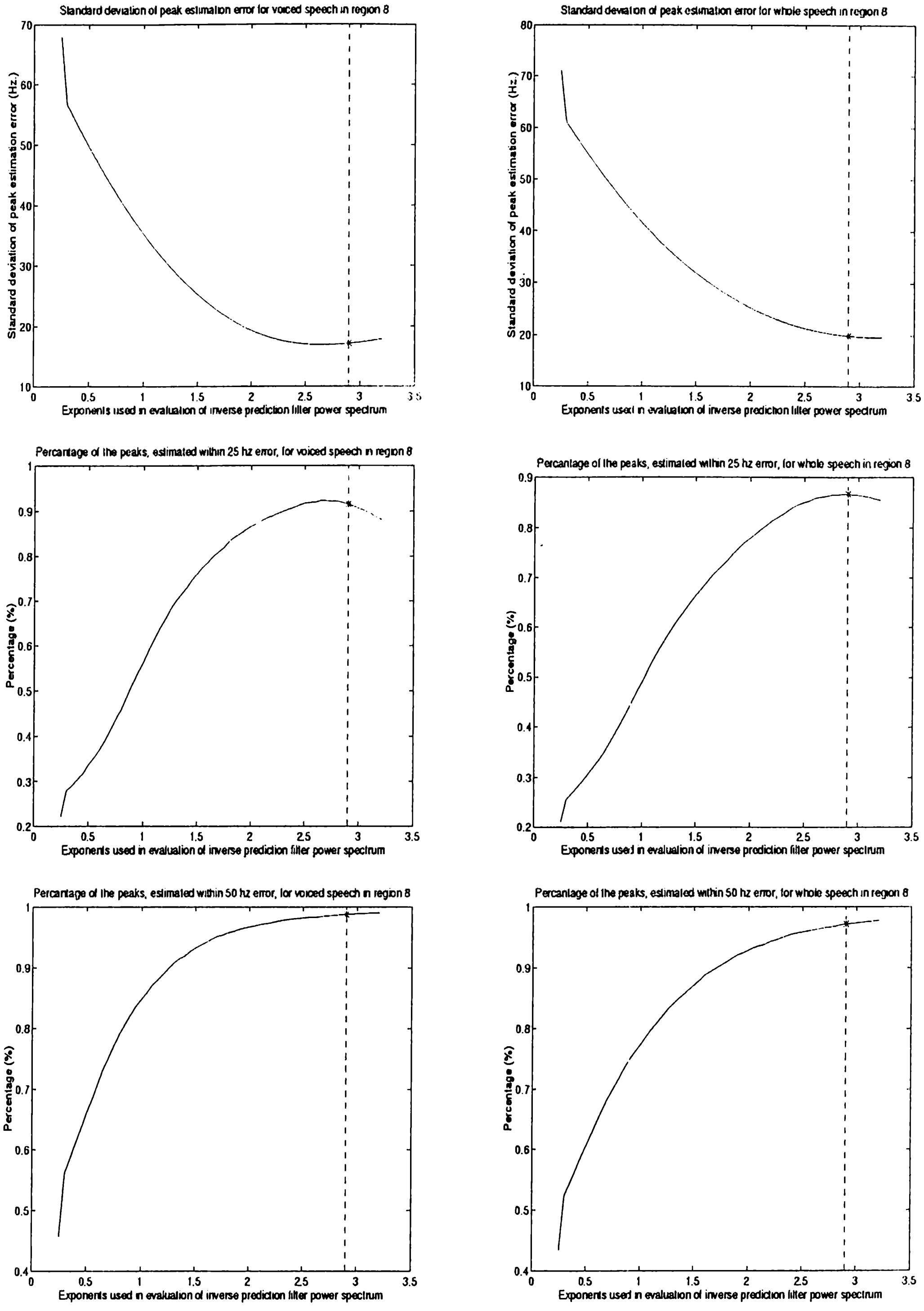


Figure B.8: Statistics of error in peak location estimation error for LSF region 8. τ_8 is selected as 2.9.

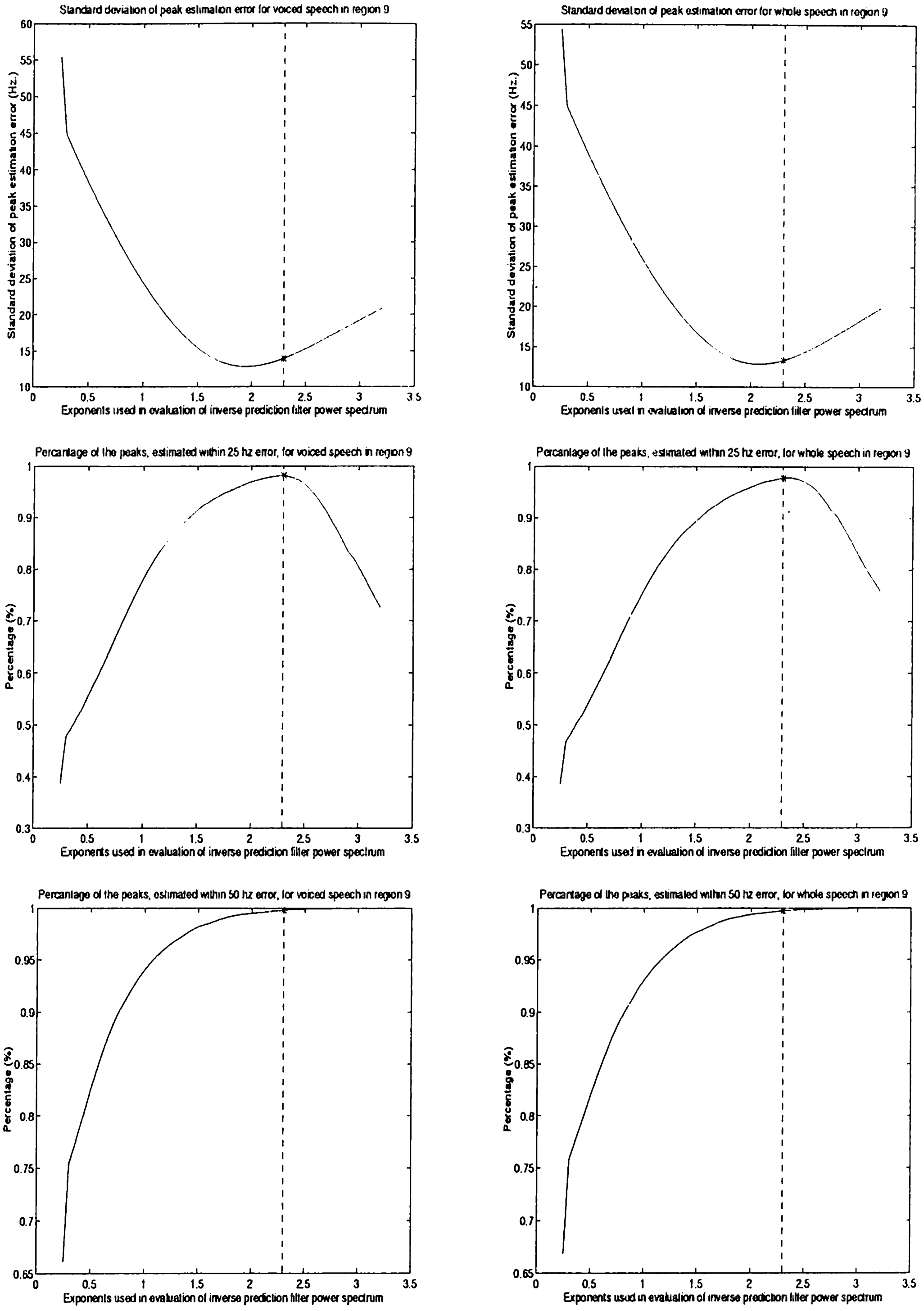


Figure B.9: Statistics of error in peak location estimation error for LSF region 9. τ_9 is selected as 2.3.

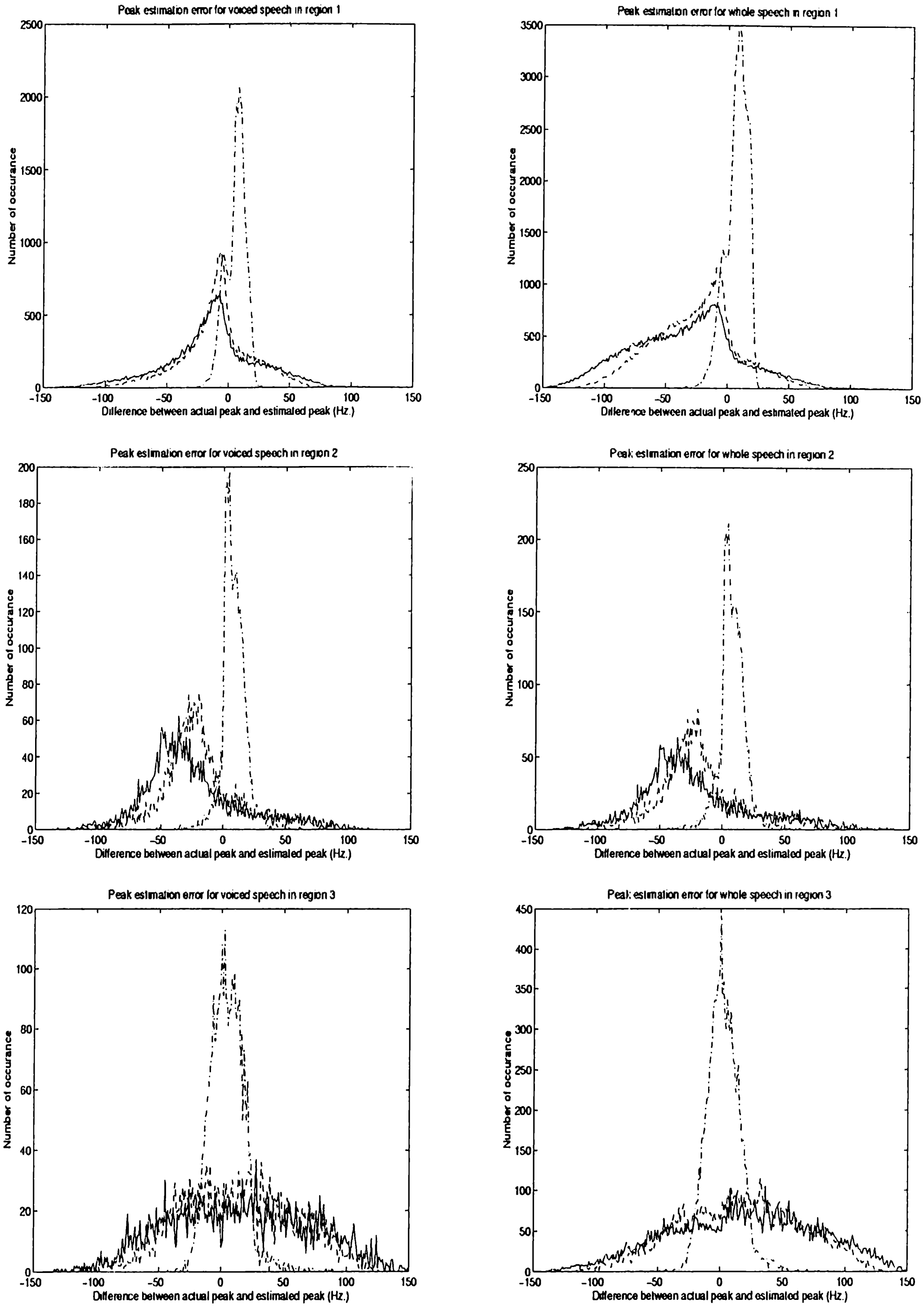


Figure B.10: Number of occurrence versus difference between original and estimated peaks for the LSF regions 1, 2 and 3. First and second column presents statistical data about voiced frames and all frames, respectively. The solid, dashed and dashed-dotted lines corresponds to simple mean, weighted mean with $\tau = 0.15$ and weighted mean with $\tau = \tau_i$, respectively. $\tau_1 = 2.5$, $\tau_2 = 2.6$ and $\tau_3 = 2.25$.

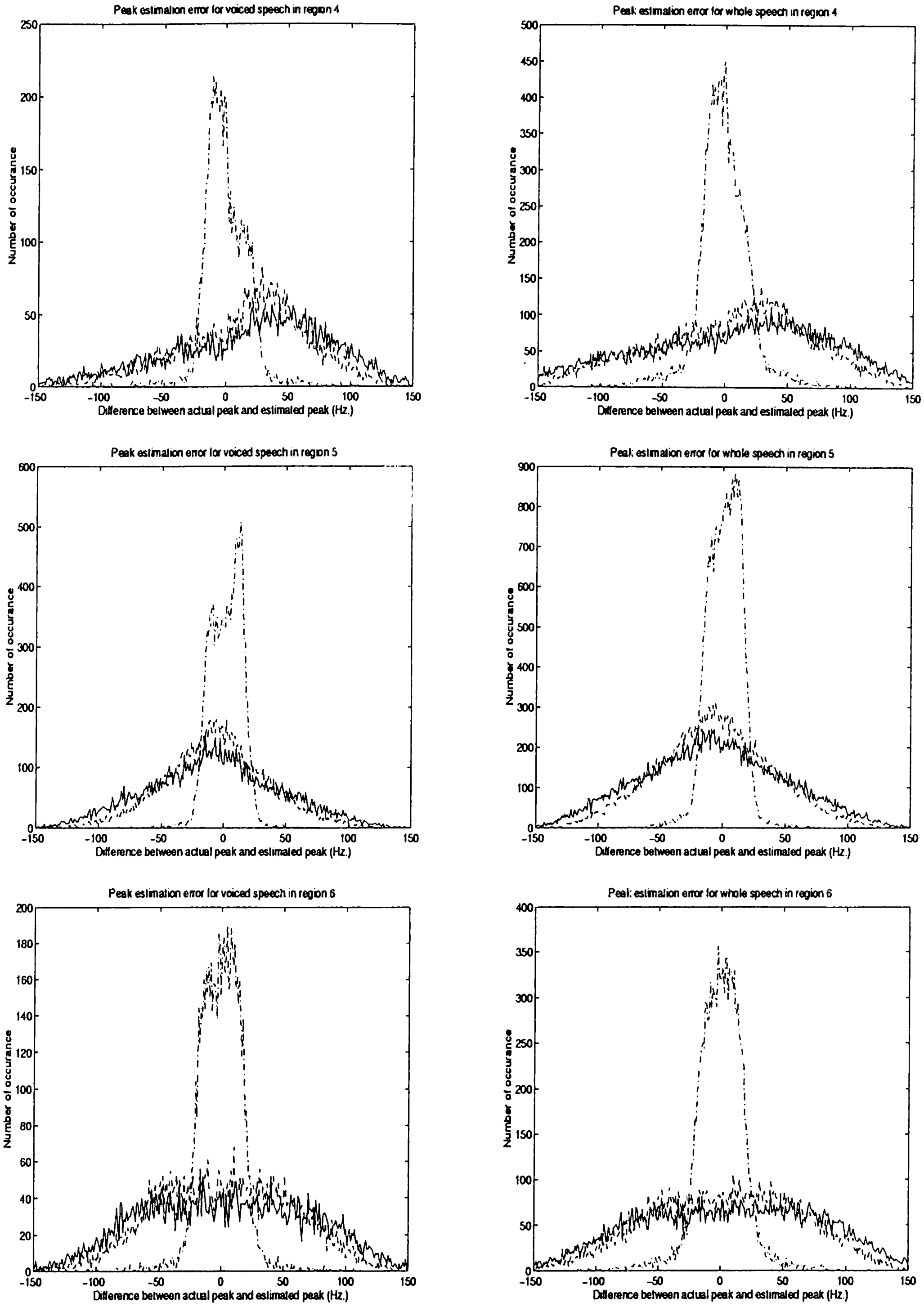


Figure B.11: Number of occurrence versus difference between original and estimated peaks for the LSF regions 4, 5 and 6. The solid, dashed and dashed-dotted lines corresponds to simple mean, weighted mean with $\tau = 0.15$ and weighted mean with $\tau = \tau_i$, respectively. $\tau_4 = 2.6$, $\tau_5 = 2.35$ and $\tau_6 = 3.0$.

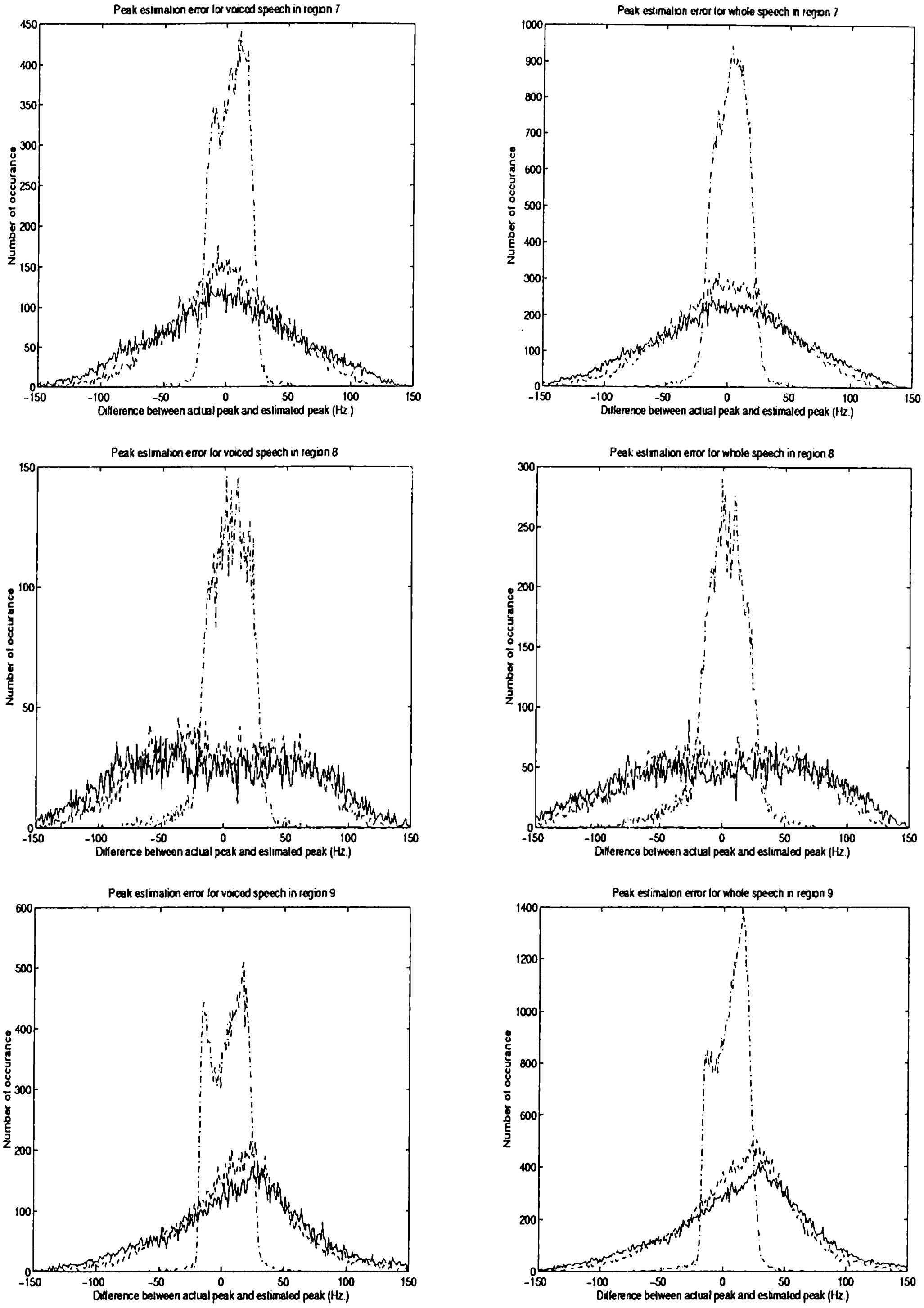


Figure B.12: Number of occurrence versus difference between original and estimated peaks for the LSF regions 7, 8 and 9. The solid, dashed and dashed-dotted lines corresponds to simple mean, weighted mean with $\tau = 0.15$ and weighted mean with $\tau = \tau_i$, respectively. $\tau_7 = 2.35$, $\tau_8 = 2.9$ and $\tau_9 = 2.3$.

Bibliography

[1] L.R. Rabiner, "Applications of voice processing to telecommunications," *Proceedings of IEEE*, vol. 82, 1994.

[2] J.P. van Hemert, "Automatic segmentation of speech," *IEEE Trans. Signal Process.*, vol. 39, pp. 1008–1012, 1991.

[3] A. Das, E. Paksoy, and A. Gersho, *Speech Coding and Synthesis*, chapter 7: Multimode and Variable-Rate Coding of Speech, Elsevier, 1995.

[4] L.F. Lamel, L.R. Rabiner, and A.E. Rosenberg, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 29, pp. 777–785, 1981.

[5] Ta-Hsin Li and Jerry D. Gibson, "Speech analysis and segmentation by parametric filtering," *IEEE Trans. Speech, Audio and Signal Process.*, vol. 4, pp. 203–213, 1996.

[6] Regine Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 36, pp. 29–40, 1988.

[7] R.J. Di Francesco, "Real-time speech segmentation using pitch and convexity jump models: Applications to variable rate speech coding," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 38, pp. 741–748, 1990.

[8] D.B. Grayden and M.S. Scordilis, "Phonetic segmentation of fluent speech," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. I73–I76, 1994.

[9] B.J. Pawate, "A new method for segmenting continuous speech," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. I53–I56, 1994.

[10] A. Ljolje and M.D. Riley, "Automatic segmentation and labelling of speech," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 473–476, 1991.

[11] S. Krishnan and P.V.S. Rao, "Segmental phoneme recognition using piecewise linear regression," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. I49–I52, 1994.

[12] V. Zue, J. Glass, M. Phillips, and S. Seneff, "Acoustic segmentation and phonetic classification in the SUMMIT system," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 389–392, 1989.

[13] Mustafa Bayindir, "Automatic segmentation and labeling of isolated turkish words," M.S. thesis, Middle East Technical University, 1997.

[14] P. Jeanrenaud and P. Peterson, "Segment vocoder based on reconstruction with natural segments," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 605–608, 1991.

[15] R.M. Gray, A. Buzo, A.H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 28, pp. 367–376, 1980.

[16] J.S. Erkelens and P.M.T. Broersen, "On the statistical properties of line spectrum pairs," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 768–771, 1995.

[17] "Specifications for the analog to digital conversion of voice by 2,400 bit/second mixed excitation linear prediction," 1998.

[18] Monson H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, Inc., 1996.

[19] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signal," *Journal of Acoustical Society of America.*, p. 535, 1975.

[20] R.P. Cohn and J.S. Collura, "Incorporating perception into lsf quantization - some experiments," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1347–1350, 1997.

[21] H.L. Vu and L. Lois, "Spectral sensitivity of lsp parameters and their transformed coefficients," *Proc. European Specch Comm. Technology, EUROSPEECH*, pp. 1251–1254, 1997.

[22] P.E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice-Hall, 1987.

[23] K.K. Paliwal, "On the use of line spectrum freqeucies parameters for speech recognition," *Digital Signal Processing*, vol. 2, pp. 80–87, 1992.

[24] A.V. McCree and J.C. De Martin, "A 1.7 kb/s MELP coder with improved analysis and quantization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 593–596, 1998.

[25] N. Sugamura, "Quantizer design in lsp speech analysis-synthesis," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 432–440, 1988.

[26] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 661–664, 1991.

[27] F.K. Soong and B.H. Juang, "Optimal quantization of lsp parameters using delayed decisions," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 185–188, 1990.

[28] M. Xie and J.P. Adoul, "Fast and low-complexity lsf quantization using algebraic vector quantizer," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 716–719, 1995.

[29] D. Chang, Cho Y, and S. Ann, "Efficient quantization of lsf parameters using classified svq and combined with conditional splitting," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 736–739, 1995.

[30] K.K Paliwal and B.S Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE Trans. Speech, Audio and Signal Process.*, vol. 1, pp. 3–14, 1993.

[31] S. Nandkumar, K. Swaminatham, and U. Bhaskar, "Robust speech mode based lsf vector quantization for low bit rate coders," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 41–44, 1998.

[32] C.S. Xydeas and C. Papanastasiou, "Efficient coding of lsp parameters using split matrix quantization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 740–743, 1995.

[33] C.C. Kuo, F.R. Jean, and H.C. Wang, "Low bit-rate quantization of lsp parameters using two-dimensional differential coding," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. I97–I100, 1992.

[34] E. Erzin and A.E. Çetin, "Interframe differential vector coding of line spectrum frequencies," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. II25–II28, 1993.

[35] C.W. Seymour and A.J. Robinson, "A low bit-rate speech coder using adaptive line spectral frequency prediction," *Proc. European Specch Comm. Technology, EUROSPEECH*, pp. 1319–1322, 1997.

[36] Y.K. Lee, K.C. Kim, and H.S. Lee, "An efficient coding of lsp parameters using multiple type frame segmentation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 753–756, 1995.

[37] A.N. Lemma, W.B. Kleijn, and E.F. Deprettere, "Lpc quantization using wavelet based temporal decomposition of the lsf," *Proc. European Specch Comm. Technology, EUROSPEECH*, pp. 1259–126, 1997.

[38] A.M. Kondoz, *Digital Speech*, John Wiley & Sons Ltd., 1994.

[39] F.K. Soong and B.H. Juang, "Line spectrum pairs (lsp) and speech data compression," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1.10.1–1.10.4, 1984.

[40] P. Kabal and R.P. Ramachandran, "The computation of line spectrum frequencies using chebyshev polynomials," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 34, pp. 1419–1425, 1986.

[41] C.H. Wu and J.H. Chen, "A novel two-level method for the computation of the lsp frequencies using a decimation-in-degree algorithm," *IEEE Trans. Speech, Audio and Signal Process.*, vol. 5, pp. 106–115, 1997.

[42] A. Goalic and S. Saoudi, "An intrinsically reliable and fast algorithm to compute the line spectrum pairs(lsp) in low bitrate celp coding," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 728–731, 1995.

[43] B.M.G. Cheetham, "Adaptive lsp filter," *Electronics Letters*, vol. 23, pp. 89–90, 1986.

[44] H.J. Coetzee and T. P. Barnwell III, "An lsp based speech quality measure," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 596–599, 1989.

[45] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *Bell Labs. Technical Journal*, vol. 47, pp. 634–647, 1970.

[46] H.L Vu and L. Lois, "A new general distance measure for quantization of lsf and their transformed coefficients," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 45–48, 1998.

[47] Bishnu S. Atal, Vladimir Cuperman, and Allen Gersho, *Advances in speech coding*, Boston : Kluwer Academic Publishers, 1991.

[48] S. Roucos, R. M. Schwartz, and J. Makhoul, "A segment vocoder at 150 b/s.," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 61–64, 1983.

[49] A. Gersho, "Advances in speech coding and audio compression," *Proceedings of IEEE*, vol. 82, 1994.

[50] A.S. Spanias, "Speech coding: A tutorail review," *Proceedings of IEEE*, vol. 82, 1994.

[51] A.V. McCree and T.P. Barnwell III, "A mixed excitation lpc vocoder model for low bit-rate speech coding," *IEEE Trans. Speech, Audio and Signal Process.*, vol. 3, pp. 242–250, 1995.

[52] V. Cuperman, B.S. Atal, and A. Gersho, *Advances in Speech Coding*, Kluwer Academic Publishers, 1991.

[53] S.V. Vaseghi, "Finite state celp for variable rate speech coding," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 37–40, 1990.

[54] L. Cellario, M. Giani, P. Blocher, D. Sereno, and K. Hellwig, "A vr-celp codec implementation for cdma mobile communications," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. I281–I284, 1994.

[55] L.A.H. Gomez, F.J.C. Quiros, C.G. Mateo, and J.O. Garcia, "Real-time implementation and evaluation of variable rate celp coders," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 585–588, 1991.

[56] R.D. De Iacovo and D. Sereno, "Embedded celp coding for variable bit-rate between 6.4 and 9.6 kbit/s," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 681–684, 1991.

[57] P. Lupini, N.B. Cox, and V. Cuperman, "A multi-mode variable rate celp coder based on frame classification," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 406–409, 1993.

[58] S. A. McClellan and J.D. Gibson, "Variable rate celp based on subband flatness," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1409–1413, 1995.

[59] P. Kroon and M. Recchione, "A low-complexity toll-quality variable bit-rate coder for cdma cellular systems," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5–8, 1995.

[60] R. Peng and V. Cuperman, "Variable-rate low-delay analysis-by-synthesis speech coding at 8-16 kb/s," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 29–32, 1991.

[61] R.D. Francesco, C. Lamblin, A. Leguyader, and D. Massaloux, "Variable rate speech coding with online segmentation and fast algebraic codes," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 233–236, 1990.

[62] S. Wang and A. Gersho, "Improved phonetically-segmented vector excitation at 3.4 kb/s," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. I349–I352, 1992.

[63] A. Shen, B. Tang, A. Alwan, and G. Pottie, "A robust variable-rate speech coder," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 249–252, 1995.

[64] E. Paksoy, A. McCree, and V. Viswanathan, "A variable-rate multimodal speech coder with gain matched analsis-by-synthesis," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 751–754, 1997.

[65] E.W.M. Yu and C.F. Chan, "Variable bit-rate mbelp speech coding via v/uv distribution dependent spectral quantization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1607–1610, 1997.

[66] S. Villette, M. Stefanovic, I. Atkinson, and A.M. Kondoz, "High quality split band lpc vocoder and its fixed point real time implementation," *Proc. European Specch Comm. Technology, EUROSPEECH*, pp. 1243–1246, 1997.

[67] S.Y. Kwon and A.J. Goldberg, "An enhanced lpc vocoder with no voiced/unvoiced switch," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, pp. 851–858, 1984.

[68] A.V. McCree and T.P. Barnwell III, "Improving performance of a mixed excitation lpc vocoder in acoustic noise," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 137–140, 1992.

[69] A.V. McCree and T.P. Barnwell III, "A new mixed excitation lpc vocoder," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 593–596, 1991.

[70] J.H. Chen and A. Gersho, "Adaptive postfilter for quality enhancement of coded speech," *IEEE Trans. Speech, Audio and Signal Process.*, vol. 3, pp. 59–71, 1995.

[71] P.A. Laurent and P. de La Nove, "A robust 2400 bps subband lpc coder," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 500–503, 1995.

[72] D.W. Griffin and J.S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 36, pp. 1223–1235, 1988.

[73] J. Makhoul, R. VisWanathan, R. Schwartz, and A.W.F. Huggins, "A mixed-source model for speech compression and synthesis," *Journal of Acoustical Society of America.*, vol. 64, pp. 1577-1581, 1978.

[74] B.S. Atal M.R Schroder and J.L Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of Acoustical Society of America.*, vol. 66, pp. 1647–1652, 1979.

[75] A.V. McCree and T.P. Barnwell III, "Implementation and evaluation of a 2400 bps mixed excitation lpc vocoder," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 159–163, 1993.

[76] C.F. Chan and E.W.M. Yu, "Frequency domain postfiltering for multiband excited linear predictive coding of speech," *Electronics Letters*, vol. 32, pp. 1061–1063, 1996.

[77] G.S. Kang and S.S. Everett, "Improvement of the excitation source in the narrowband linear prediction vocoder," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 33, pp. 377–386, 1985.

[78] M.R. Sambur, A.E. Rosenburg, L.R. Rabiner, and C.A. McGonegal, "On reducing the buzz in lpc synthesizer," *Journal of Acoustical Society of America.*, vol. 63, pp. 918–924, 1978.

[79] A.E. Rosenburg, "Effects of glottal pulse shape on the quality of natural vowels," *Journal of Acoustical Society of America.*, vol. 49, pp. 583–590, 1971.

[80] L.M. Supplee, R.P. Cohn, J.S. Collura, and A.V. McCree, "MELP: The new federal standard at 2400 bps," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1591–1594, 1997.

[81] M.A. Kohler, "A comparison of the new 2400 bps MELP federal standard with other standard coders," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1587–1590, 1997.

[82] E.S. Dermetas, N.D. Fakotakis, and G.K. Kokinakis, "Fast endpoint detection algorithm for isolated word recognition in office environment," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 733–736, 1991.

[83] H. Kobatake, K. Tawa, and A. Ishida, "Speech/nonspeech discrimination for speech recognition systems under real life noise environment," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 365–368, 1989.

[84] L.R. Rabiner and M.R. Sambur, "An algorithm for determining endpoints of isolated utterances," *Bell Labs. Technical Journal*, vol. 54, pp. 297–315, 1975.

[85] G.S. Ying, C.D. Mitchell, and L.H. Jamieson, "Endpoint detection of isolated utterances based on a modified teager energy measure," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 732–735, 1993.

[86] Engin Erzin, *New Methods for Robust Speech Recognition*, Ph.D. thesis, Bilkent University, 1995.

[87] Y. Yatsuzuka, "Highly sensitive speech detector and high-speed voiceband data discriminator in dsi-adpcm systems," *IEEE Trans. on Communications*, vol. 30, pp. 739–750, 1982.

[88] P.T. Brady, "A technique for investigating on-off patterns of speech," *Bell Labs. Technical Journal*, vol. 44, pp. 1–22, 1965.

[89] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 369–372, 1989.

[90] E. Paksoy, K. Srinivasan, and A. Gersho, "Variable rate speech coding with phonetic segmentation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 155–158, 1993.

[91] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 365–368, 1998.

[92] A. Cavallaro, F. Beritelli, and S. Casale, "A fuzzy logic-based speech detection algorithm for communications in noisy environments," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 565–568, 1998.