ABSTRACT

Title of Document: MAXIMUM LIKELIHOOD PITCH ESTIMATION USING SINUSOIDAL MODELING

Vijay Mahadevan, Master of Science, 2010

Directed By: Dr. Carol Y. Espy-Wilson
Department of Electrical and Computer Engineering

The aim of the work presented in this thesis is to automatically extract the fundamental frequency of a periodic signal from noisy observations, a task commonly referred to as pitch estimation. An algorithm for optimal pitch estimation using a maximum likelihood formulation is presented. The speech waveform is modeled using sinusoidal basis functions that are harmonically tied together to explicitly capture the periodic structure of voiced speech. The problem of pitch estimation is casted as a model selection problem and the Akaike Information Criterion is used to estimate the pitch. The algorithm is compared with several existing pitch detection algorithms (PDAs) on a reference pitch database. The results indicate the superior performance of the algorithm in comparison with most of the PDAs. The application of parametric modeling in single channel speech segregation and the use of mel-frequency cepstral coefficients for sequential grouping are analyzed in the speech separation challenge database.

MAXIMUM LIKELIHOOD PITCH ESTIMATION USING SINUSOIDAL MODELING

By

Vijay Mahadevan

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2010

Advisory Committee:
Dr. Carol Y. Espy-Wilson, Chair/Advisor
Dr. Rama Chellappa
Dr. Min Wu

Acknowledgements

I would like to thank my mother, Kulakodi Arthanari and father, Tarrakad Vaidyanathan Mahadevan for their love and support. I want to thank my brother, Karthikeyan Mahadevan and my sister-in-law, Veena Adityan for their support and guidance. Karthik has always been there for me at every step.

I am very grateful to my advisor, Prof. Carol Y. Espy-Wilson for her guidance and support. I wish to express my immense gratitude to her for giving me a wonderful opportunity to learn. She has been a great source of inspiration and an excellent mentor. I am extremely privileged to have her as my advisor.

I would like to express my gratitude to the members of thesis committee, Prof Rama Chellappa and Prof Min Wu for sparing their invaluable time in reviewing the manuscript.

Many thanks to all the members of the Speech Communication Lab: Srikanth, Daniel, Vikramjit, Xinhui, Tarun and Jing Ting. Srikanth has been a second mentor who helped me with numerous technical discussions and comments. I would like to thank my roommates Shalabh, Nitesh, Ramaswamy, Ishaan and Ashish for their support. Special thanks to Karthik Ravirajan, Ashwin Swaminathan and Balaji Vasan for their invaluable help in numerous occasions.

Last but not least, thanks to Jayashree K. Seshadri for her support and understanding.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Background

1.1 Objective

The aim of the work presented in this thesis is to automatically extract the fundamental frequency of a periodic signal from noisy observations, a task commonly referred to as pitch estimation. A Fourier series is a decomposition of a periodic signal into a sum of a set of simple oscillating functions, namely sines and cosines or complex exponentials. These sinusoids all repeat over the same interval, meaning they have frequencies which are integer multiples of a fundamental frequency. This thesis is about the estimation of the fundamental frequency for speech signals through parametric signal modeling. We also study some novel applications of the signal model for challenging problems like speech enhancement and speech segregation.

1.2 Pitch Vs Fundamental Frequency

The terms fundamental frequency and pitch will be used interchangeably in this thesis, although there is a fine distinction between the two. The former is a mathematical term that describes the periodicity in the signal whereas pitch can be thought of as a perceived fundamental frequency. The American National Standards defines the term pitch as “that attribute of the auditory sensation in terms of which sounds may be ordered on a scale extending from low to high”. Hence the term pitch has more to do with auditory sensation than the physical attribute of the signal. In spite of this difference, throughout this thesis we will use pitch synonymously with

fundamental frequency to refer to the physical attribute of the signals associated with the Fourier series.

1.3 Pitch Detection Algorithms (PDA)

The algorithms that aim at extracting the pitch are referred to as pitch detection algorithms (PDA). The problem of pitch estimation can be viewed differently from pitch detection which is a hypothesis testing problem. In most of the work presented we do not make any distinction about pitch detection and estimation. Naturally, one has to decide if the observed speech signal is voiced (periodic) or unvoiced (aperiodic) and if it is voiced, then we need to estimate the pitch. Hence both these terms are tied together in a generic PDA. Typically, pitch determination requires a search of different possible candidate frequencies over an analysis window. A cost function is defined for every pitch candidate and the estimated frequency is chosen to be the one that gives an optimum cost. We will now briefly discuss some of the popular techniques for pitch detection (Christensen et al., 2008; Hess, 1983; Rabiner, 1976).

1.4 Non-parametric Methods

There exists many non-parametric methods, based on, for example, the autocorrelation, cross-correlation, averaged magnitude difference function or the cepstrum. Most of these methods define the cost function to measure some sense of similarity of the signal and its delayed version. For example, the autocorrelation based pitch detector can be formally viewed as minimizing, over possible pitch

periods the mean squared error between the signal and its delayed version. It is essentially a measure of self-similarity and we expect to observe peaks near the actual period. Another example of a non-parametric method is the harmonic product spectrum. All these methods suffer from a common problem of non-uniqueness in pitch estimation even in the ideal case i.e., there exists multiple lags for which the signal is similar to itself. A detailed study of various non-parametric approaches is presented in Hess, 1983.

1.5 Human Auditory System models

Another class of methods that estimate pitch frequency is based on models of the human auditory system. Instead of taking their starting point in the properties of the signal, these methods are based on the properties of the human ear and brain. The motivation is that the human auditory system has a remarkable property of identifying multiple pitches simultaneously, and separates various sources despite the background noise. The hope is that by mimicking the auditory signal processing, we can design a system that works as well as humans. For examples of such methods, references therein and overview of all things related to pitch perception are discussed in Plack et al., 2005.

1.6 Parametric Models

The objective of the above methods is quite different from the method presented in this thesis. We are concerned with finding the parameters that are most likely to explain the observed signal and this is generally a different concept altogether than

modeling the peculiarities of the auditory system. In parametric approach towards pitch estimation, a signal model is proposed which aim at explaining the observation with few finite parameters. In particular, we would like to infer from the observation the parameters of the model. R.A. Fisher (1922) discussed three aspects of the general problem of valid inference: (1) model specification, (2) estimation of model parameters, and (3) estimation of precision. The model specification is partitioned into two components: formulation of a set of candidate models and selection of a model to be used in making inferences. Among the statistical parametric estimation methods, the two philosophies namely Maximum likelihood (ML) and Maximum a posteriori (MAP) methods are analyzed. The parametric models discussed are based on sinusoidal modeling of the observed signal. In particular, we present the work where the parameters are assumed to be fixed on the duration of the signal that is analyzed. An ML estimation framework is presented for estimating the parameters of the model. In order to estimate the fundamental frequency, the Akaike information criterion (AIC) is applied to regularize the parameter estimation process. A closely related work towards MAP estimation for pitch tracking is presented in Tabrikian et al., 2004. The prior is imposed on the fundamental frequencies which are assumed as a Markov sequence and the MAP estimation of the fundamental frequency is implemented using a dynamic programming procedure. In this work we do not assume any distribution on the transition probability density function (pdf) of the fundamental frequencies and each frame operates independently.

1.7 Thesis Outline

The parametric modeling of the speech signal and its application towards pitch estimation is presented in Chapter 2. The major contribution in this chapter is the optimal estimation of fundamental frequency through ML formulation and AIC model selection framework. Perhaps the earliest work which approximates the speech signal by a finite Fourier series is the PDA by Steiglitz et al., 1975. However, the algorithm was analyzed in a very limited setting for one male speaker and there was no extensive results reported across a database. The problem of pitch estimation using the signal models typically suffer from over fitting using ML methods Wise et al., 1976 which is brought to attention in this chapter. A detailed analysis of the proposed algorithm with results from a publicly available pitch database is presented.

Chapter three discusses the important applications of parametric modeling described in this thesis. There are two major applications that will be discussed highlighting the potential and use of this method. The major contribution in this chapter is in signal separation by using regularized least squares method.

Chapter four extends the analyses towards a single channel speech segregation system and the objective is to track multiple pitch frequencies across time. This problem is called sequential grouping in co-channel speech (Wang & Brown, 2006). In order to achieve meaningful separation of the speech signals using the pitch frequencies, it is desired to group the speech that belongs to the target speaker into one stream and the masker in to another stream. A detailed analysis is presented on the problem with

results using different features that are used for grouping. The primary contribution in this chapter is the use of the Mel-frequency cepstral coefficients (MFCCs) to perform sequential grouping.

Chapter five presents the conclusion and directions for future work. The general framework of parametric modeling has extensive applications and some of these were presented in Chapter four. Interesting ideas on future directions for improvement in pitch estimation and signal enhancement are presented in this chapter.

Chapter 2: Pitch Detection Algorithm

The statistical method for pitch tracking presented in this work (Mahadevan & Espy-Wilson, 2011) can be viewed as a generalization of the discrete Fourier transform representation. It is also a special case of a sinusoidal speech model where all the sinusoidal components are assumed to be harmonically related, i.e. integer multiples of the fundamental frequency. The system outputs a pitch estimate for every frame that is detected to be voiced. We follow a metric that estimates the local signal to noise ratio (SNR) and decide on the voicing probability (Quatieri, 2002). The voice activity detection is an integral part of the algorithm which is measured by the goodness of the model fit to the observation. The statistical method for pitch tracking presented in this work follows the maximum likelihood estimation of the parameters. We follow a regression framework and decide on the pitch frequency using the Akaike Information Criteria (Burnham & Anderson, 2002). We consider three principal parts of the mathematical model i.e. the conceptual, analytic and computational aspects in sections 2.1 through 2.5. The voicing detection block is outlined in section 2.6. A description of the database used in the evaluation is presented in section 2.7 and the performance comparison with several existing PDAs is discussed in section 2.8.

2.1 Motivation

For a stationary speech signal, pitch can be defined as the perception of a fundamental frequency of a pure harmonic template which optimally fits a successive harmonic component pattern of the speech signal (Goldstein, 1973). We follow a signal model

that explicitly captures the periodic structure of the speech signal. This approach towards estimating pitch is referred as Harmonic Structure Matching Pitch Estimation (HSMPE) in Gong and Haton, 1987. In our work, we explicitly model the time domain signal using sinusoidal basis functions that are harmonically tied together.

2.2 Mathematical Formulation

We start with the basic Fourier series representation of a stationary periodic signal. The windowed speech waveform is represented by a sum of sinusoidal functions with fixed amplitudes, frequencies and phases (McAulay & Quatieri, 1986). This approach can be viewed as a generalization of the discrete Fourier transform, i.e. the period of the signal is arbitrary and not necessarily equal to the length of the signal. This framework was used in (Arruda, 2010) the name of regressive discrete Fourier series and it is well known in the statistical literature as least squares spectral analysis. Under this condition, the windowed speech signal *s[n]* is represented as,

$$\mathrm{s[n]} = \sum_{k=1}^{M(f_0)} (a_k \cos(2\pi f_0 kn + \varphi_k)) + \varepsilon[\mathrm{n}] \qquad (2.1)$$

where $1 \leq n \leq N$, a_k, φ_k, f_0 and $M(f_0)$ represent the amplitude, phase, fundamental frequency and the number of harmonics respectively and *ε[n]* represents the residual error from the model. Equation 1 can be compactly written in matrix form as,

$$\mathbf{s} = \mathbf{A}\,(\mathrm{f_0}) * \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \qquad (2.2)$$

$$\boldsymbol{A}(f_0) = \begin{pmatrix} e^{-1i\omega_0} & e^{-1i2\omega_0} \dots & e^{-1iM\omega_0} \\ e^{-ni\omega_0} & e^{-ni2\omega_0} \dots & e^{-niM\omega_0} \\ e^{-Ni\omega_0} & e^{-Ni2\omega_0} \dots & e^{-NiM\omega_0} \end{pmatrix}$$

where the matrix **A** contains complex exponentials at the multiples of $\omega_0 = 2\pi f_0$ and is of size NxM(f_0). The harmonic amplitude and phase information is captured in $\boldsymbol{\gamma}$. The residual error is assumed to be additive white Gaussian noise with zero mean and covariance matrix $R = \sigma^2 I$. Hence the unknown parameters in the model are f_0, $\boldsymbol{\gamma}$ and σ^2 which we wish to estimate from the observed signal.

2.3 Maximum Likelihood Estimation

The likelihood of observing the data given the parameters is,

$$g(\boldsymbol{s}|f_0, \sigma^2, \boldsymbol{\gamma}) \sim N(\boldsymbol{A}(f_0) * \boldsymbol{\gamma}, \sigma^2 \boldsymbol{I}) \tag{2.3}$$

and the log-likelihood function L($\boldsymbol{\theta}$) with $\boldsymbol{\theta} = [\sigma^2, \gamma, f_0]$ containing all the unknown parameters is given by,

$$L(\sigma^2, \boldsymbol{\gamma}, f_0) = \frac{N}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}([\boldsymbol{s} - \boldsymbol{A}(f_0)\boldsymbol{\gamma}]^H[\boldsymbol{s} - \boldsymbol{A}(f_0)\boldsymbol{\gamma}]) \tag{2.4}$$

The maximum likelihood parameter estimate is found by maximizing (4),

$$\widehat{\boldsymbol{\theta}} = \arg max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}) \tag{2.5}$$

The log-likelihood function is non-linear in f_0 and the usual optimization methods will yield local maxima. However, the parameter space for f_0 is restricted to the possible pitch frequency for humans and therefore we do a global brute force approach for estimating f_0. To do so, we fix $f_0 = f_0$' and observe that the optimization problem is quadratic in $\boldsymbol{\gamma}$ and the solution is given by Moore-Penrose pseudo inverse of $\mathbf{A}(f_0')$ denoted as $\boldsymbol{A}^{+}(f_0') = (\mathbf{A}(f_0')^{\mathbf{T}} * \mathbf{A}(f_0'))^{-\mathbf{1}} * \mathbf{A}(f_0')$. The well known optimal estimates is noted below for $\boldsymbol{\gamma}$ and σ^2,

$$\widehat{\boldsymbol{\gamma}} = \boldsymbol{A}^{+}(f_0') * \boldsymbol{s} \tag{2.6}$$

$$\widehat{\sigma^2} = [s - \mathbf{A}(f_0') * \widehat{\boldsymbol{\gamma}}]^{\mathbf{T}}[s - \mathbf{A}(f_0') * \widehat{\boldsymbol{\gamma}}]/\boldsymbol{n} \tag{2.7}$$

The estimated signal $\hat{s}$ is given by the projection of the observation on the space spanned by the columns of $\mathbf{A}(f_0')$,

$$\hat{\boldsymbol{s}} = P_{\mathbf{A}(f_0')} * \mathbf{s} \tag{2.8}$$

$$P_{\mathbf{A}(f_0')} = \mathbf{A}(\mathbf{A}^{\boldsymbol{T}}\mathbf{A})^{-\mathbf{1}}\boldsymbol{A}|_{\boldsymbol{f_0}'} \tag{2.9}$$

The maximized value of the log-likelihood function ignoring the additive constants is then given by,

$$L(\hat{\theta}) = \frac{N}{2} \ln \left(\frac{1}{\widehat{\sigma^2}}\right) \tag{2.10}$$

The problem formulation is reduced to minimizing the residual sum of squares. The column space of $\mathbf{A}$ $(\frac{f_0}{2})$ is a superset of $\mathbf{A}$ (f_0) and therefore the residual error variance will follow $\widehat{\sigma^2}_{f_0/2} \leq \widehat{\sigma^2}_{f_0}$. It can be seen that choosing f_0 that maximizes $\mathbf{L}(\theta)$ in (10) will result in pitch halving error almost always when $\frac{f_0}{2}$ is in the parameter space. This should come as no surprise as we are simply doing a regression on the data using different models indexed by f_0. Therefore we need a tradeoff on the number of parameters used to describe the model, i.e. the complexity of the model and the goodness of fit from the model. This is achieved using the AIC described in the next section.

2.4 Model Selection

The AIC model selection stems from the Kullback- Leibler (K-L) information loss (Rao et al., 2008; Burnham & Anderson, 2002) to choose the best model from a set of candidates. In our case, the different models are indexed by the fundamental frequency. The tradeoff between the model complexity and the goodness of fit as given by AIC is,

$$AIC(model) = -2 * \left(Maximized\ value\ of\ the\ likelihood \Big/ model\right) + 2 * Number\ of\ parameters\ in\ the\ model \tag{2.11}$$

$$AIC(f_0) = N\ln\left(\widehat{\sigma^2}(f_0)\right) + 2 * M(f_0) \quad (2.12)$$

We have the maximized log-likelihood value using the templates of projection matrices indexed by f_0. The number of parameters in the model is equal to the number of regressors used, i.e. the dimension of the harmonic coefficients M(f_0). We choose the f_0 that gives the lowest AIC score. A scenario illustrating the pitch halving error through ML model selection which is corrected using AIC information criteria is show in Figure 2.1. The algorithm provides high resolution in estimating the pitch frequency as we are not restricted to work with integer periods with resolution dictated by the sampling interval. The effect of pitch resolution in computational complexity is analyzed in the following section.

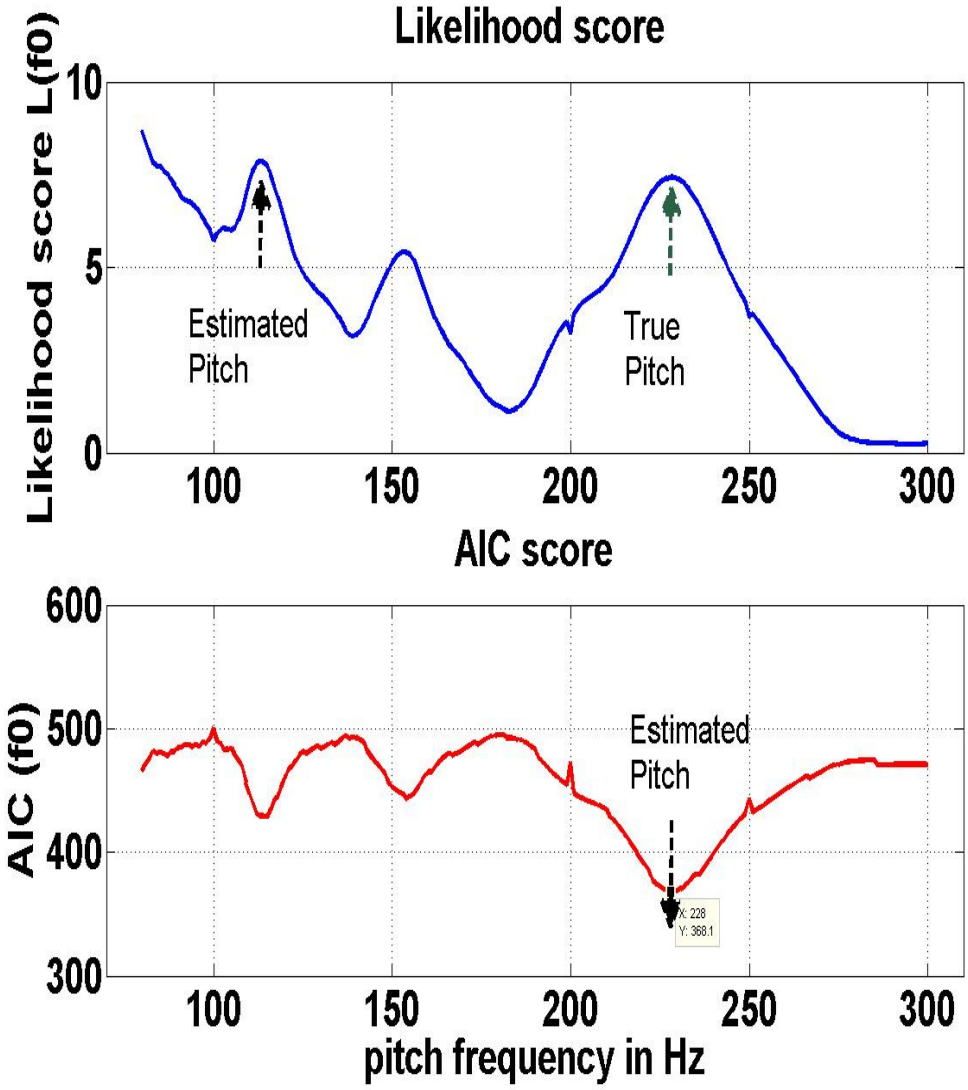


Figure 2.1: Illustration of pitch halving error using maximum likelihood pitch estimation

2.5 Computational Complexity

In the problem of pitch estimation we are essentially solving a system of linear equations through projection templates. The storage complexity of these templates requires a memory space of the order (Big- O notation) $O(T * N^2)$ where T denotes the cardinality of the f_0 parameter search space. The number of computations done per

candidate model is $O(N^2)$ and therefore for T models we have a total of $O(T * N^2)$. It should be noted that other minimum mean squared error methods based on similarity measures like the autocorrelation and the Average Magnitude Difference Function (AMDF) require $O(N)$ for every candidate pitch period (brute force approach) and therefore a total of $O(T * N)$ computational load.

The algorithm can be easily scaled to meet the computational requirements with a tradeoff on the accuracy of the pitch estimates. By computing the pitch frequency in the first voiced frame, gradient search techniques can be used to estimate the fundamental frequency in the successive frames. There can be various strategies to efficiently search the pitch grid starting from a coarse resolution and then tuning it to a finer resolution according to the required level of accuracy. Figure 2.2 illustrates the computational time required to process a signal of length 1.35*s* sampled at 8 kHz at 10*ms* frame rate in 3GHz Intel processor. The computational time further scales with the sampling frequency of the signal. If we down sample the signal by a factor of L, the computational complexity scales by a factor of L^2, i.e. the load for T models is $O\left(T * \left(\frac{N}{L}\right)^2\right)$.

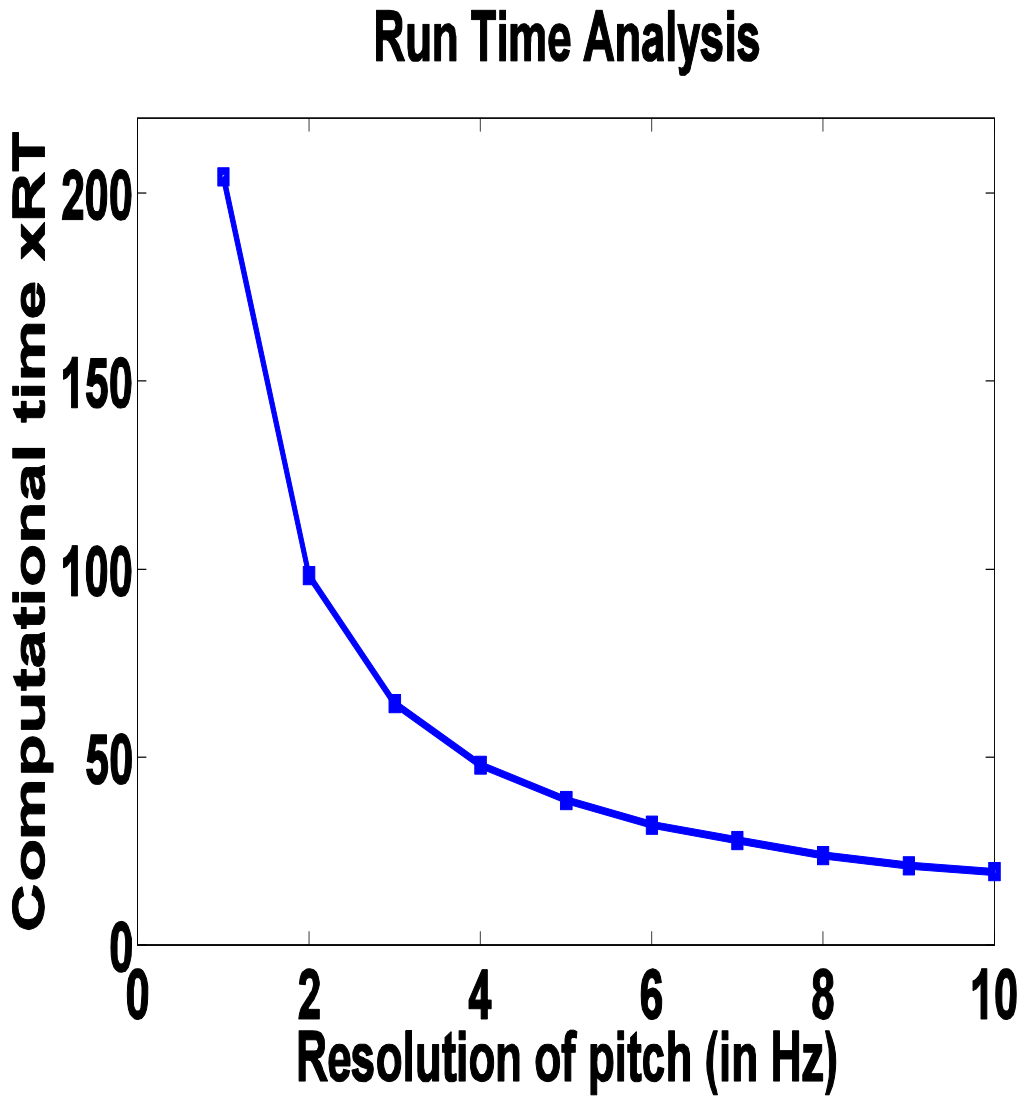


Figure 2.2: Run time analysis of the pitch detection algorithm as a function of pitch resolution

2.6 Voice Activity Detection

Voice activity detection is an integral part of the algorithm which is measured by the goodness of the model fit to the observation. The estimated speech signal $\hat{s}$ and the residual ε can be used to arrive at a measure of local SNR as follows,

$$\boldsymbol{\varepsilon = s - \hat{s}} \tag{2.13}$$

$$SNR = 10log_{10}\left(\frac{\hat{s}^T\hat{s}}{\varepsilon^T\varepsilon}\right) \tag{2.14}$$

The voicing decision can be based on the SNR level and one approach indicated in (Quatieri, 2002) is,

$$\left\{P_v = \left|\begin{matrix} 1, SNR > 10dB \\ \frac{1}{6}(SNR - 4), 4dB \leq SNR \leq 10dB \\ 0, SNR < 4dB \end{matrix}\right.\right\} \tag{2.15}$$

2.7 CSTR Database

Performance evaluation is done on the publicly available database provided by the Center for Speech Technology Research (CSTR) at University of Edinburgh, Scotland, UK. The database includes 50 sentences each from a male and female speaker. The database was biased towards utterances containing voiced fricatives, nasals, liquids and glides, since PDAs generally find these difficult to analyze (Bagshaw, 1994). The analysis window length was fixed at 25ms at 20 kHz sampling frequency and a frame rate of 6.4ms was followed. The pitch range analyzed was between 80-400Hz for both male and female speakers. There was no pre-processing stage to filter the speech signal.

2.8 Performance Comparison

The sentences were recorded with the use of laryngograph so that a reference laryngeal frequency contour can be obtained. The laryngograph measures the impedance between the two electrodes placed bilaterally across the larynx. The measured impedance decreases with the increased vocal fold contact. The glottal closure is marked in the laryngograph signal by a sharp rise to peak. The laryngograph data provides a simple and accurate method of producing a f_0 contour with which all other contours can be compared. The method to extract the $f_0$value from the laryngograph data is outlined in Bagshaw et al., 1993. Every $f_0$value in the reference file had a time label which was used to align the estimated pitch value (P_{est}) with the reference pitch (P_{ref}). A nearest neighbor interpolation was used to compare the two pitch values at the time label where the algorithm estimated the pitch. The error measures computed for performance evaluation are the same as specified in Bagshaw et al., 1993. When the estimated and reference pitch represent voiced speech, we have two error measures namely, gross errors and fine errors. The gross error high (GEH) is counted if $P_{est} > 1.2 * P_{ref}$ and gross error low (GEL) is counted if $P_{est} < 0.8 * P_{ref}$ for the duration when both represent voiced speech. Net gross error (GE) is the sum of GEL and GEH. Fine errors in pitch estimation are defined on the frames where$\left|P_{est} - P_{ref}\right| \leq 0.2$. The duration of unvoiced or silent regions incorrectly classified as voiced by the PDA is noted as *unvoiced in error*. This result is accumulated over all the utterances for a speaker and noted as a percentage of total unvoiced (or silent) duration. Similarly, we have *voiced in error* for the duration of voiced speech that are erroneously classified as unvoiced. The

statistics of the absolute deviation in the fine pitch errors are reported in mean and population standard deviation (p.s.d). The list of PDAs used in the comparison is,

- Cepstrum pitch determination (CPD), Noll, 1967
- Feature-based pitch tracker (FBPT), Phillips, 1985
- Harmonic product spectrum (HPS), Schroeder, 1968
- Integrated pitch tracking algorithm (IPTA), Secrest & Doddington, 1983
- Parallel processing method (PP), Gold & Rabiner, 1969
- Super resolution pitch determinator (SRPD), Medan et al., 1991
- Enhanced version of SRPD (eSRPD), Bagshaw et al., 1993
- Modified AMDF-based PDA with probabilistic error correction (mAMDFp), Ying et al., 1996
- Pitch determination algorithm based on sub-harmonic to harmonic ratio (SHR), Sun, 2000
- Maximum likelihood pitch detection (ML-AIC), Mahadevan & Espy-Wilson, 2011
    - Raw pitch results (raw)
    - Post-processed by median filter (filtered)

The results for the first 7 PDAs are taken from Bagshaw et al., 1993 where eSRPD was shown to perform superior to the rest. The raw pitch estimates from the ML-AIC algorithm were post-processed with a 5 point median filter. The results plotted in Figures 2.3 through 2.8 indicate that the performance of the algorithm is comparable

to or better than most of the PDAs listed. The net gross error which is the sum of GEH and GEL values illustrated in Figures 2.7 and 2.8 reveal the comparison of several existing PDAs with the proposed algorithm.

The GEL values for ML-AIC are quite high as compared to GEH which can be noted in Figures 2.5 and 2.6. The explanation for such bias in error is due to model over fitting. Detailed analyses on these errors on ML-AIC (raw) reveal that 75.86% of the GEL for male and 76.74% of the GEL for female occur due to pitch halving or sub multiple error i.e. $\left|z * P_{est} - P_{ref}\right| \leq 0.2,\ z \in \{2,3,4\}$. Most of the deletion errors (voiced in error) occur in the first few frames or last few frames of a voiced segment. When three frames in the beginning and end of a continuous voiced segment (i.e. no pause or silence in between) were excluded from the analysis, the deletion errors dropped to 3.51% for male and 4.99% for female. Overall the results for the raw pitch estimates indicate that the performance of the algorithm is comparable to (eSRPD) or better than most of the methods in gross errors and fine pitch errors. Median filtering reduced the insertion and deletion errors to some extent as seen in Figures 2.3 and 2.4. The tradeoff for reduction in VAD errors is reflected in fine error measures. The mean absolute deviation and p.s.d show an increase in their values after smoothing in Figures 2.9 and 2.10.

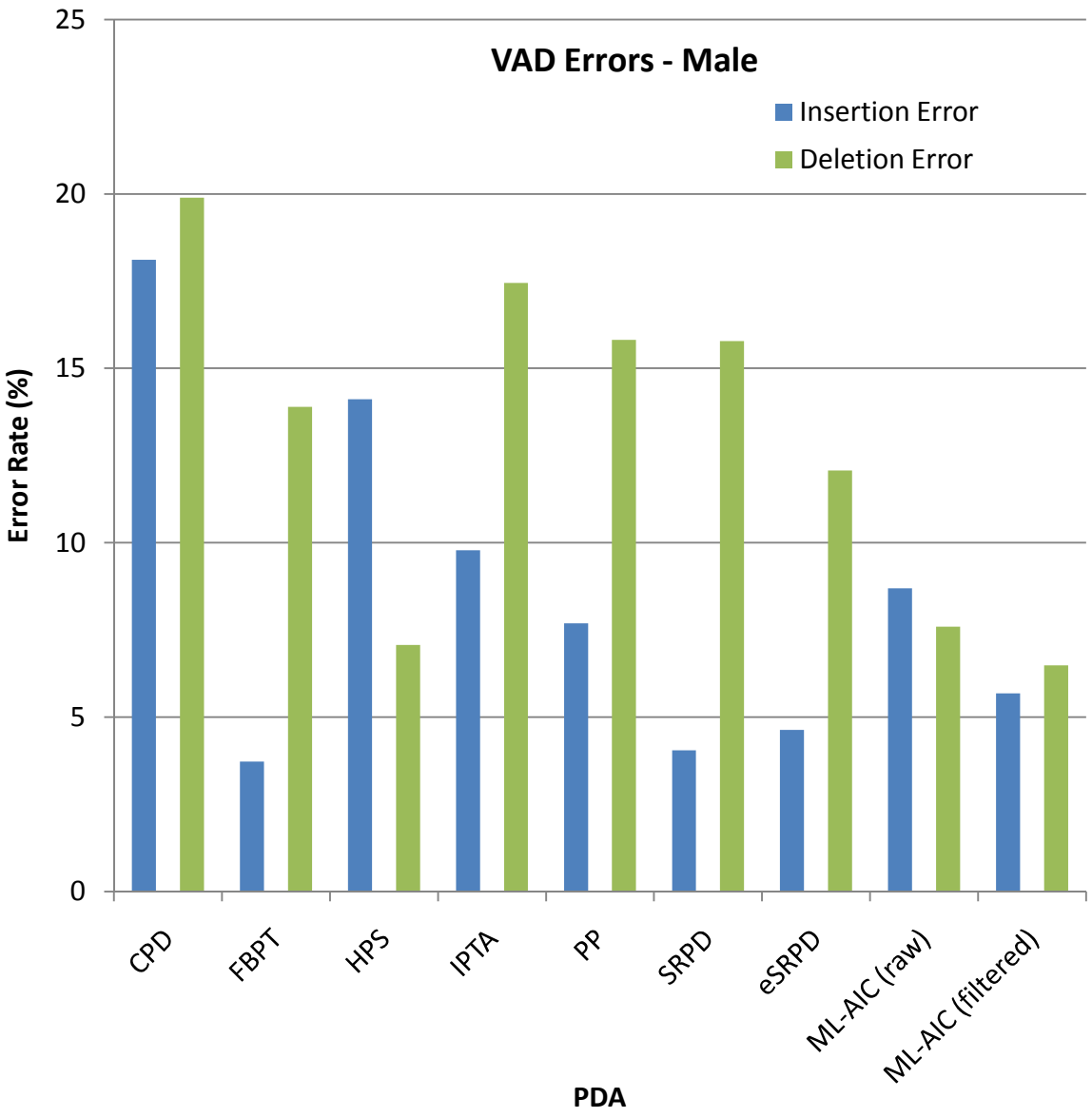


Figure 2.3: Voice activity detection (VAD) errors in PDA for male speaker

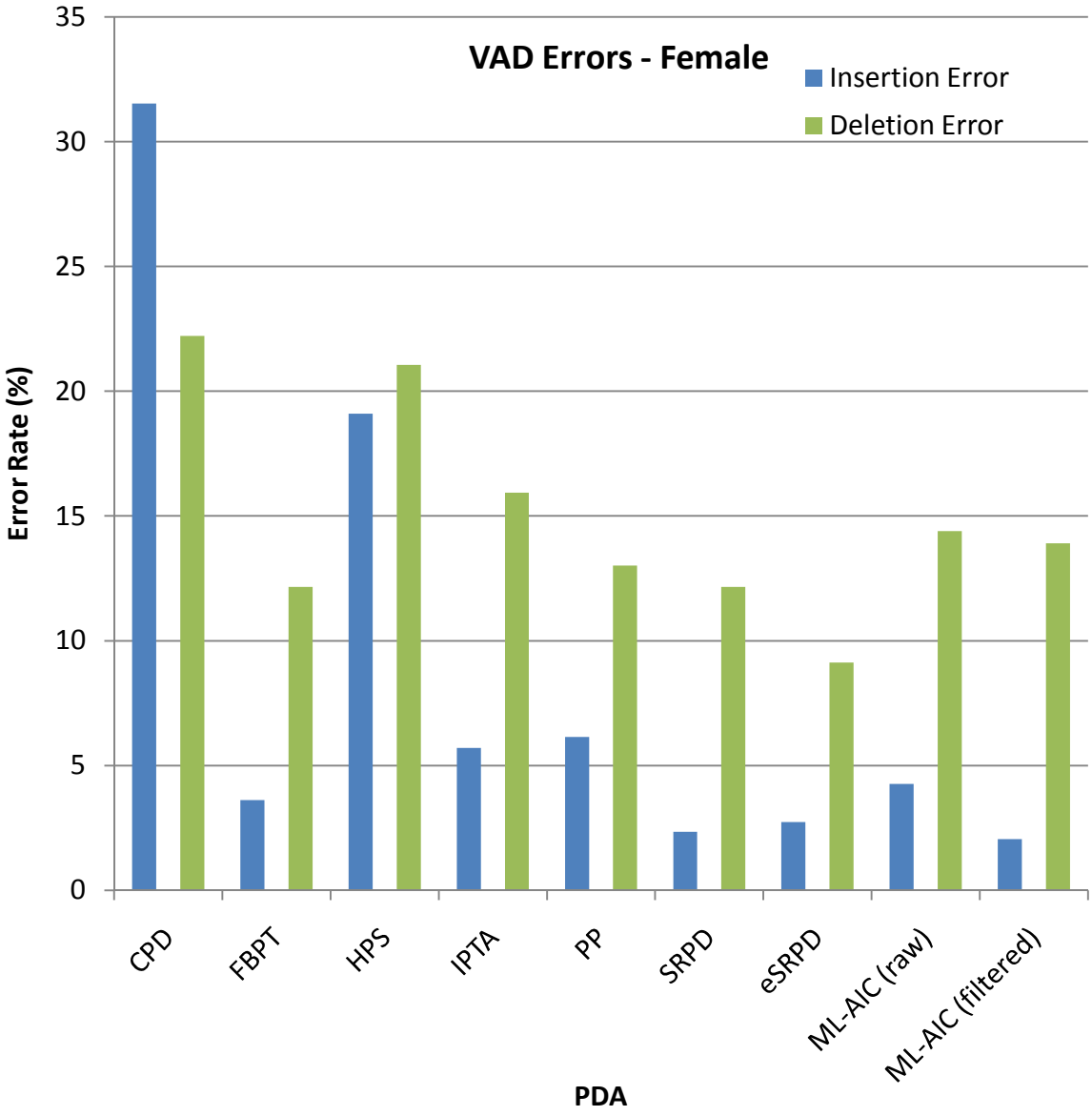


Figure 2.4: Voice activity detection (VAD) errors in PDA for female speaker

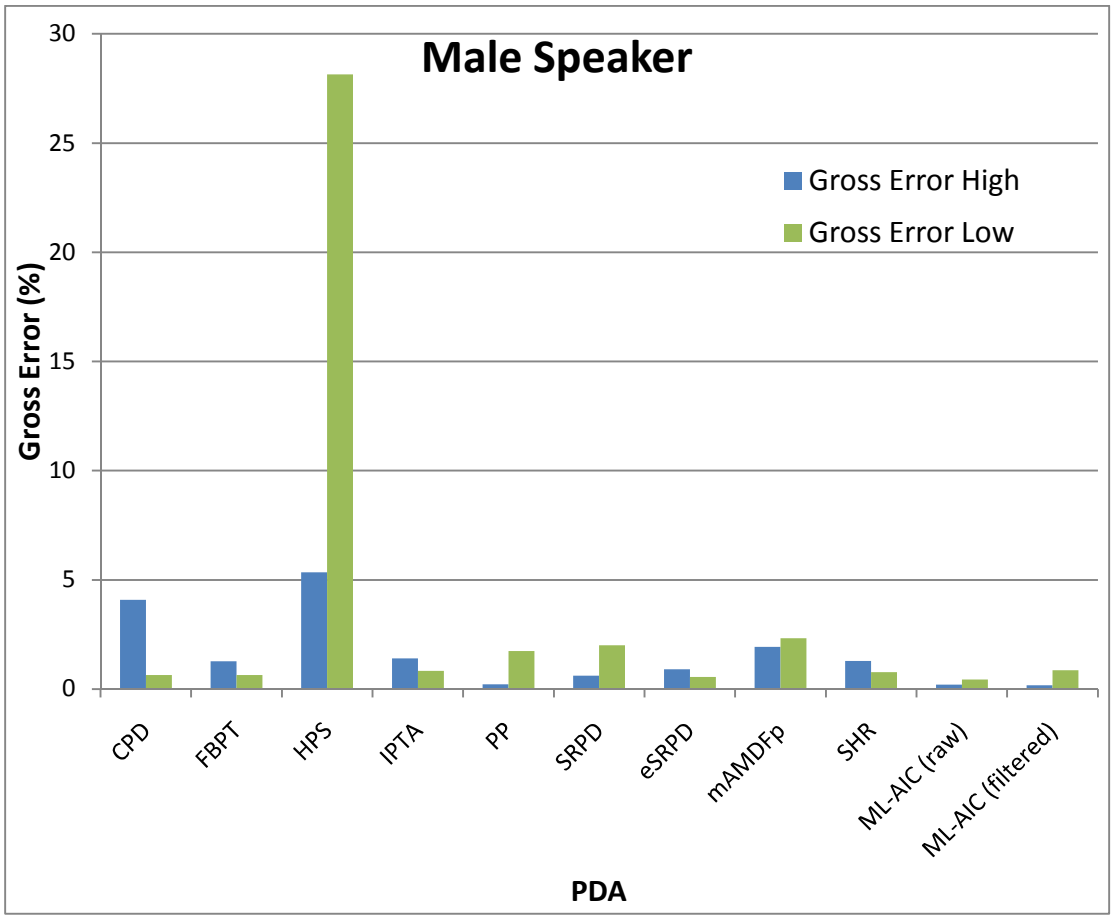


Figure 2.5: Comparison of gross error high and gross error low in PDA for male speaker

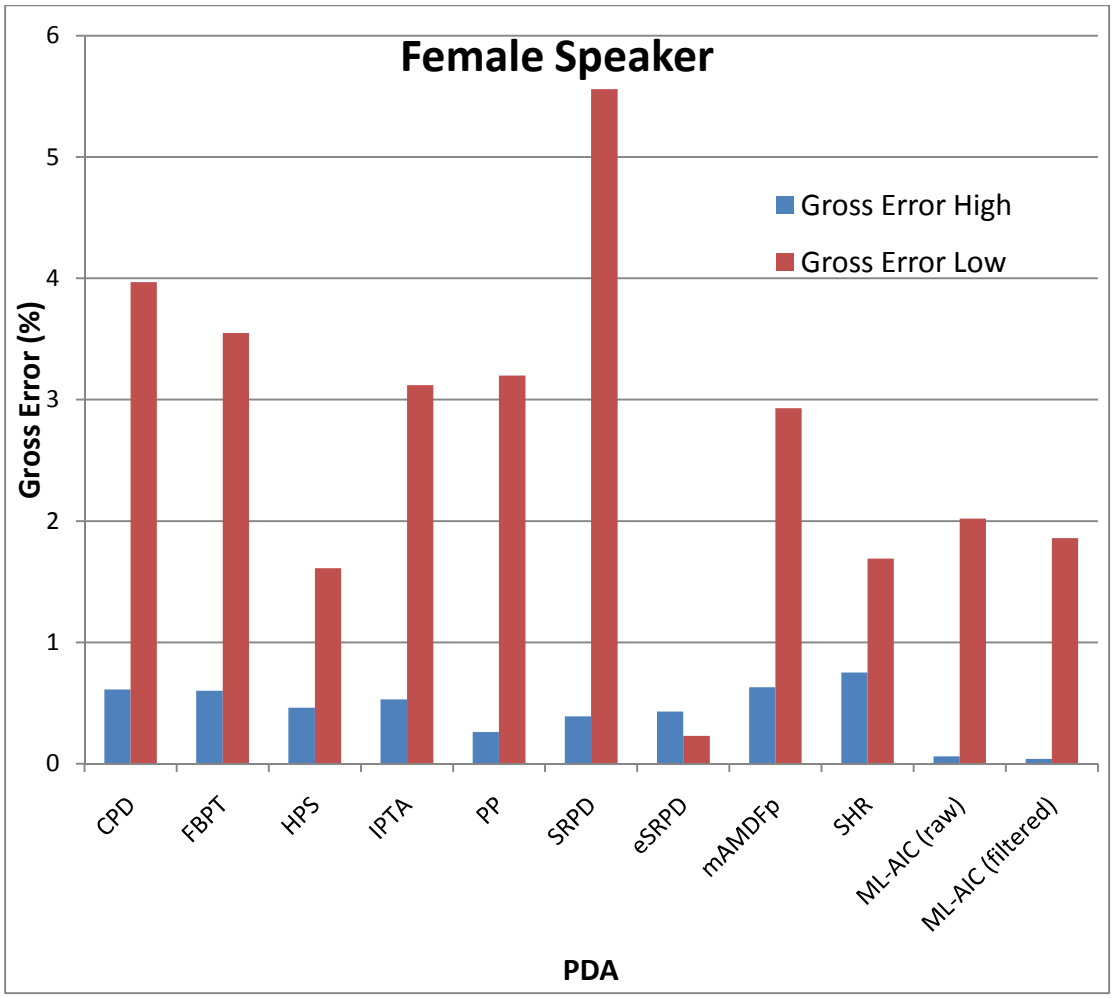


Figure 2.6: Comparison of gross error high and gross error low in PDA for female speaker

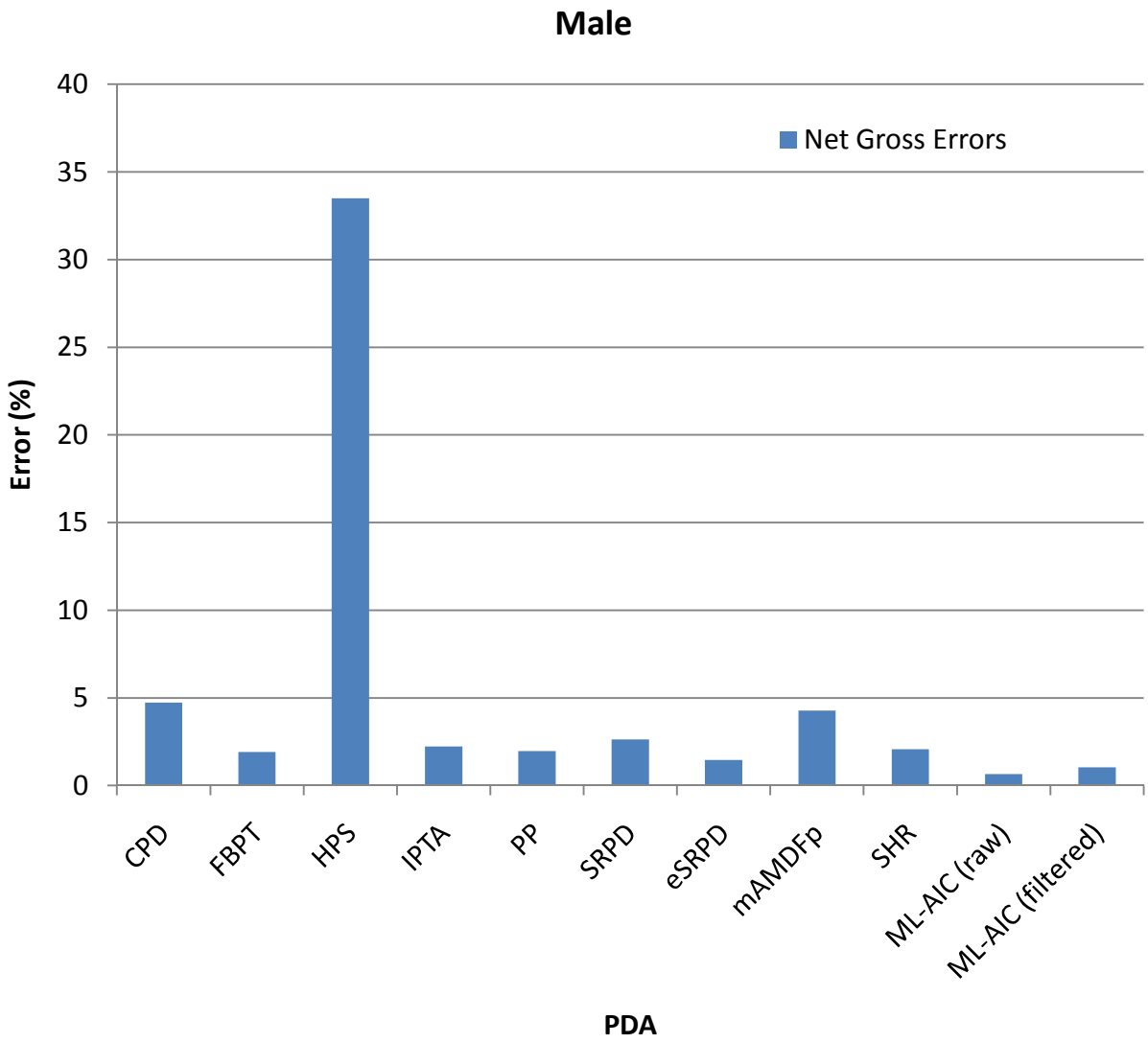


Figure 2.7: Net gross errors in PDA for male speaker

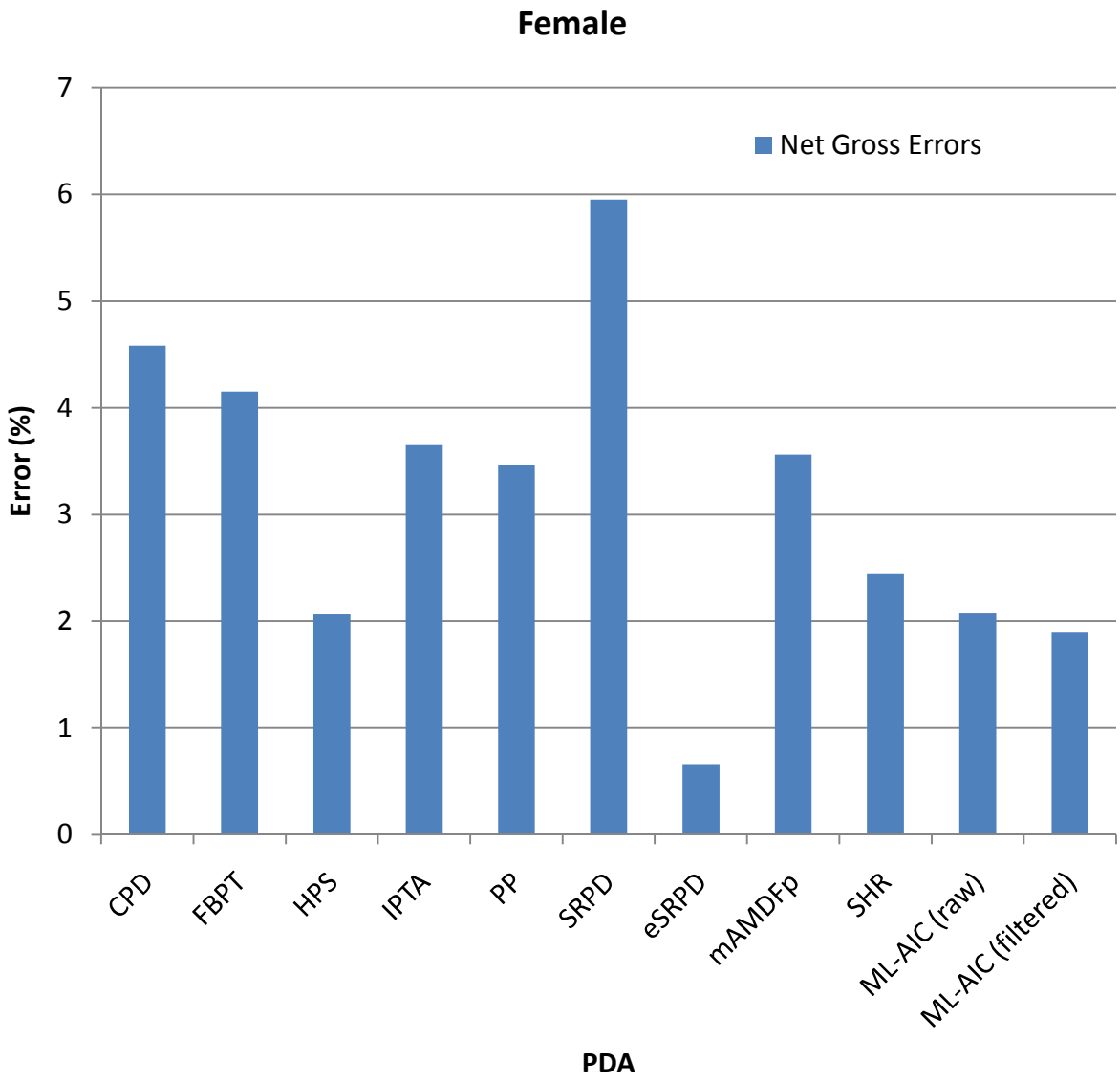


Figure 2.8: Net gross errors in PDA for female speaker

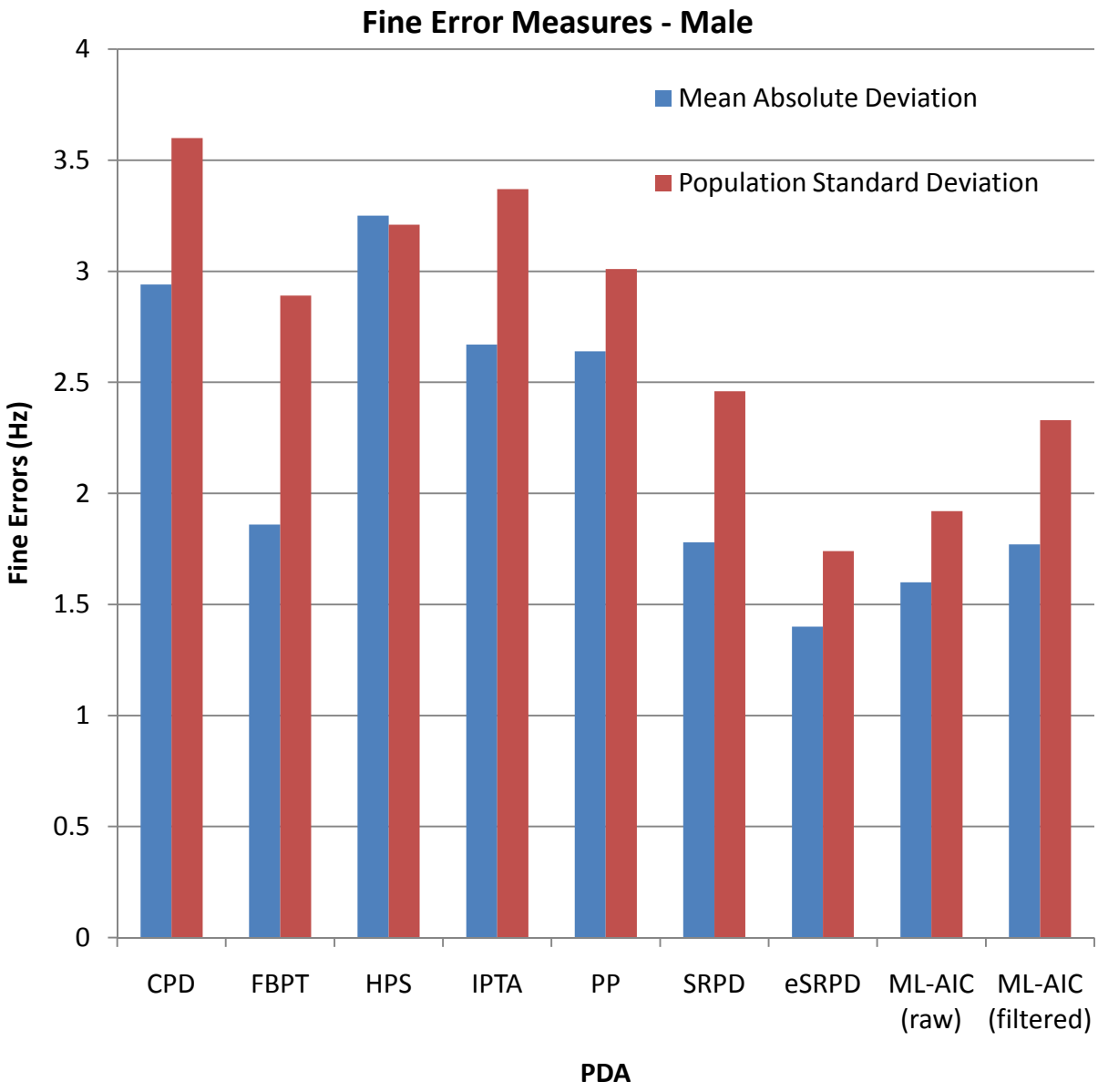


Figure 2.9: Fine errors in PDA for male speaker

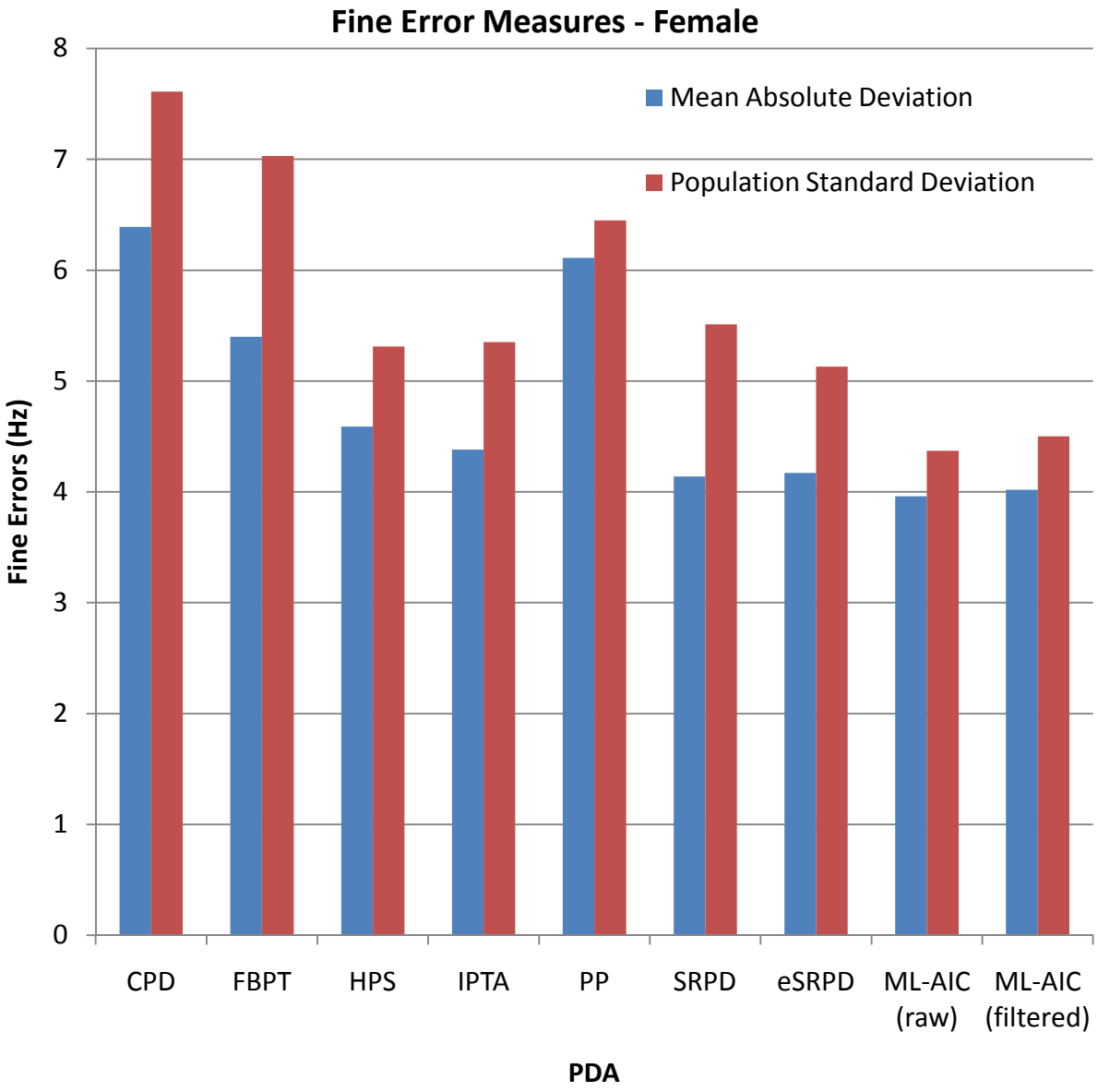


Figure 2.10: Fine errors in PDA for female speaker

Figures 3 and 4 compare the reference pitch with the estimated pitch contour for a male and a female speaker respectively. The reference pitch values were linearly interpolated in the voiced segments at the frame rate followed in the algorithm. The post processed pitch estimates are shown in blue.

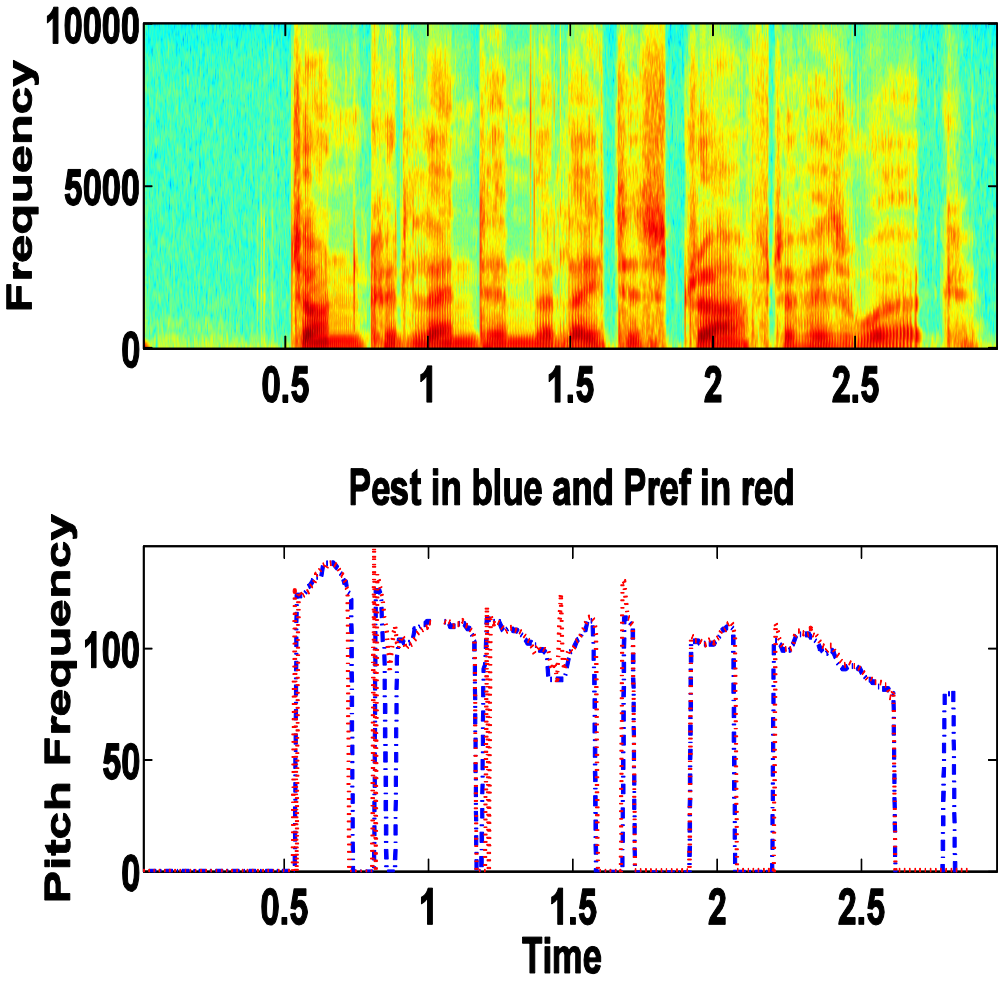


Figure 2.11: (top) Spectrogram and (bottom) comparison of P_{est}(blue) with P_{ref} (red) for male

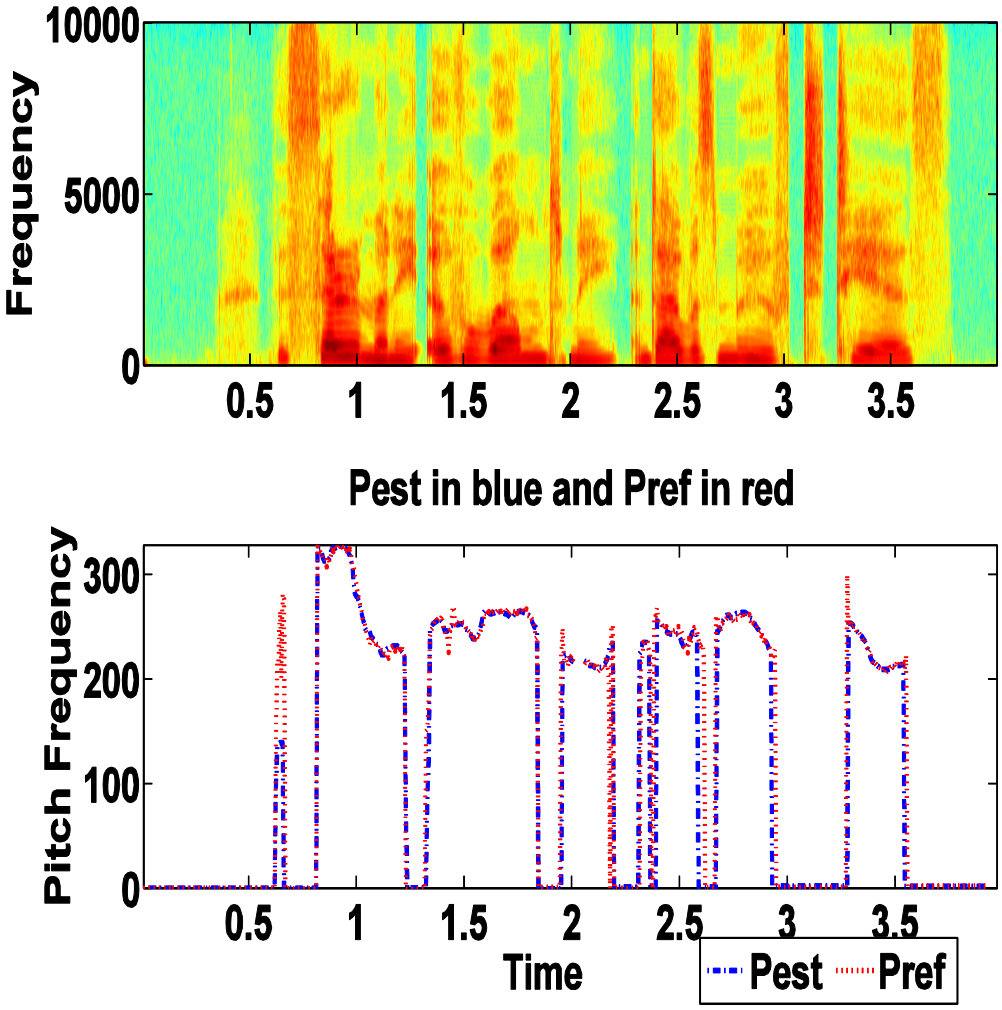


Figure 2.12: (top) Spectrogram and (bottom) comparison of P_{est}(blue) with P_{ref} (red) for female

Pitch estimation is one of the key applications of parametric modeling. In this chapter, we have studied the performance of the pitch detection algorithm and the results reveal the superior performance of the algorithm in comparison with most of the existing techniques. The extension of this model based approach towards multi-

pitch estimation is straightforward but the computational time to search the two dimensional grid space would be extremely high and prohibits its use for real time applications. As a next step towards understanding the application of pitch in signal modeling, chapter three explores the use of parametric models in speech enhancement and speech segregation. The pitch values from the clean utterances were extracted using Wavesurfer in the analysis that follows.

Chapter 3: Applications of Parametric Modeling

3.1 Speech Enhancement

There exist a number of single channel speech enhancement techniques which are discussed in the work by Loizou, 2007. Signal enhancement is a direct application of parametric modeling described in this thesis. The observed signal is projected into the space spanned by the columns of the data model defined by the pitch frequency. A good speech enhancement system based on parametric modeling takes into account the characteristics of noise and the properties of the signal of interest in estimating the parameters of the model. However, estimating the noise and speech characteristics over time requires an adaptive framework which is quite challenging to implement. The noise in the observation is assumed to be additive white Gaussian with zero mean and covariance matrix $R = \sigma^2 I$. If we don't make any assumptions on the prior model for harmonic coefficients, the resulting estimator is given by the ordinary least squares solution.

3.2 Speech Segregation

A parameterization of the signal into components allows for a natural separation of sources if the signal components have a close relation to the sources. In the case of periodic signals, the model discussed in this thesis allows a direct representation of different sources provided their pitch frequencies are known. The parameters of the target and masker can be estimated using least squares solution. A detailed description of the segregation process is explained in Vishnubhotla & Espy-Wilson,

2009. Let us consider the following case of two overlapping sources which are represented by the signal model discussed in Chapter 2.

$$\mathbf{s_1} = \mathbf{A}\,(\mathrm{f}_{0,1}) * \boldsymbol{\gamma_1} + \boldsymbol{\varepsilon_1} \tag{3.1}$$

$$\mathbf{s_2} = \mathbf{A}\,(\mathrm{f}_{0,2}) * \boldsymbol{\gamma_2} + \boldsymbol{\varepsilon_2} \tag{3.2}$$

$$\mathbf{s} = \mathbf{s_1} + \mathbf{s_2} = \left[\mathbf{A}\left(\mathrm{f}_{0,1}\right) \mid \mathbf{A}\left(\mathrm{f}_{0,2}\right)\right] * \begin{bmatrix}\boldsymbol{\gamma_1}\\ \boldsymbol{\gamma_2}\end{bmatrix} + \boldsymbol{\epsilon} = \mathbf{A}\left(\mathrm{f}_{0,1}, \mathrm{f}_{0,2}\right) * \boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{3.3}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon_1} + \boldsymbol{\varepsilon_2} \tag{3.4}$$

Now, observe that the overall signal model is represented by concatenating the data matrix of speaker A with pitch frequency $\mathrm{f}_{0,1}$ and speaker B with pitch frequency $\mathrm{f}_{0,2}$**.** The signal separation framework is achieved by minimizing the energy of the residual error $\boldsymbol{\varepsilon}$. The OLS estimate for $\boldsymbol{\gamma} = \begin{bmatrix}\boldsymbol{\gamma_1}\\ \boldsymbol{\gamma_2}\end{bmatrix}$ is given by the pseudo-inverse of $\mathbf{A}\left(\mathrm{f}_{0,1}, \mathrm{f}_{0,2}\right)$**.** For sake of simplicity we drop the arguments in **A.** The least squares estimate for $\boldsymbol{\gamma}$ is given below,

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{A}^{+} * \boldsymbol{s} \tag{3.5}$$

$$\boldsymbol{A}^{+} = (\mathbf{A}^{\mathbf{T}} * \mathbf{A})^{-1} * \mathbf{A}^{\mathbf{T}} \tag{3.6}$$

The harmonic coefficient estimates for the individual speakers is obtained from $\hat{\boldsymbol{\gamma}}$ by picking the entries that belong to speaker A i.e. the first $\mathrm{M}(\mathrm{f}_{0,1})$ coefficients which represents the number of harmonics for speaker A. Similarly we can find the contribution for speaker B. The process of estimating $\boldsymbol{\gamma}$ requires inverting a matrix which will cause problems if the matrix is not well-conditioned. A challenging problem in speech separation is when we have overlapping harmonics from the two speakers (Danieswicz & Quatieri, 1998; Quatieri and Danieswicz, 2000). At which point, the resulting estimate for $\boldsymbol{\gamma}$ using OLS becomes unreliable. When the data matrix becomes ill-conditioned or singular, the resulting solution for $\boldsymbol{\gamma}$ is no longer unique. In order to give preference to a particular solution with desired properties we need to include an additional penalty in the cost function (Foster, 1961; Sayed, 2008).

3.3 Regularized Least Squares

In regularized least squares (RLS) the cost function is modified to include a penalty on the L^2 norm of the parameters. A formal expression for the cost function is described below:

$$L_{RLS}(\boldsymbol{\gamma}) = \left|\left|\boldsymbol{s} - \mathbf{A}(\mathrm{f}_{0,1}, \mathrm{f}_{0,2}) * \boldsymbol{\gamma}\right|\right|_2 + ||\boldsymbol{\Gamma} * \boldsymbol{\gamma}||_2 = L_{OLS}(\boldsymbol{\gamma}) + ||\boldsymbol{\Gamma} * \boldsymbol{\gamma}||_2 \qquad (3.7)$$

The first term in the cost function is the residual sum of squares; the second term is the regularization term, where $\boldsymbol{\Gamma}$ is called the Tikhonov matrix. The choice of $\boldsymbol{\Gamma}$ as the identity matrix gives preference to solutions with smaller L^2 norms. The Tikhonov matrix $\Gamma = \alpha * \mathrm{I}$, where α is the Tikhonov factor. The value of α decides the tradeoff

between minimizing the residual sum of squares and minimizing the norm of the estimate. For example, setting α to zero implies there is no regularization term and the solution is same as the OLS solution and for $\alpha \rightarrow \infty$, the estimated $\boldsymbol{\gamma}$ approaches zero. For intermediate values of α, the estimated $\boldsymbol{\gamma}$ is shrunk towards zero compared to OLS estimate. The estimate for $\boldsymbol{\gamma}$ using the RLS approach is given by,

$$\widehat{\boldsymbol{\gamma}}_{RLS} = (\mathbf{A}^{\mathbf{T}} * \mathbf{A} + \alpha * \mathbf{I})^{-\mathbf{1}} * \mathbf{A}^{\mathbf{T}} * \mathbf{s} \tag{3.8}$$

Even if the data matrix $\mathbf{A}$ is rank deficient, so that $\mathbf{A}^{\mathbf{T}} * \mathbf{A}$ is singular, the regularized matrix $\mathbf{A}^{\mathbf{T}} * \mathbf{A} + \alpha * \mathbf{I}$ is non-singular for any non-zero value of α and hence a stable solution is guaranteed. In order to study this problem, a synthetic speech mixture voiced frame was created using the signal model. The pitch frequencies used in the model were 110Hz and 275Hz for speaker A and speaker B respectively. Hence the harmonics of speaker A include {110, 220, 330, 440, **550**, 660 ,.., **1100**,..} and the harmonics of speaker B include {275, **550**, 825, **1100** ,..}. Evidently there are overlapping harmonics present in the data matrix and therefore the data matrix is ill-conditioned. The solution from OLS will be unstable which is illustrated in the figure below. A comparison is made with the RLS method where Tikhonov matrix $\Gamma = \mathrm{I}$. The target to masker ratio (TMR) for the synthetic mixture was 12dB.

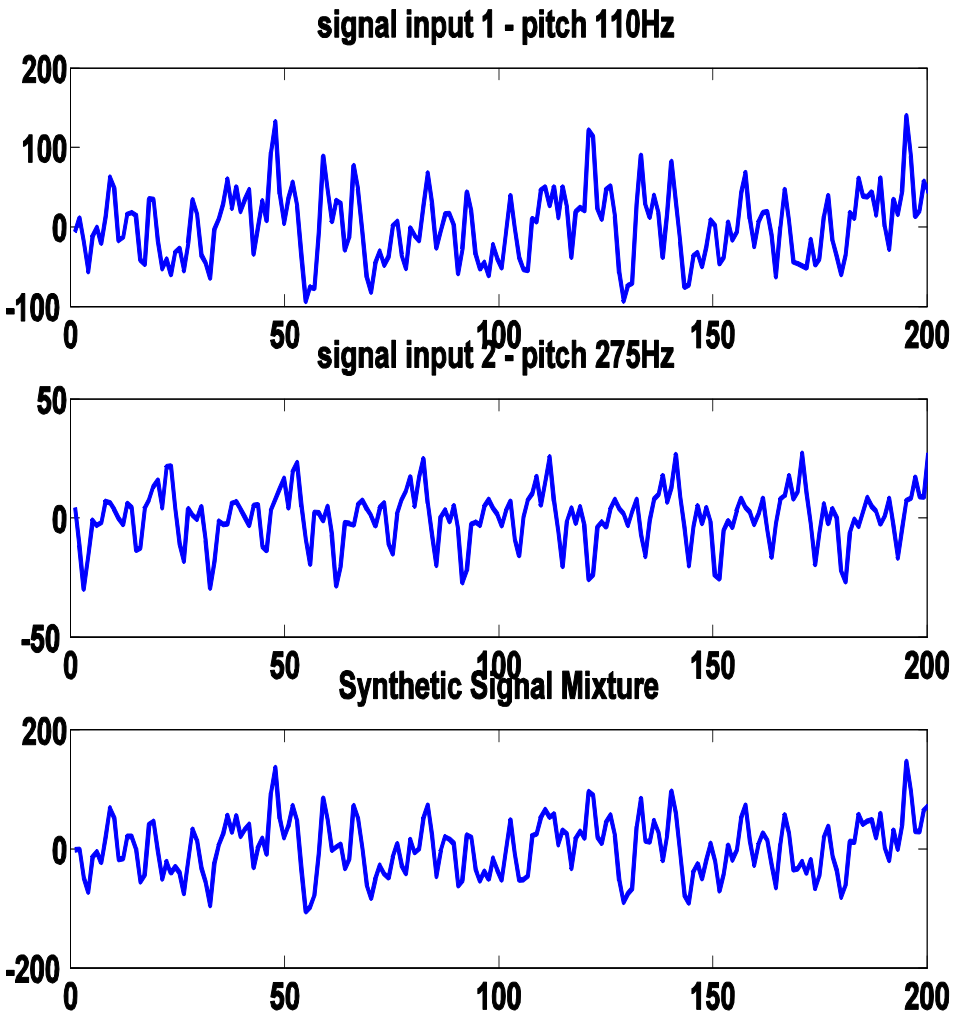


Figure 3.1: Synthetic signals analyzed to compare OLS and RLS

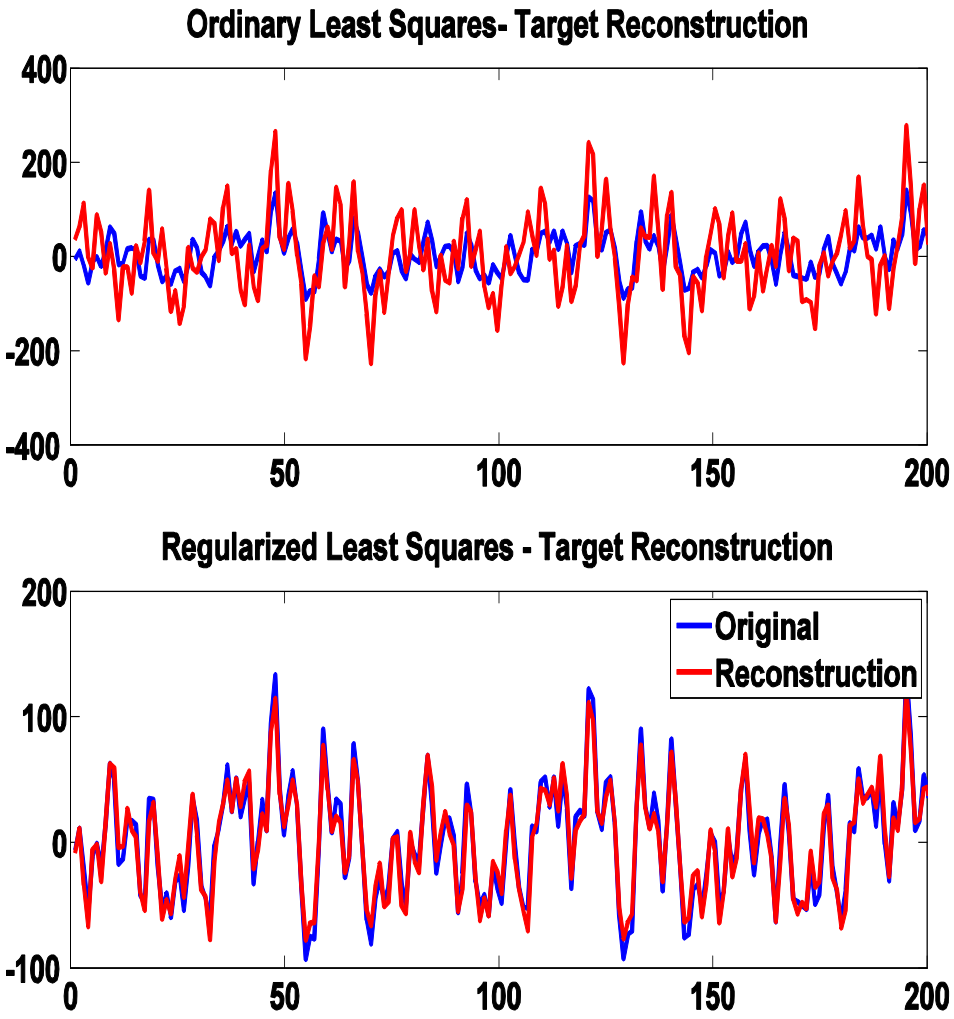


Figure 3.2: Synthetic target signal reconstruction (pitch 110Hz) using OLS and RLS

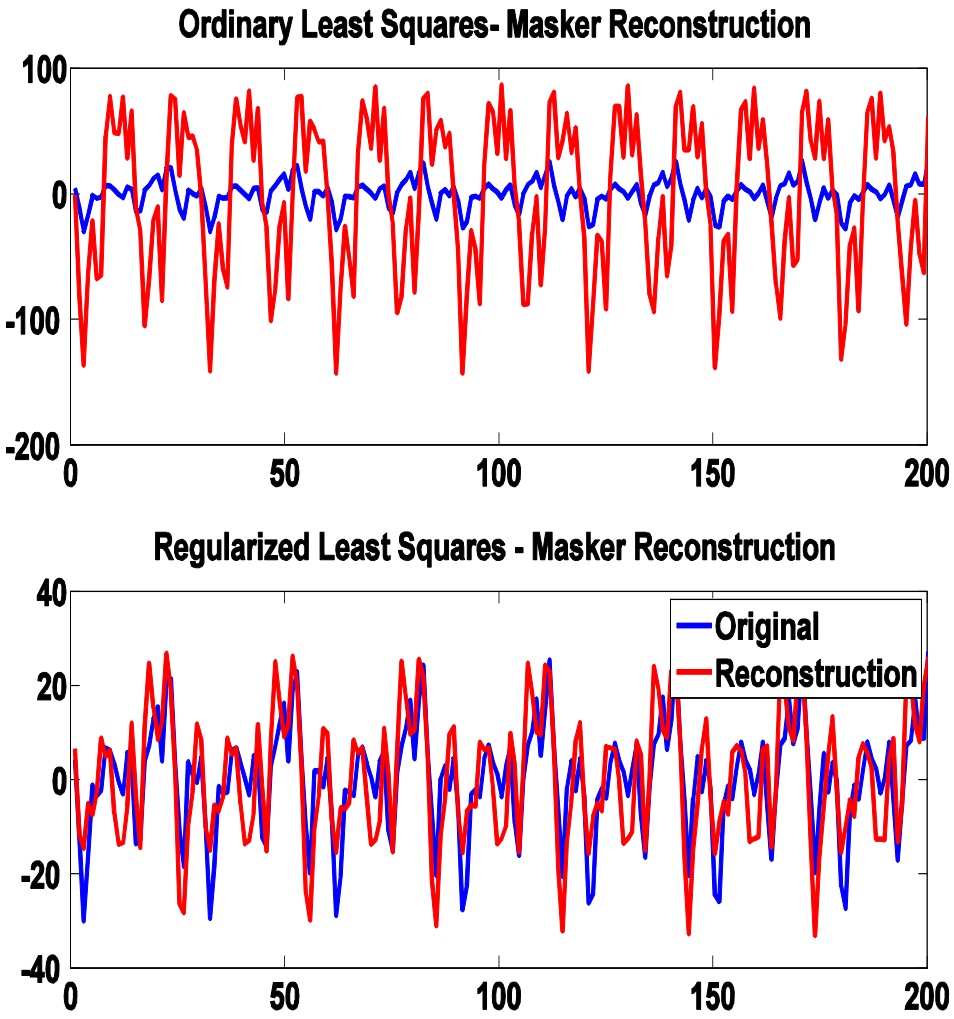


Figure 3.3: Synthetic masker signal reconstruction (pitch 275Hz) using OLS and RLS

3.4 SSC Database

The speech separation challenge (SSC) database (Cooke et al., 2010) was used to evaluate the performance of the algorithm. All the speech files are single-channel "wav" data sampled at 25 kHz. The files were down sampled to 8 kHz in all the

processing and analysis. The categories of speech mixtures analyzed belonged to different talker files with gender categories being *"MaleMale, FemaleMale, MaleFemale and FemaleFemale"*. The database has a total of 1200 clean files from 34 speakers of which 18 are male and 16 are female. These files were mixed at different target to masker ratios (TMRs) to analyze the robustness of the algorithms. The TMRs analyzed are 0dB, 3dB and 6dB from the SSC database. There are a total of 379 files per TMR in the different talker category (includes same gender and different gender files only).

3.5 Experimental Results

A real speech mixture frame was analyzed with the pre-mixing pitch values of 110Hz and 120Hz for target and masker respectively. A similar unstable behavior of OLS was observed and the estimates are compared with RLS which reveals the importance of regularization in speech separation. To evaluate the comparison of OLS and RLS in speech separation, an extensive evaluation on the segregated signals was done on the speech separation challenge database. Perceptual evaluation of speech quality (PESQ) was used as the objective measure to compare the performance of the algorithms (Rix et al., 2001). PESQ is a standard used for comparing the quality of the speech signals transmitted over the telephone network. The PESQ score ranges from 0.5 (highly degraded) to 4.5 (high quality). The pitch values were taken from Wavesurfer before the utterances were mixed.

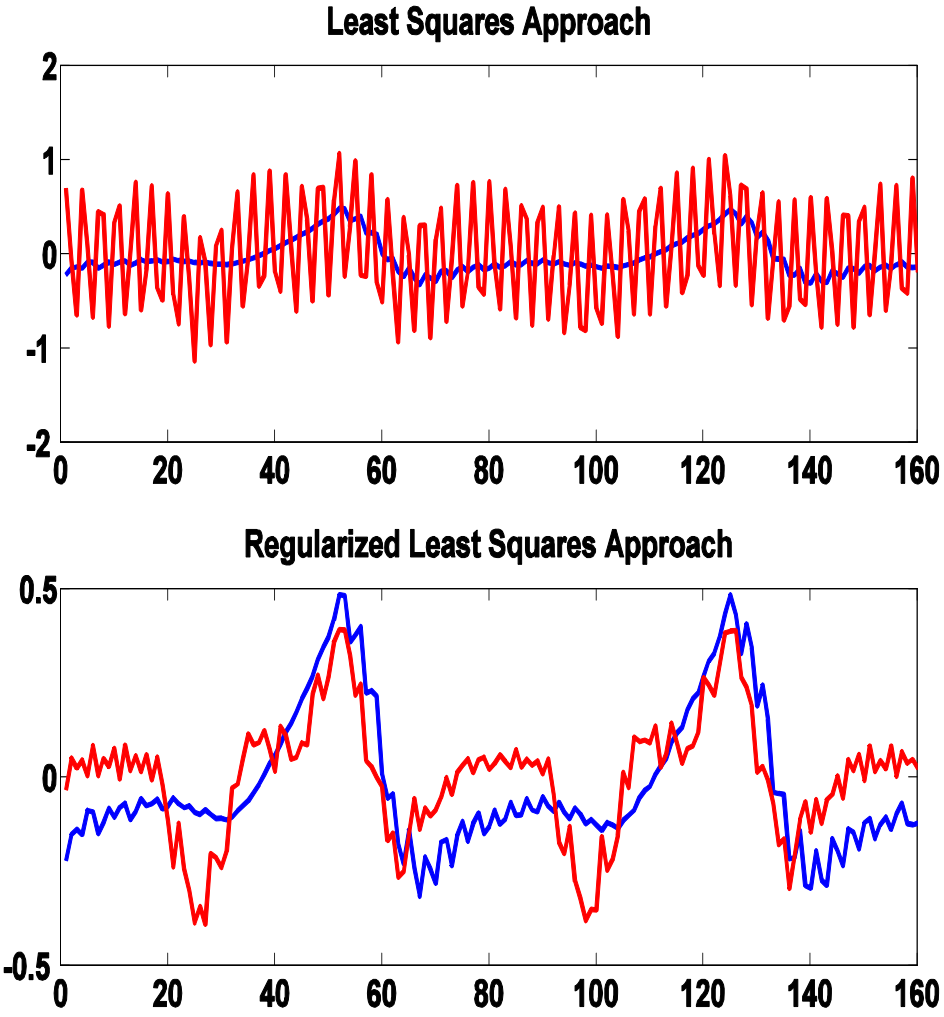


Figure 3.4: Target signal reconstruction (pitch 110Hz) using OLS and RLS with the original signal in blue and reconstructed signal in red

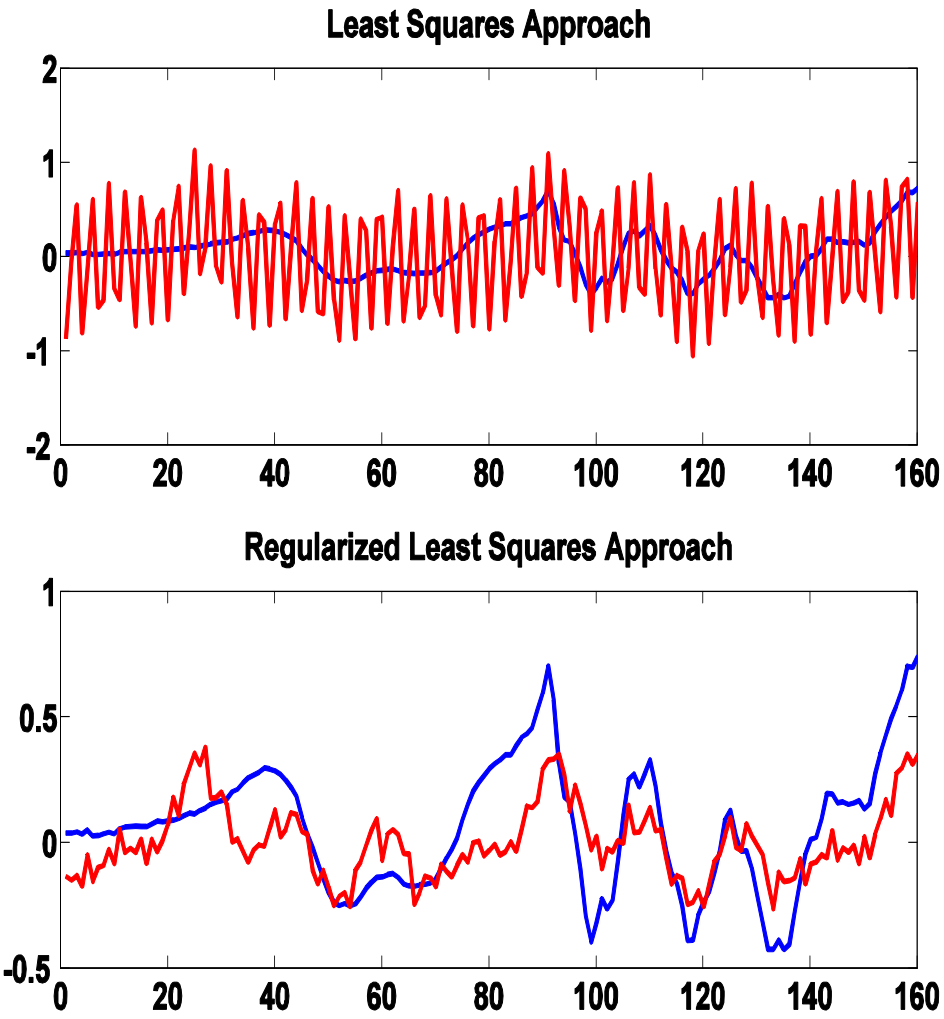


Figure 3.5: Masker signal reconstruction (pitch 120Hz) using OLS and RLS with the original signal in blue and reconstructed signal in red

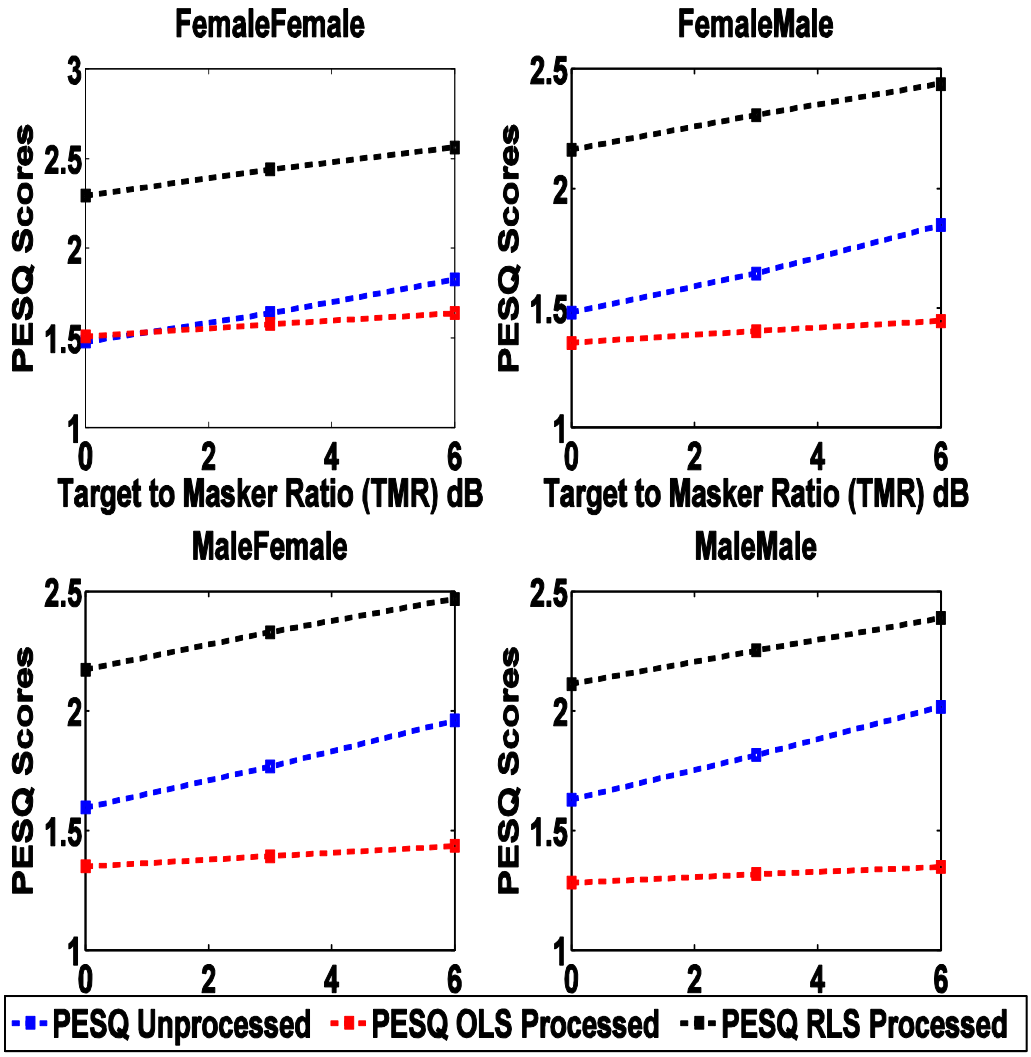


Figure 3.6: PESQ scores for Target Speaker comparing the segregation performance of OLS and RLS

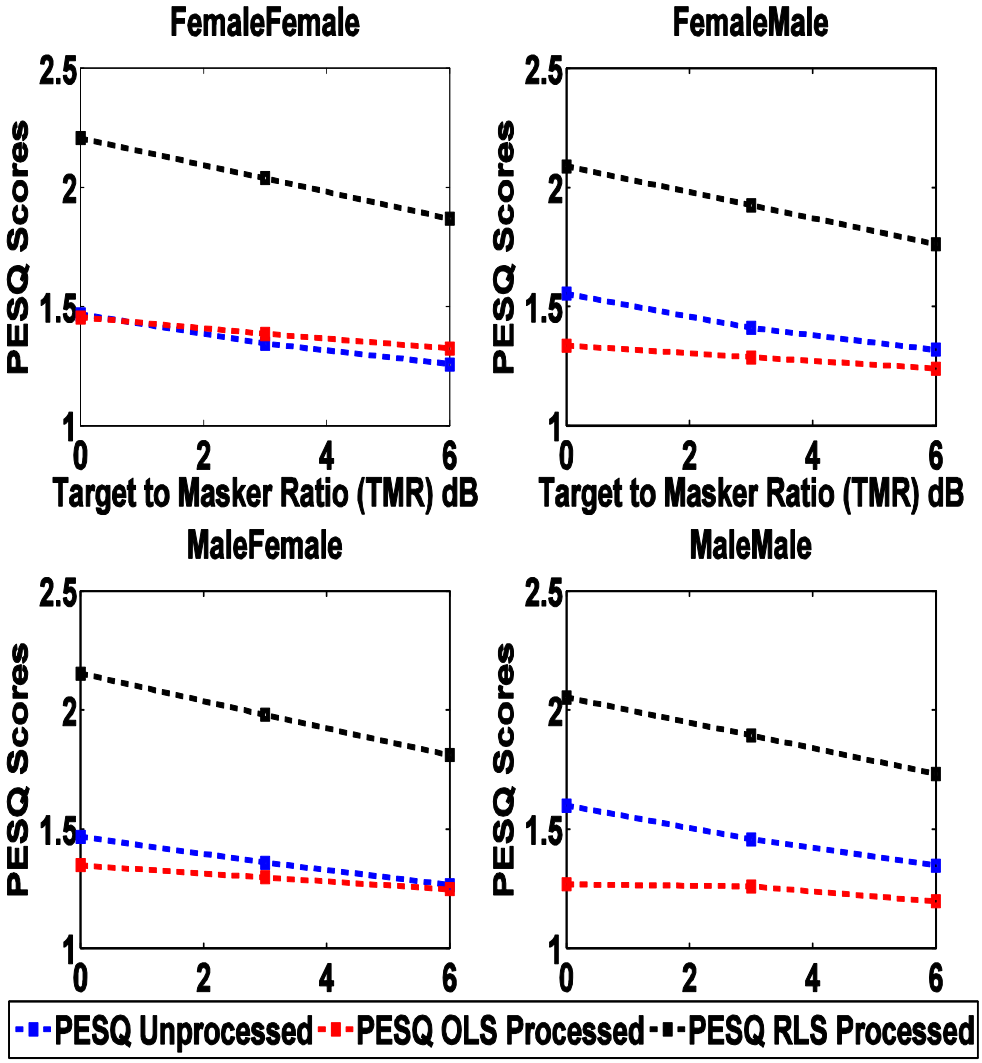


Figure 3.7: PESQ scores for Masker Speaker comparing the segregation performance of OLS and RLS

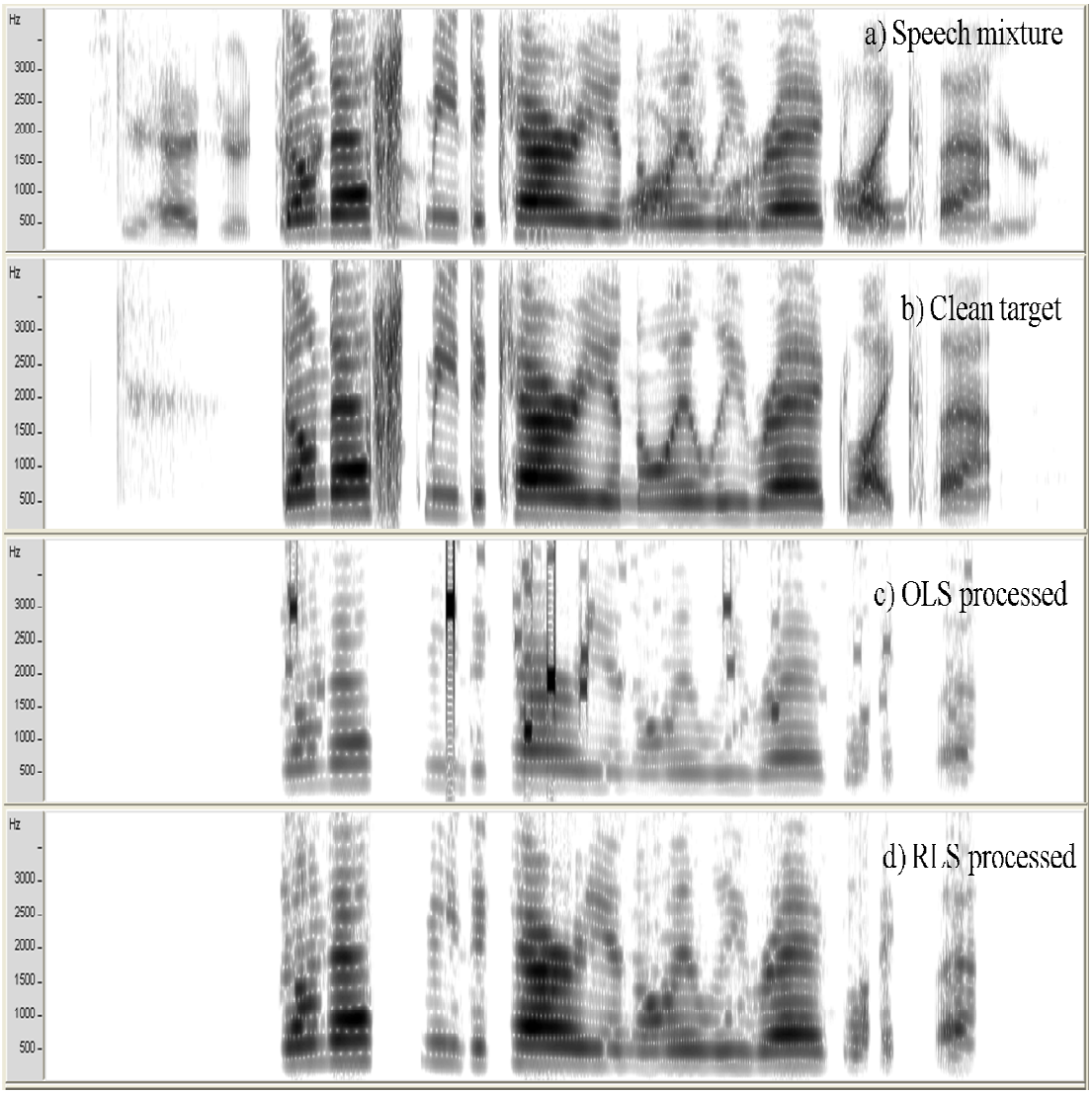


Figure 3.8: Spectrograms comparing the performance of RLS and OLS for target signal. Panel a): Speech mixture at 6dB TMR (PESQ = 2.28), Panel b): Clean target, Panel c): Extracted target using OLS (PESQ = 1.50), Panel d): Extracted target using RLS (PESQ = 2.84)

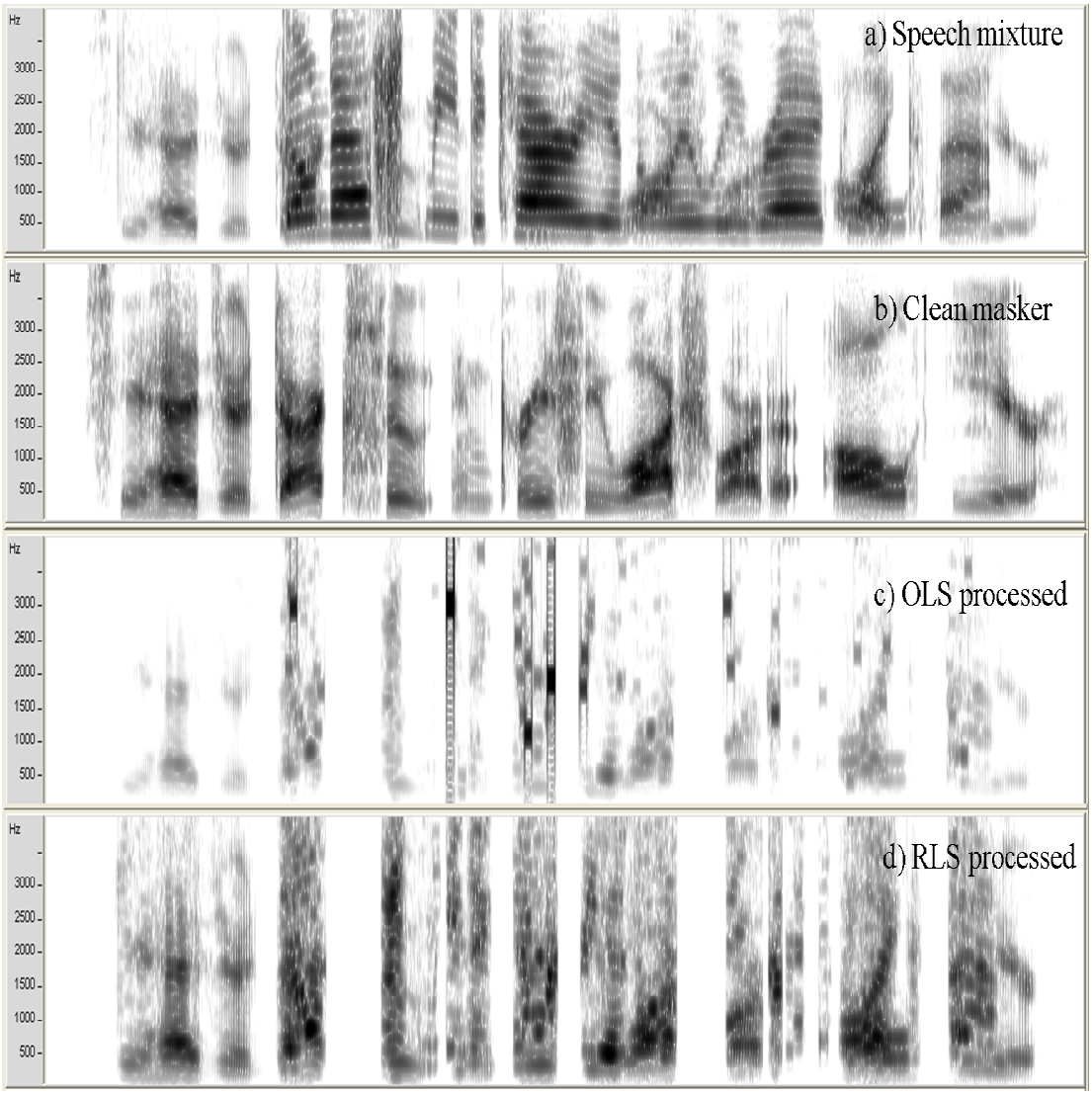


Figure 3.9: Spectrograms comparing the performance of RLS and OLS for masker signal. Panel a): Speech mixture at 6dB TMR (PESQ = 1.34), Panel b): Clean masker, Panel c): Extracted masker using OLS (PESQ = 1.13), Panel d): Extracted masker using RLS (PESQ = 1.50)

3.6 Critical Region Analysis

The critical regions where OLS will result in unstable solution are when the two pitch values come close or when some harmonics of the two speakers come close together.

In these regions the system matrix is ill-conditioned and therefore we suggested the use of RLS. An interesting experiment to perform is to understand if RLS provides any improvement as compared to not processing the frames in critical region. A frame will be flagged as critical if the system matrix describing the observation has a condition number greater than a threshold value, t_c (100). There were two cases analyzed for these frames,

1) Leave the frames as they were in the mixture for both target and masker extracted speech i.e. no processing (NP).
2) Process the frames using RLS and extract the target and masker contributions.

The claim here is that using RLS for speech segregation provides less leakage as opposed to leaving the critical frames unprocessed. Since, RLS aims at extracting the speech only the harmonics of the pitch, any overlapping harmonics will have equal energies distributed between them and the non-overlapping harmonics are segregated. In order to quantify the performance of the above two cases, PESQ scores were analyzed on the segregated speech signals across the SSC database. It can be observed that the improvement in PESQ is very minimal for different gender category for both target and masker and noticeable for the same gender category. The difference between the two can be observed in the spectrograms shown in Figures 3.11 and 3.12 for target and masker speaker respectively. The dark red ellipse highlights the critical region frames for the female-female mixture at 0dB TMR. As the plots reveal, the original mixture contents is retained for both target and masker on the highlighted region in panel 3. In panel 4, the extent of leakage from either speaker is minimized as seen in the spectrogram.

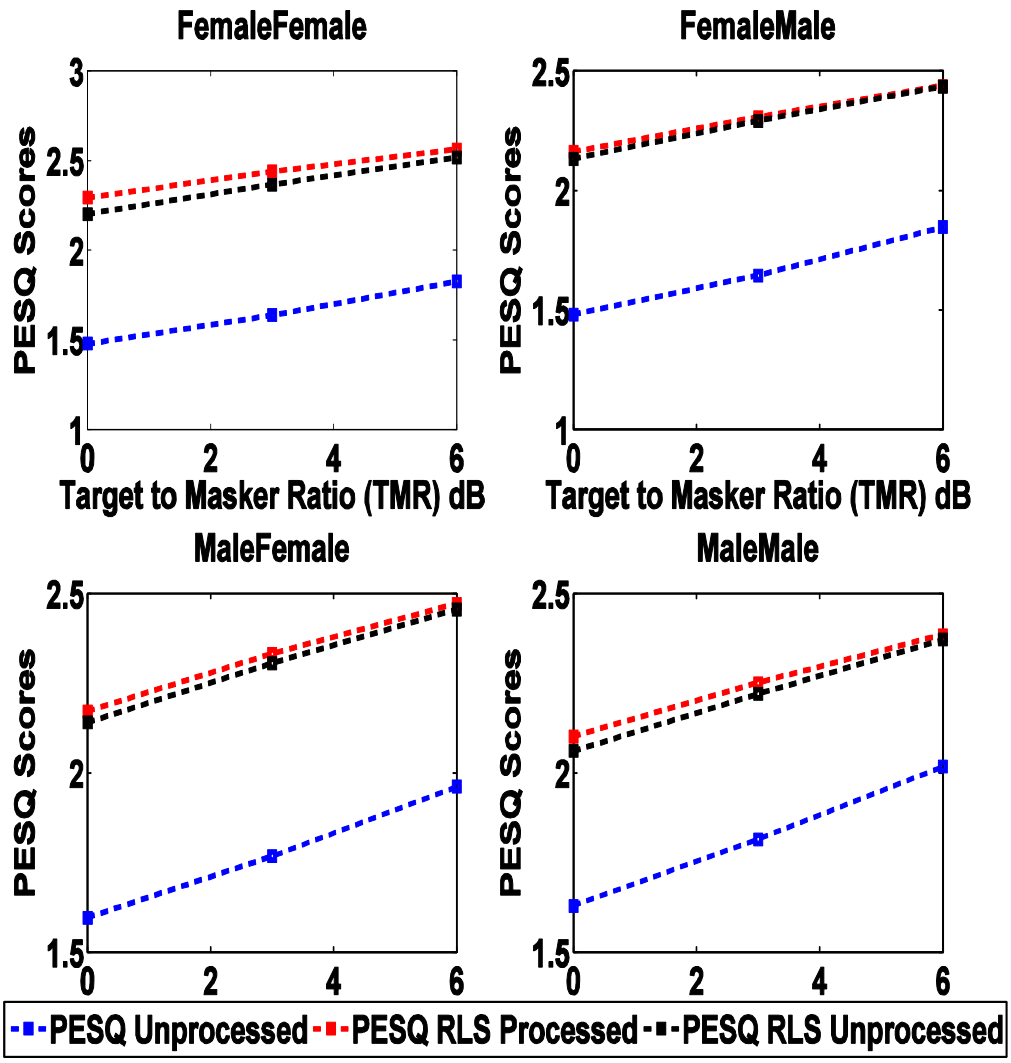


Figure 3.10: PESQ scores for target speaker comparing the segregation performance on critical regions using RLS.

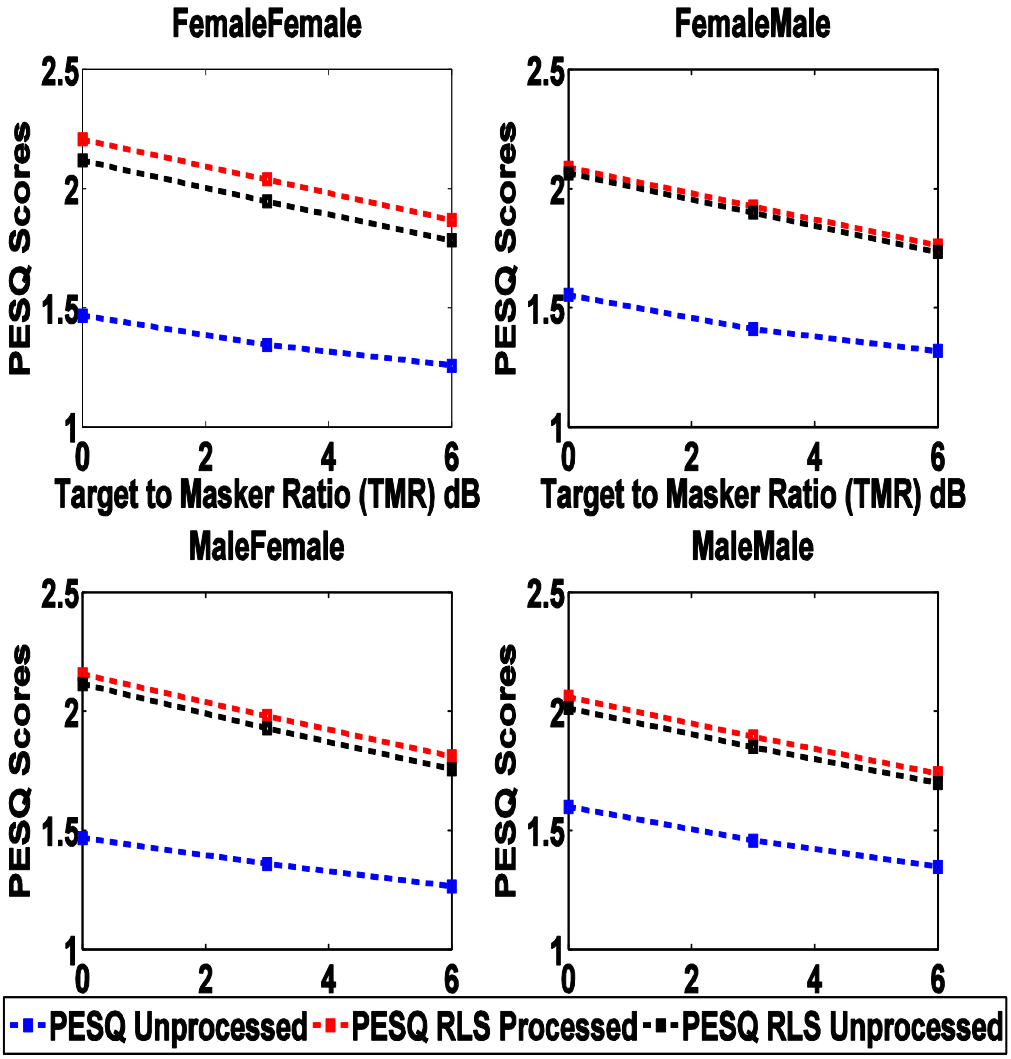


Figure 3.11: PESQ scores for masker speaker comparing the segregation performance on critical regions using RLS.

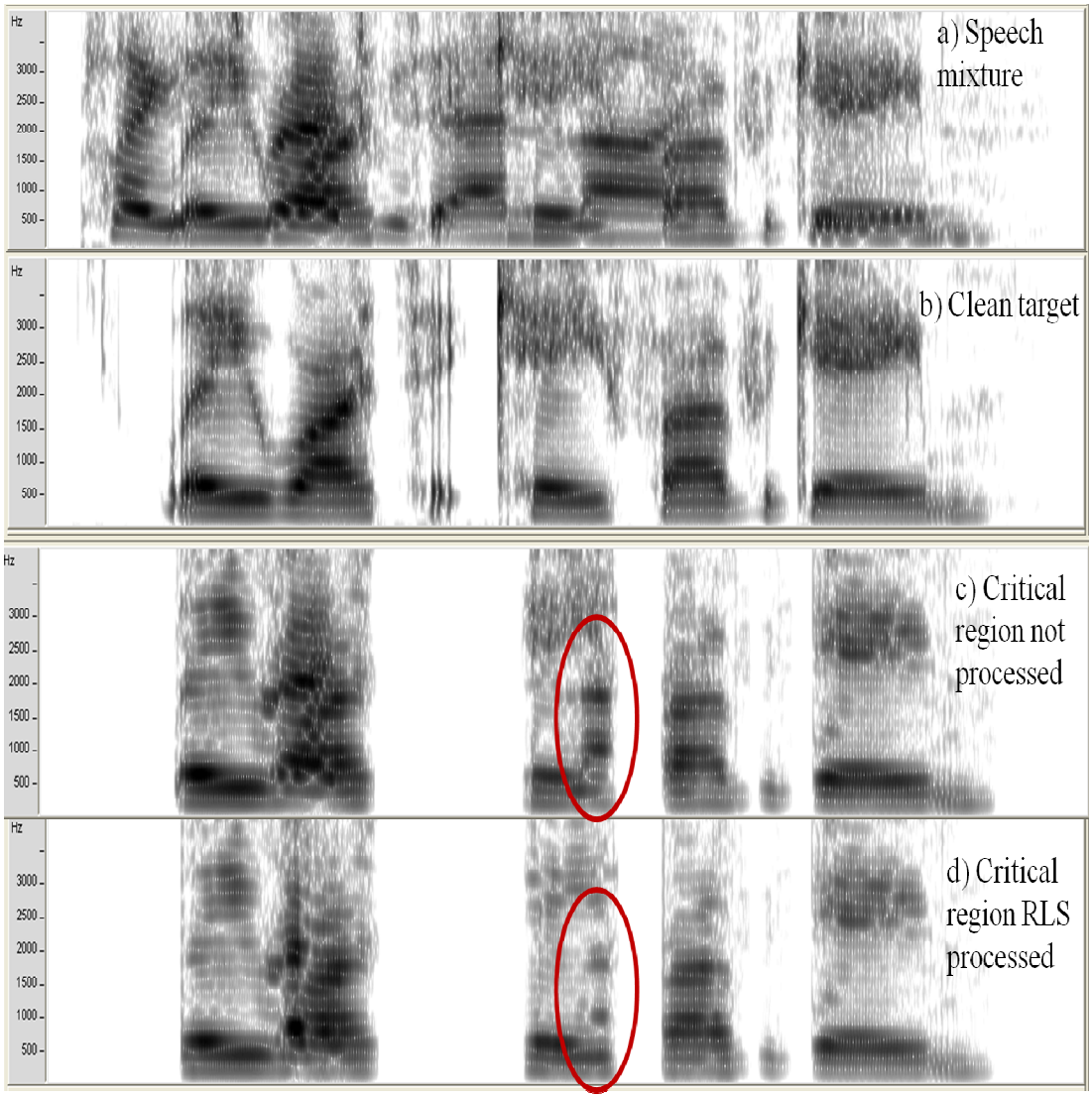


Figure 3.12: Spectrograms comparing the performance of RLS and NP for target signal. Panel a): Speech mixture at 6dB TMR (PESQ = 1.51), Panel b): Clean target, Panel c): Extracted target using RLS with critical region unprocessed (PESQ = 1.87), Panel d): Extracted target using RLS processing on critical region (PESQ = 2.14).

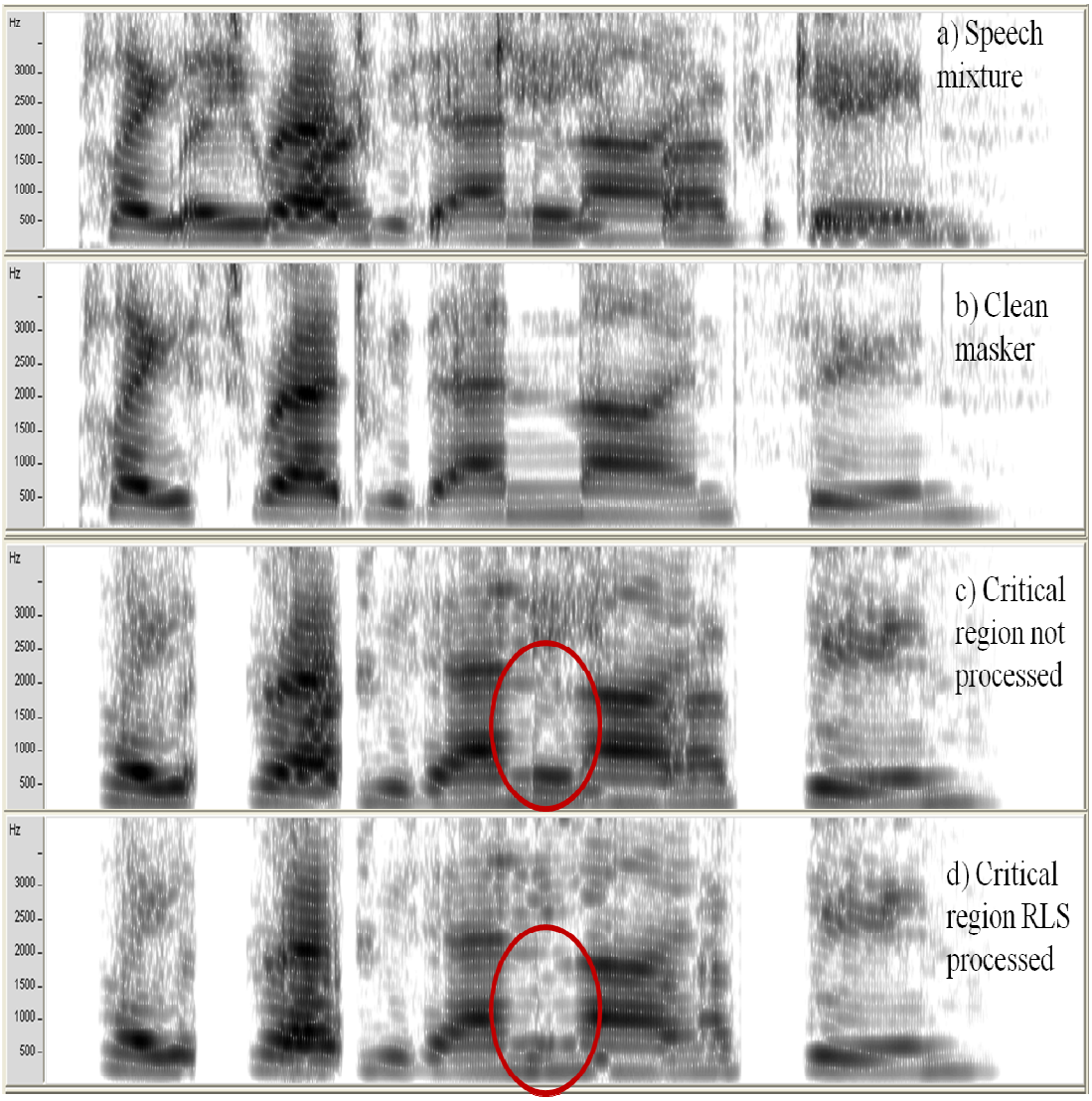


Figure 3.13: Spectrograms comparing the performance of RLS and NP for masker signal. Panel a): Speech mixture at 6dB TMR (PESQ = 1.51), Panel b): Clean masker, Panel c): Extracted masker using RLS with critical region unprocessed (PESQ = 2.09) and Panel d): Extracted masker using RLS processing on critical region (PESQ = 2.24).

Chapter 4: Sequential Grouping in Co-channel Speech

4.1 Speech Segregation System

The block diagram shown in Figure 4.1 describes the working of the speech segregation system that is studied in this thesis. This system has three major components namely: 1) Multi-pitch detector, 2) Segregation block and 3) Sequential grouping block. The segregation system is designed to separate two overlapping sources, and starts by analyzing the input signal into a number of channels. The signals through these channels are used to estimate the pitch of both participating speakers. These pitch estimates are used to set up a least-squares matrix equation, the solution for which yields the harmonic amplitudes and phases of both the speakers (Vishnubhotla & Espy-Wilson, 2009). The next stage assigns the segregated speech to the appropriate speaker for each time frame. Finally, overlap-add synthesis over multiple frames yields the final reconstructions of both speaker streams.

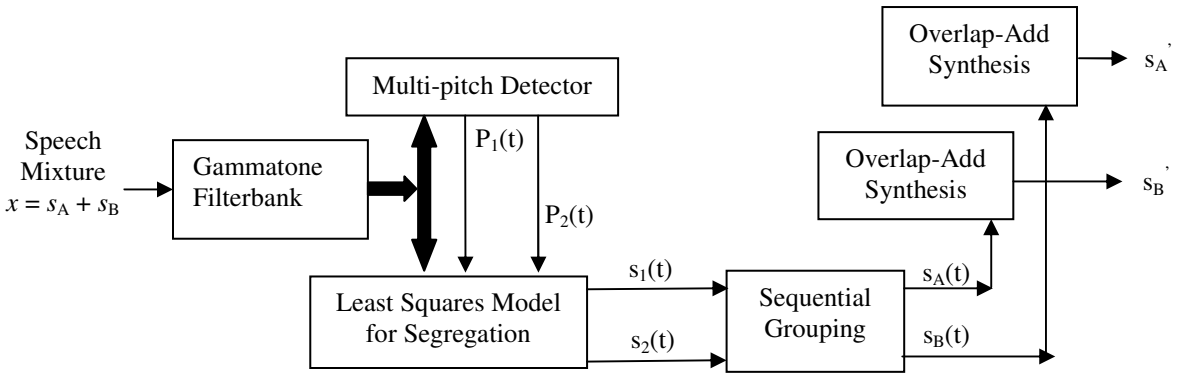


Figure 4.1: Block diagram of the speech segregation system

4.2 Multi-pitch Detector

The pitch of both speakers is next estimated using the channel outputs from the gammatone filterbank (Hohmann, 2002) using the multi-pitch detector algorithm described in (Vishnubhotla & Espy-Wilson, 2008). This algorithm relies on the 2-dimensional Average Magnitude Difference Function (2-D AMDF) as the pitch-estimation feature, and gives robust pitch estimates for both speakers. As described in (Zissman & Seward, 1992; Zissman, 1991), using pitch labels may not be the best cue to use for speaker assignment. Thus, in this work, the algorithm is used only to estimate the numerical values of both pitches, but not to assign either pitch to either speaker - this assignment problem is instead solved in a later section.

4.3 Least Squares Model for Segregation

This block was extensively discussed in the previous chapter where we proposed the use of RLS instead of OLS in frames where pitch values come close together or one is a multiple of another or when the two speakers have overlapping harmonics. By virtue of this segregation block, we end up with an extracted speech frame that has a one to one correspondence with the pitch value. Hence, given we have two pitch values at a frame, the output of this stage will be two extracted speech signals that represent the signal contributions of the target and masker on that frame. The least squares segregation therefore operates on a frame level basis yielding signal outputs that map one to one with the pitch estimates. The question of connecting the speech frames that belong to the target speaker into one stream and those belonging to masker speaker into another stream is addressed in the following section.

4.4 Sequential Grouping Block

Even though the multi-pitch detector yields the numerical pitch estimates of the two speakers, and the segregation algorithm yields the two constituent speech signals for a given frame, these are not yet assigned to any speaker. In particular, for two speakers A and B, and two segregated signals $s_1(t)$ and $s_2(t)$, the question of which segregated signal $s_i(t)$ should be assigned to speaker A will be answered in this section and the next. In the proposed algorithm, the well known Mel frequency cepstral coefficients (MFCCs) are used as the features for the speaker assignment problem.

4.5 Motivation

The motivation behind using the MFCCs is to capture the vocal tract features using the extracted harmonic coefficients from the co-channel speech. In speech analysis, we usually estimate parameters of an assumed speech-production model. The most common model views speech as the output of a linear, time-varying system (the vocal tract) excited by either quasi-periodic pulses or random noise. The deconvolution of the system and source components is feasible for speech because the convolved signals have very different spectra. Since the spectral envelope (characteristic of vocal tract) varies slowly with time, the possibility of using this information in sequential grouping is explored. The Mel frequency cepstral coefficients provide an alternative representation of the speech spectra which incorporates some aspects of audition (O'Shaughnessy, 2000). The C_0 coefficient represents the average energy in the speech frame. The value of C_1 reflects the energy balance between low and high frequencies. The coefficients C_2- C_{12} capture the finer details of the spectrum.

4.6 Classification of Sequential Grouping

The speaker assignment problem is divided into two classes namely: Intra-segment sequential grouping and Inter-segment sequential grouping. A segment is defined as a region of continuously voiced co-channel speech without any pauses, silence or unvoiced speech. In intra-segment sequential grouping, the problem is to connect the speech frames within the continuously voiced segment. Inter-segment grouping addresses the issue when we have a voiced-unvoiced-voiced transition in the co-channel speech. Then the problem is to link the voiced speech regions across the unvoiced region. The motivation for this classification is due to the inherent continuity in the acoustic features within a voiced segment.

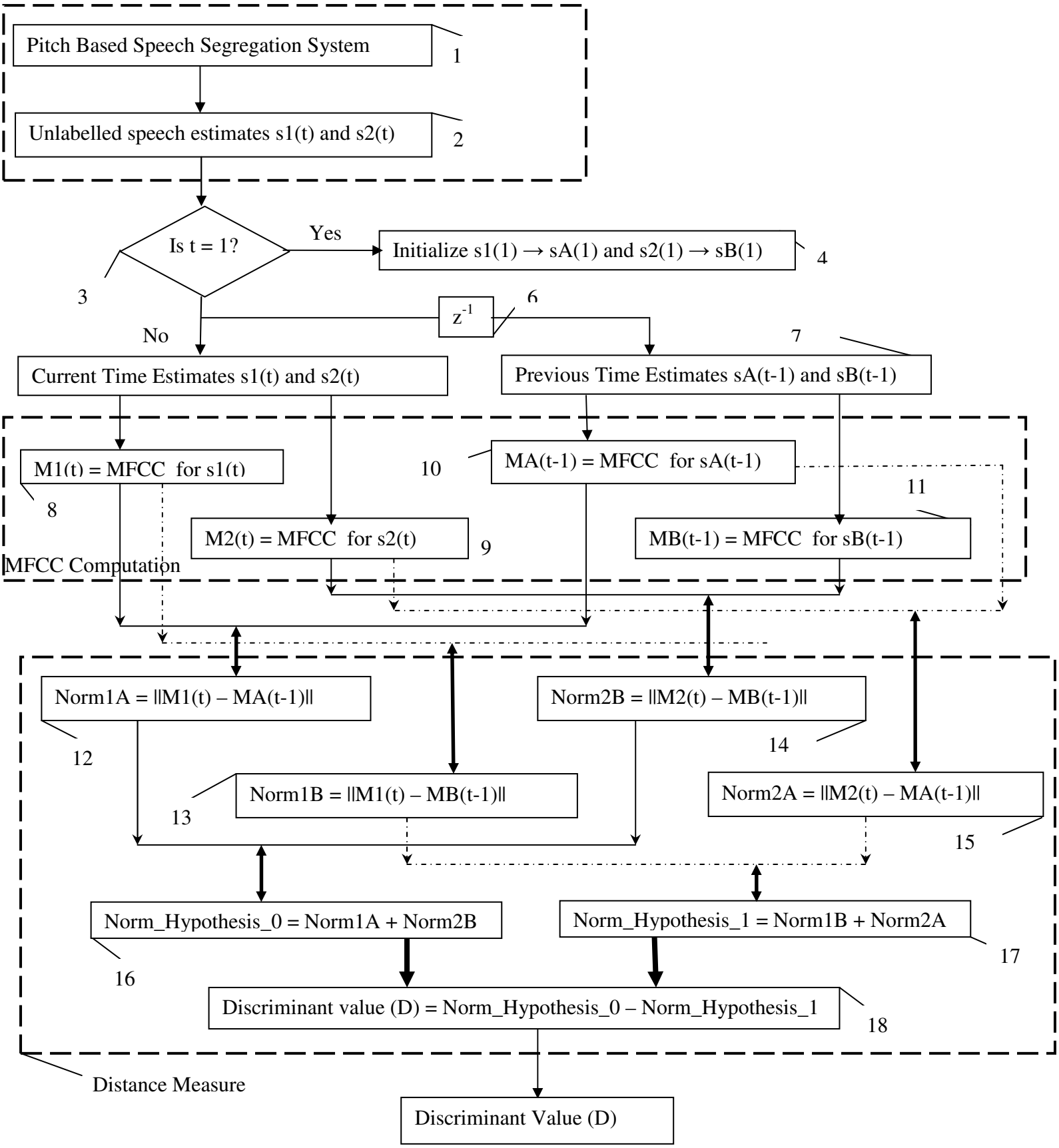
4.7 Intra-segment Sequential Grouping

In the proposed algorithm, the well known MFCCs are used as the features for the speaker assignment problem. The MFCCs of each of the segregated streams $s_1(t)$ and $s_2(t)$ are evaluated to get the features $M_1(t)$ and $M_2(t)$, respectively. These features are then used in combination with the features from the speech of the two speakers A and B in the previous frames, i.e. with $M_A(t - 1)$ and $M_B(t - 1)$. The order of the MFCCs used in the proposed system is 13, including the energy coefficient but not including the difference coefficients. The algorithm is described in the flow chart below. The norm computed in the distance measure block is the Euclidean distance or L^2 norm. Hence, we have a distance measure for every possible connection of the speech frames in the current time with the speech frames in the previous time. In our case,

given we have 2 speech frames in time *t* and *t – 1*, there can be two hypothesis to test namely,

1. Hypothesis 0 = s1(t) → sA(t-1) and s2(t) → sB(t-1)
2. Hypothesis 1 = s1(t) → sB(t-1) and s2(t) → sA(t-1)

Based on the norm values, a discriminant feature D is computed. The sign of D and its magnitude are used in making the decision at a frame. The choice of the hypothesis is decided based on the lowest norm in the assignment. This is shown in the second flowchart in the nearest neighbor classifier block. The reliability threshold δ on the magnitude of D was set to zero in all the experiments and analysis. In the analysis, frames were analyzed in pairs and error locations were noted for each pair.


Pitch Based Speech Segregation System
1
Unlabelled speech estimates s1(t) and s2(t)
2
Is t = 1?
3
Yes
Initialize s1(1) → sA(1) and s2(1) → sB(1)
4
No
z^{-1}
6
Current Time Estimates s1(t) and s2(t)
7
Previous Time Estimates sA(t-1) and sB(t-1)
M1(t) = MFCC for s1(t)
8
M2(t) = MFCC for s2(t)
9
MA(t-1) = MFCC for sA(t-1)
10
MB(t-1) = MFCC for sB(t-1)
11
MFCC Computation
Norm1A = ||M1(t) – MA(t-1)||
12
Norm1B = ||M1(t) – MB(t-1)||
13
Norm2B = ||M2(t) – MB(t-1)||
14
Norm2A = ||M2(t) – MA(t-1)||
15
Norm_Hypothesis_0 = Norm1A + Norm2B
16
Norm_Hypothesis_1 = Norm1B + Norm2A
17
Discriminant value (D) = Norm_Hypothesis_0 – Norm_Hypothesis_1
18
Distance Measure
Discriminant Value (D)

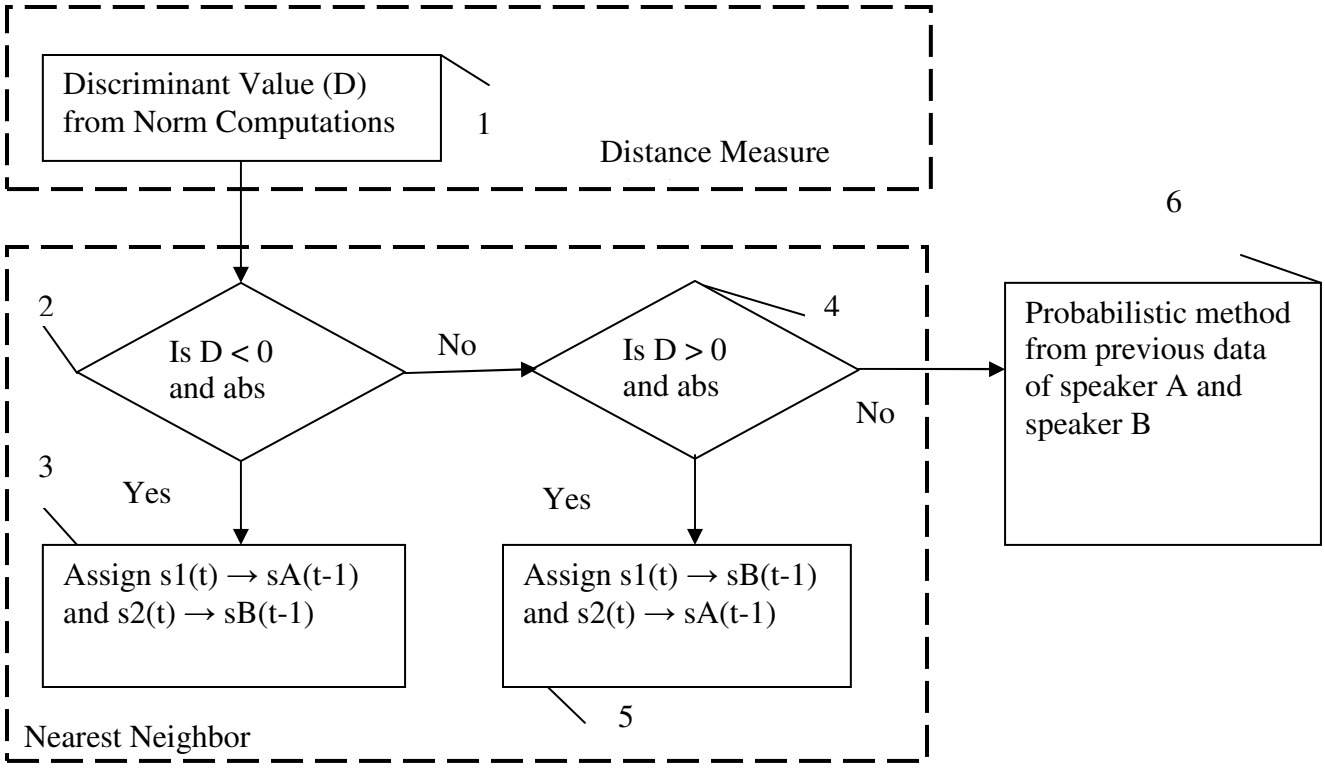


Figure 4.2: Flow chart of the sequential grouping algorithm

4.8 Experiments

4.8.1 Pitch Tracking Algorithm

A sequential grouping algorithm that uses only the pitch values to connect the speech frames have been studied in the literature (Parsons, 1976; Zissman & Seward, 1992). A common problem with this grouping is when we have the pitch tracks crossing; the nearest neighbor rule to assign the pitch values almost always results in an error. In the report by Zissman, they study the assignment on the boundaries of these cross over regions i.e. when the pitch separation between target and masker is below a

threshold (typically 7 or 12Hz). The pitch values that are before and after these critical regions are connected by making a decision on which hypothesis to follow (cross-over of the tracks or no cross-over) using pitch track interpolation. Further, the pitch values within the critical region are assigned randomly to the two speakers as there was no hope for separation of the signal when the pitch values come arbitrarily close. This was a major limitation in the speech segregation system developed by Quatieri and Danisewicz (1990). The SSC database analysis reveals that for the same gender category the average length of the critical region is about 65 ms. In the report by Zissman & Seward (1992), there was no analysis within the critical region. In the algorithm presented, we analyze these regions and identify the accuracy of using the MFCC coefficients in making the assignment. An important reason for the algorithm to perform in the critical region is due to the stable nature of the RLS algorithm in speech segregation.

4.8.2 Analysis of the Algorithm

The use of MFCCs as the features to perform grouping is studied extensively across the SSC database. An interesting observation was made using pre-mixing MFCC features, i.e. MFCCs from the clean pre-mixing speech frames. These features provided near perfect grouping performance in the continuously voiced regions of the target and masker and less than 0.25% error in the transition frames where the signal was beginning to be voiced. Whereas the use of clean pitch values from the pre-mixing utterances provided 98% accurate grouping. These errors were predominantly due to pitch tracks coming close and when they cross over. Hence, there was a strong

motivation to analyze the use of MFCC features from the segregated speech signals (extracted MFCC). It was observed that using the extracted MFCCs from the segregated speech based on clean pitch values provided at least 97% accurate grouping performance. The discussion that follows will analyze the error locations and limitations of MFCC grouping. In Figure 4.3, the brown ellipse highlights a target speaker's voiced segment and the magenta rectangle highlights a masker speaker's voiced segment. The black vertical lines mark the frames where the MFCC norm assignment results in an error. The first error frame between 0.2 and 0.4 seconds occurs in the first 1/3rd of the target speaker's voiced segment and in the last 1/3rd of the masker speaker's voiced segment. In a similar fashion, all the error frames are broken down in to a two dimensional grid with the X-axis corresponding to target speaker's voiced segment and Y-axis being masker speaker's voiced segment. It was found that most of the errors were concentrated towards the beginning and end of the voiced segments. There are three sets of features analyzed in the algorithm namely,

a) Pitch values

b) MFCC Coefficients ($C_0 - C_{12}$)

c) MFCC Coefficients and pitch value (C_0 - $C_{12} - f_0$)

The pitch values used in the experiments below are taken from pre-mixing utterances using Wavesurfer (true pitch values) or estimated from the speech mixture using the multi-pitch algorithm. The MFCC features are computed using the segregated speech signals using the pitch values that are input to the algorithm. The last set of features is a concatenation of both MFCC and pitch together into a single vector and its

combined performance is analyzed. The histogram of these errors relative to the total number of frames analyzed is studied in the following sections.

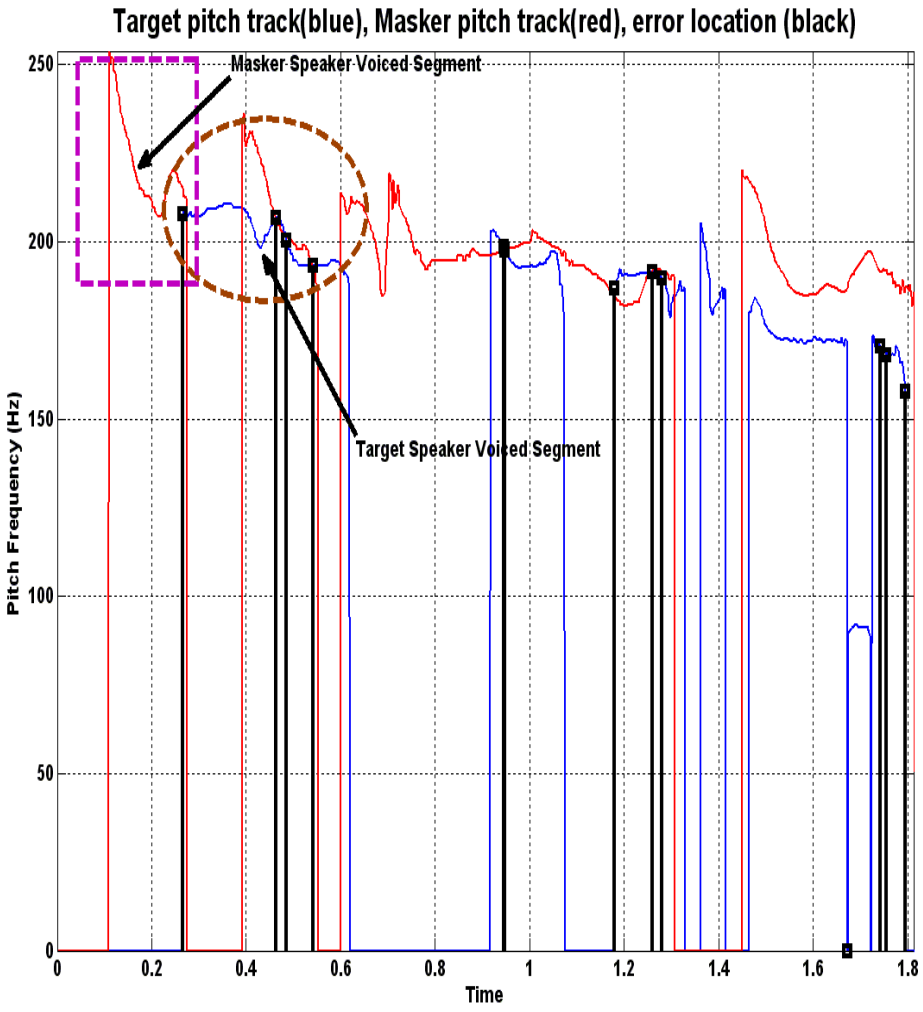


Figure 4.3: Illustration of the error location analysis in sequential grouping algorithm

4.8.3 Experimental Results from True Pitch Values

The pitch feature used as the input to the algorithm is from Wavesurfer. The tables in the appendix B.1 reveal the actual numbers for the error in different regions of the voiced segment. We can observe from Figures 4.4 and 4.5 that the use of MFCC features ($C_0 - C_{12}$) has the error concentration more towards the first one-third and last one-third of the voiced segment for both target and masker speaker. The trend of V shape in the plot reveals the drop in error in the middle of a voiced segment. The combination of MFCC with pitch value was critical in bringing down the error in transition region (first 1/3rd and last 1/3rd) by a factor of two in the same gender category and it was reduced close to zero in different gender case. It should be noted that in all the above analysis we have the pre-mixing pitch values given to us. This is not a practical scenario in which we expect to operate but this provides a ceiling performance of what we can hope to achieve using the algorithm. Further, any error made in one frame at the beginning of a voiced segment will propagate indefinitely till a subsequent change in the assignment is made. This is true with any online tracking algorithm and it is recognized in the algorithm discussed. The numbers reported above are for a particular pair of frames analyzed and the decision made on that pair. This is different from the actual number of frames that are assigned incorrectly which depends on the duration till which a subsequent error is made in the tracking.

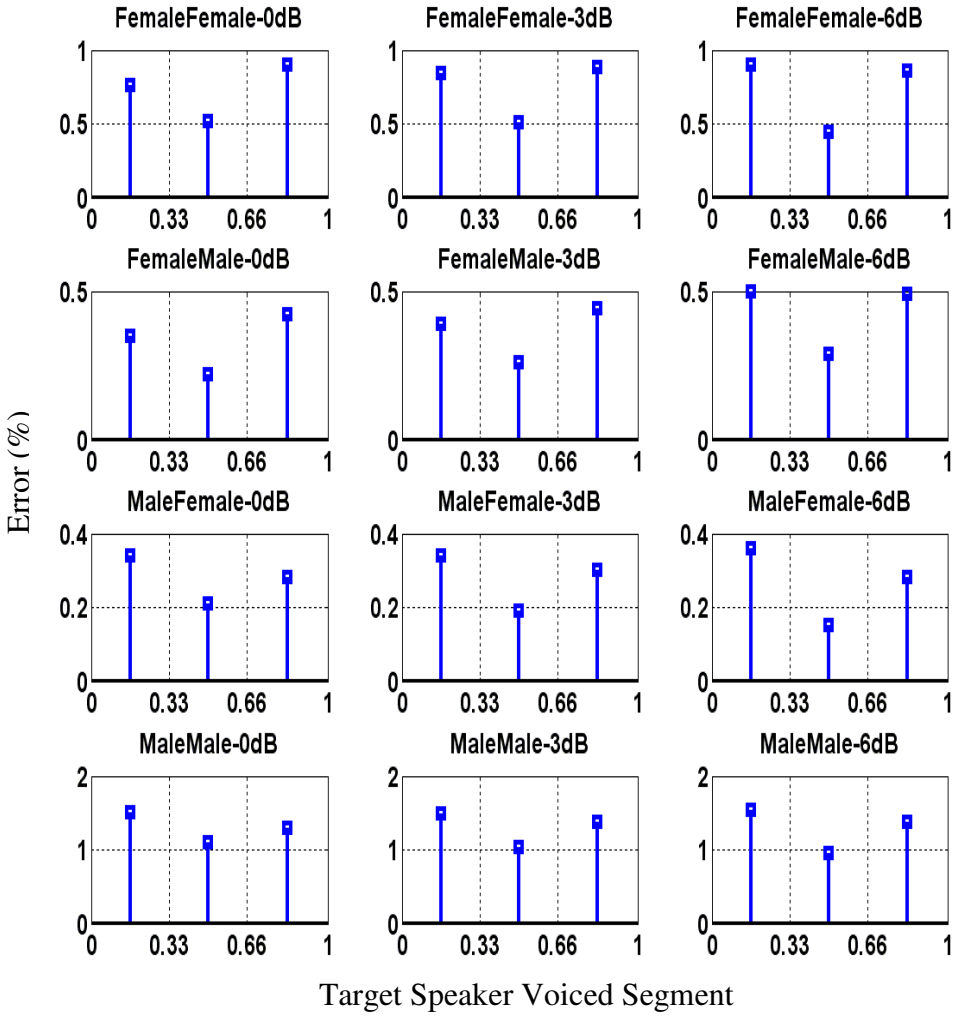


Figure 4.4: Target speaker summary analysis of error location using MFCC ($C_0 - C_{12}$) from true pitch values

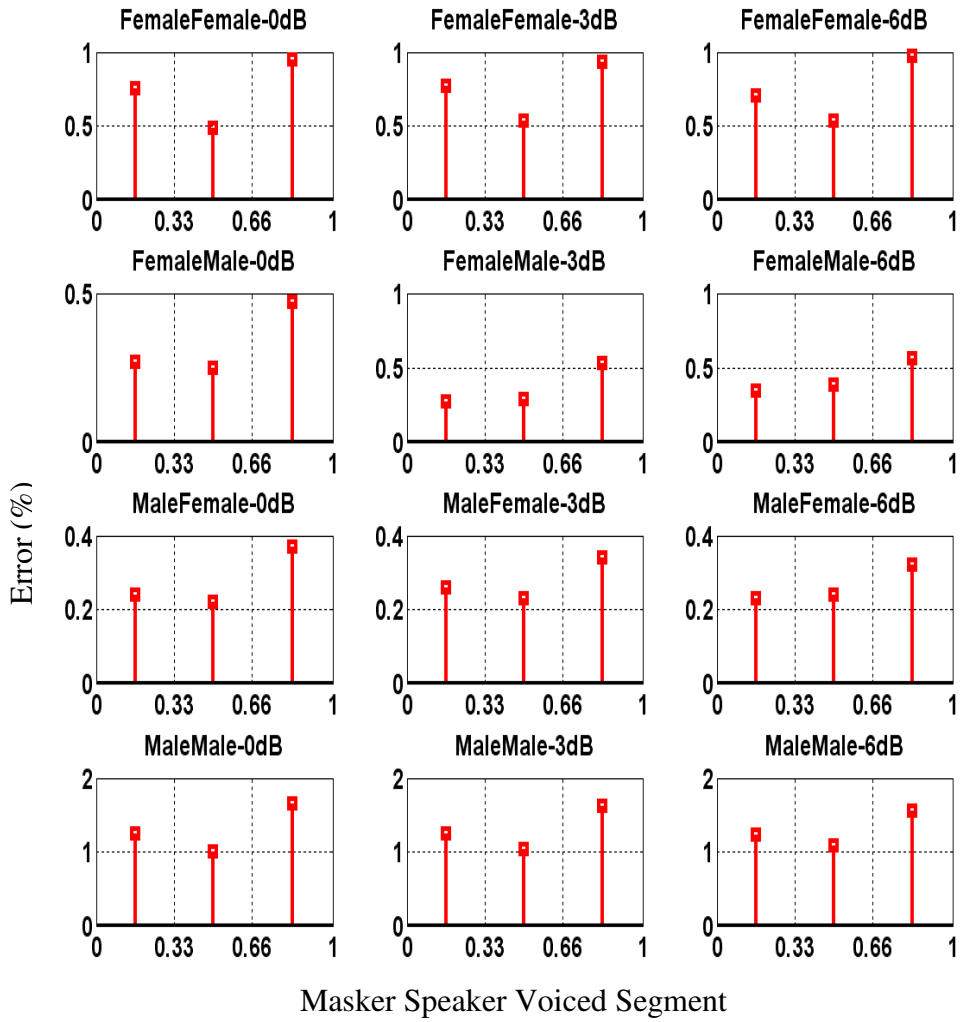


Figure 4.5: Masker speaker summary analysis of error location using MFCC (C_0 – C_{12}) with true pitch values

4.8.4 Experimental Results from Estimated Pitch Values

The results with the estimated pitch values using multi-pitch algorithm is analyzed after the ground truth assignment was made on the estimated pitch tracks using the pre-mixing clean pitch tracks. There is a significant increase in the number of frames that are assigned incorrectly based on the assumed ground truth. The V trend that was

observed with true pitch values is consistent with the estimated pitch values as seen in Figures 4.6 and 4.7. However, the combination of MFCC feature with the estimated pitch values resulted in an increase in assignment error. This is primarily due to pitch octave errors from the multi-pitch estimation algorithm. The table with actual numbers for the error is shown in Appendix B.2

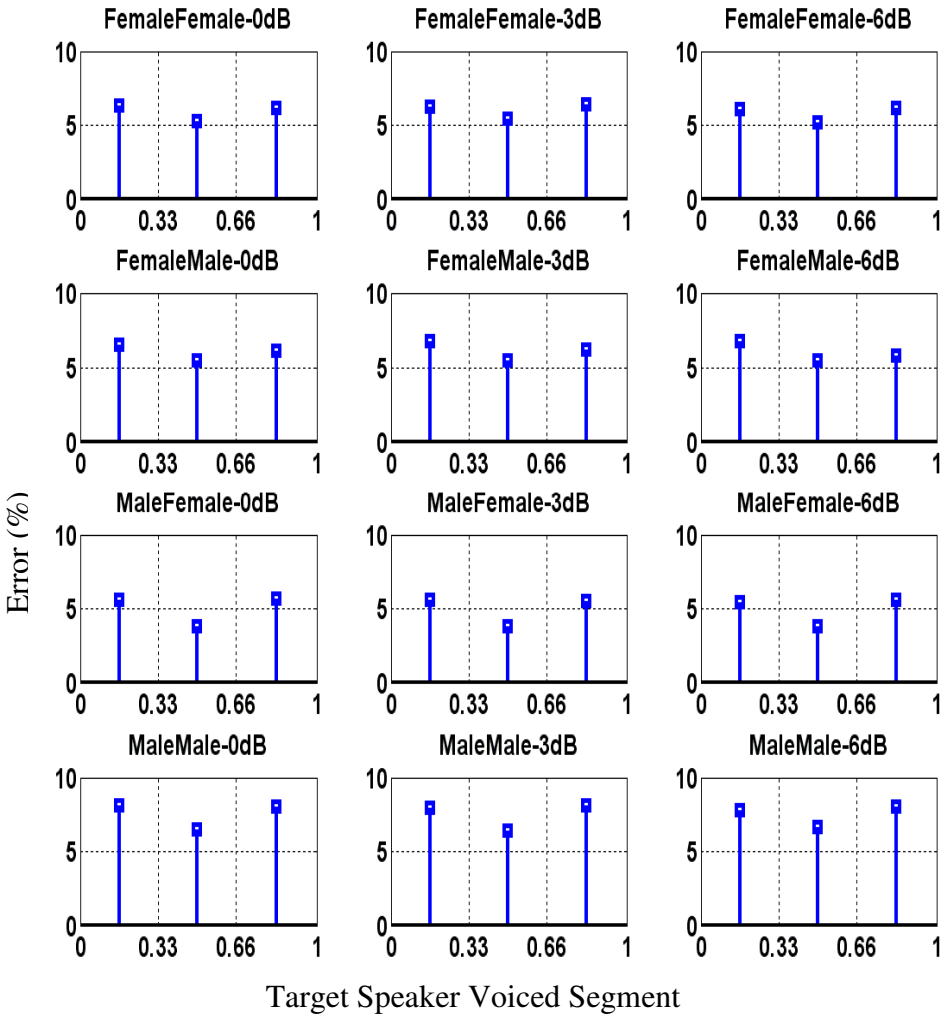


Figure 4.6: Target speaker summary analysis of error location using MFCC ($C_0 - C_{12}$) from estimated pitch values

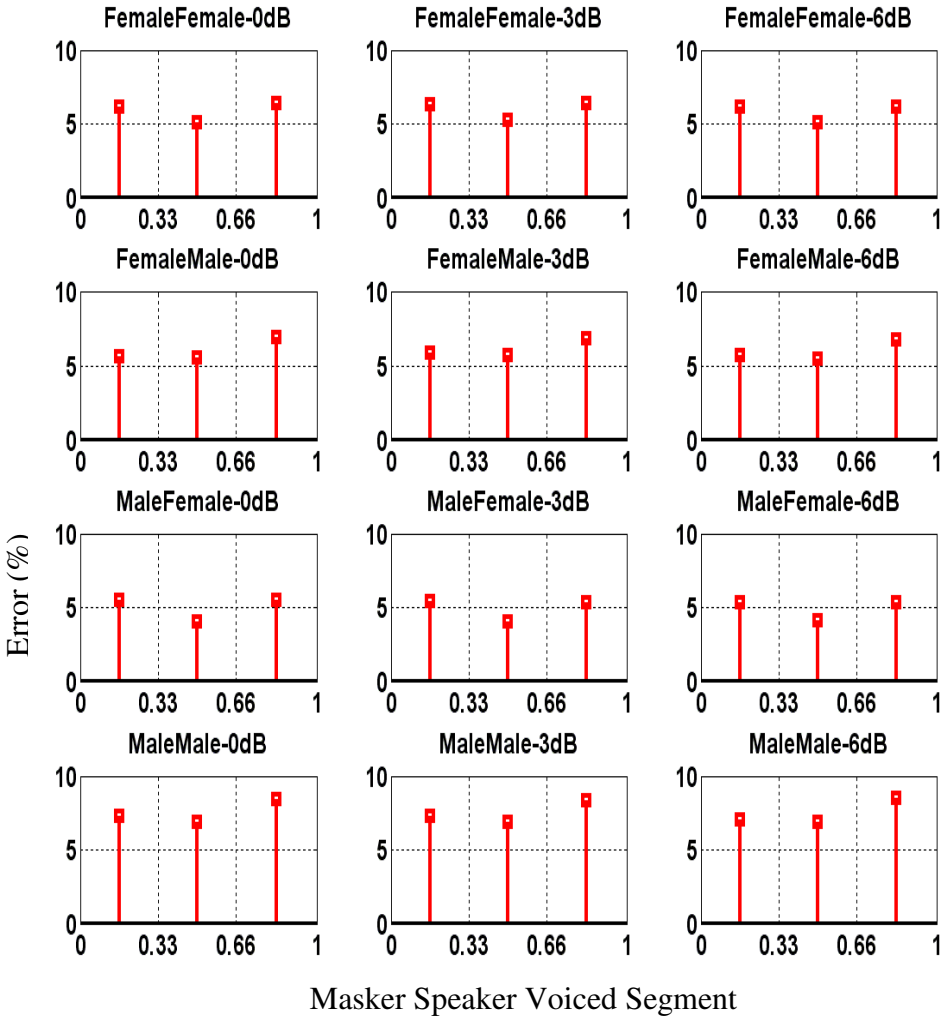


Figure 4.7: Masker speaker summary analysis of error location using MFCC (C_0 – C_{12}) from estimated pitch values

4.8.5 Experimental Results from True Pitch Values in Critical Regions

One of the primary problems in making the assignment is when the pitch tracks come really close. On those frames, the estimated speech signals have more leakage and as a result the estimated MFCC coefficients are not reliable. The error contribution in the different gender category is primarily due to the fact that the female speaker's

pitch is a multiple of male speaker's pitch frequency. In order to narrow down the performance of the system at closely spaced pitch frequencies, the same algorithm was analyzed only on those frames where the two pitch values are separated by less than 8Hz. The system addressed in Zissman and Seward (1992) does not account for any separation on these frames and they make random assignments of the pitch tracks on this region. They highlight the use of cepstral coefficients to connect the speech frames on these critical regions but there was no results reported on those frames. In the summary plots below we study the results for the assignment on these frames using MFCCs ($C_0 - C_{12}$).

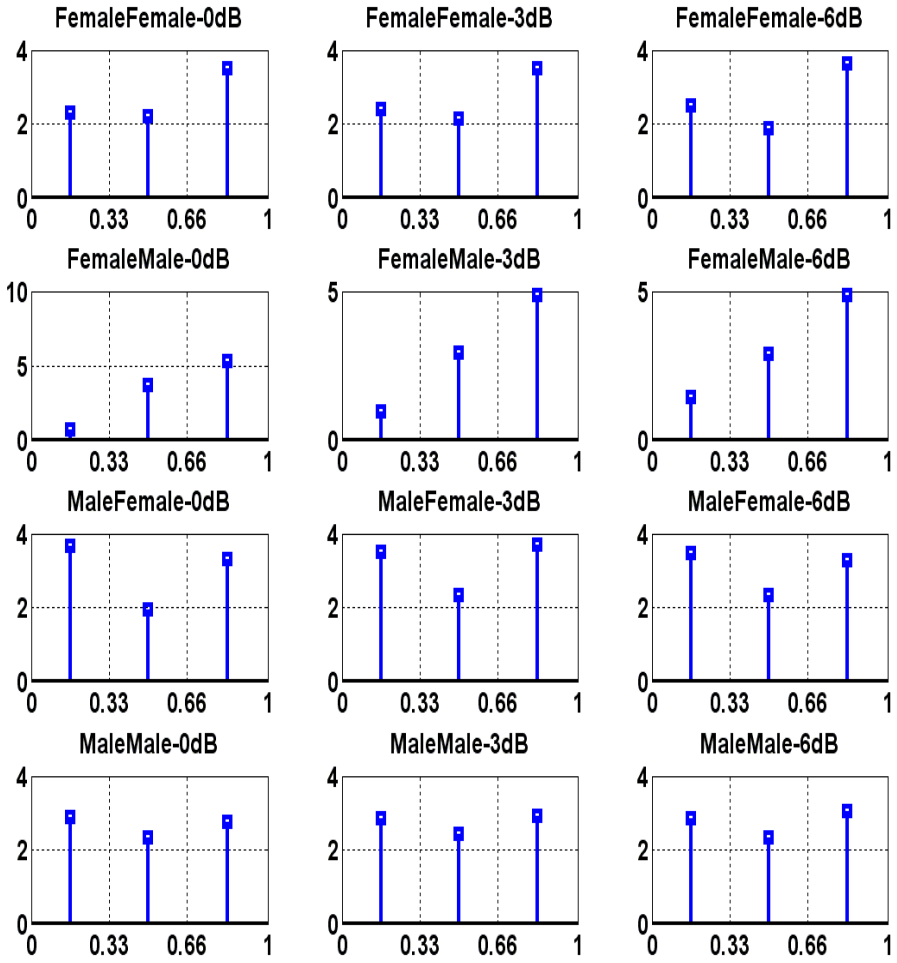


Figure 4.8: Target speaker summary analysis of error location using MFCC ($C_0 - C_{12}$) from true pitch values in the critical regions with pitch separation threshold of 8 Hz.

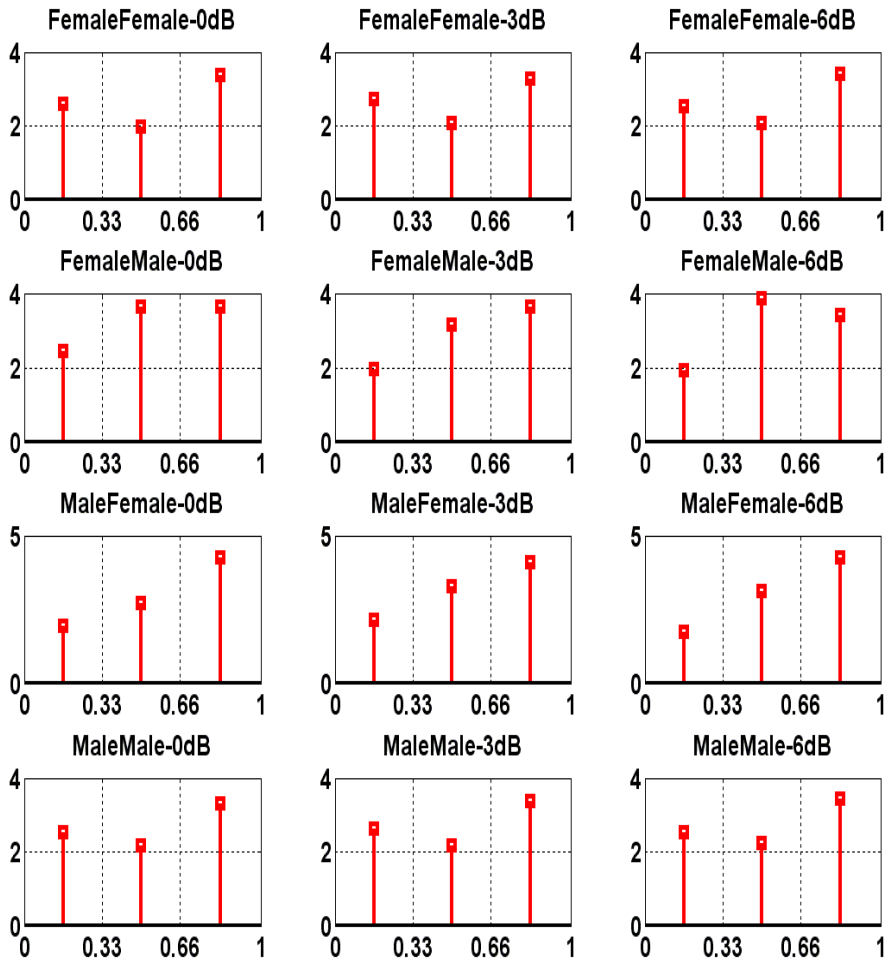


Figure 4.9: Masker speaker summary analysis of error location using MFCC (C_0 – C_{12}) from true pitch values in the critical regions with pitch separation threshold of 8 Hz.

4.9 Discussion

We can observe that using true pitch values in making the assignment has less than 10% error in these critical region frames. However, when the estimated pitch values are used in these regions, the error is at least 50% on all of the above features. This

reveals the difficulty in tracking these regions as estimating the two pitch frequencies on these frames is a challenging task. The weakness of the algorithm is 1) It cannot distinguish which pitch track belongs to target and masker (speaker verification) 2) The assignments made on one frame is not independent of the subsequent assignments i.e. any error made in the first few frames propagates indefinitely.

4.10 Inter-segment Sequential Grouping

In this thesis, we only present an overview of the problem and some literature on the existing methods that address this issue. Inter-segment sequential grouping is the problem of connecting segments of speech that belong to target into one stream and masker into another stream. This problem has been studied in the literature in the name of speaker diarization (Tranter & Reynolds, 2006). Typically in speaker diarization, we have clean non-overlapping speech segments that belong to different speakers and the task is to correctly identify and group the speech segments that belong to a particular speaker. This is inherently difficult when there are no prior speaker models available (usually a Universal background model is available) and a clustering algorithm is employed after all the segments have been analyzed. For example, if we have segments $S = \{S_1, S_2, \ldots, S_k\}$, k is the total number of segments. Each S_i can originate from speaker A or B and the objective is to classify each of these segments into speaker A or speaker B stream. Basically, we look for a partition of S into S_A and S_B which is addressed in multiple hypotheses tracking by Shao and Wang, 2006. They employ prior speaker models and use only the speech segments that are non-overlapping in the co-channel speech.

Chapter 5: Conclusion and Future Work

This thesis has focused on parametric modeling of speech signals and studied its application towards pitch estimation, speech enhancement and speech segregation. A novel pitch estimation algorithm is proposed and the results indicate the superior performance of the algorithm in comparison with several existing PDAs. The speech segregation system was studied in detail and the use of regularized least squares approach towards speech separation was proposed. The RLS algorithm was compared extensively with OLS on the SSC database and the PESQ scores reveal the potential of RLS. We have studied sequential grouping in co-channel speech into two broad categories i.e. intra-segment and inter-segment sequential grouping. An algorithm for intra-segment sequential grouping was explored using MFCC features which provide an alternative to the pitch feature in the grouping process. The potential benefits and pitfalls of the algorithm were analyzed. As the results show, there is a significant difference between pitch tracking results based on a priori versus jointly estimated pitch tracks. Some directions of future research work on these aspects are highlighted below.

5.1 Pitch Detection Algorithm

The use of AIC for regularization mitigates some of the pitch halving error problems but there still remains significant contribution of these errors. This suggests the use of prior information to enforce continuity on the tracks as well as other post processing schemes which can be done by allowing suitable latency. The future work can be directed towards,

- Testing the robustness of the algorithm in the presence of noise.
- Exploring regularization methods to reduce the pitch halving errors.
- Improving the computational performance.

5.2 Gradient Search for Pitch Estimation

Raw pitch estimates indicate high level of accuracy. We observe that there is a great potential to reduce the computational time through intelligent search techniques. The algorithm proposed in the work performs a global brute force search for every speech frame. If we have a good initial estimate of pitch value for the first voiced frame, then we can do a local gradient search for the minima in the residual error. A basic approach towards gradient descent for pitch estimation can be formulated as,

$$f_{0,i} = f_{0,i-1} - \beta * \nabla E(f_{0,i-1}) \tag{5.1}$$

$$E(f_0) = \widehat{\sigma^2(f_0)} = [s - \mathbf{A}(f_0) * \widehat{\boldsymbol{\gamma}}]^{\mathbf{T}}[s - \mathbf{A}(f_0) * \widehat{\boldsymbol{\gamma}}]/\boldsymbol{n} \tag{5.2}$$

$$\nabla E(f_0) = [E(f_0 + \Delta) - E(f_0 - \Delta)]/[2 * \Delta] \tag{5.3}$$

where *i* and *i – 1* are current and previous iteration values, $\beta > 0$ is the step factor in the direction of descent, $E(f_0)$ represents the residual error variance and the gradient of the error variance $\nabla E(f_0)$ is approximated by finite central difference by taking small steps around either side of f_0. The rate of convergence to the local minima is

controlled using the step factor β. For the initial guess for f_0 in every frame, we can use the pitch frequency value estimated in the previous frame. For the first frame in the voiced segment, there can be a global search across the entire pitch range which will guarantee the subsequent minima from the gradient search to be the global minimum. This approach can be extended to multi-pitch estimation as well, where we have two fundamental frequencies to update instead of one. The same approach can be extended to have a vector update instead of a scalar as shown above. This approach was presented in the report by Danieswicz and Quatieri, 1988.

5.3 Two Talker Detection

A fundamental detection problem in co-channel speech is to identify the number of speakers present. In a multi-pitch detector, the number of pitch values estimated from the algorithm can be a used to identify the number of speakers. However, like any PDA there will be insertion errors due to false alarms. Further, the pitch detector is also prone to octave errors like pitch doubling and pitch halving. Given we have two pitch estimates in a frame; we can create a number of hypothesis for different octaves of the estimated pitch frequencies. For example, if we have pitch estimates for a particular frame to be P1 = 150 Hz, P2 = 220 Hz, then we can create the following hypothesis.

Hypothesis	P1 (Hz)	P2 (Hz)
1	150	220
2	75	220
3	300	220
4	150	110
5	75	110
6	300	110
7	150	0
8	75	0
9	300	0
10	0	220
11	0	110

Table 5.1: Illustration of the various hypotheses in a multi-pitch detector

The hypothesis testing problem can be validated using the well-known model selection framework like AIC, BIC or MAP criterion. The ML criterion was shown to have over-fitting problem and needs to be regularized. Further, if the pitch estimation algorithm has deletion errors then there is no hope of identifying it through this method. Hence, the threshold for VAD in the PDA should be tuned to minimize the deletion errors and maximize insertion errors which can hopefully be corrected using the model selection framework outlined in chapter two.

5.4 Sequential Grouping

The system described in this thesis cannot handle unvoiced speech or silence nor can it identify which stream belongs to target and masker. Extension in the direction of inter-segment sequential grouping with a priori speaker models can be a logical next step. An important challenge with online systems is to group the speech on a frame level basis. In this scenario, even if we have plenty of training data for different speakers the test data is typically few milliseconds (typically 20-50ms latency). Further, the maximum length of a test segment can be approximately few hundred milliseconds long (< 400-500ms). Traditionally, speaker identification or verification systems based on the vocal tract system features (MFCC) follow statistical approach (Reynolds et al, 2000). The statistical methods capture the speaker variability in terms of the probability density function (pdf) of the feature vectors of the speaker in the feature space. The performance of these systems depends on the amount of data available for both training and testing. If the data available is small, then the distribution of the feature vectors is sparse, and hence the recognition performance is poor during testing. In the work by Mahadeva Prasanna et al., 2006 they study the use of source and system features for speaker identification and perform extensive comparison of using only source or system features for different sizes of train and test dataset. For small test data size of 1-5 seconds, their study indicate that the source features from the LPC residual captured the speaker information better in recognition accuracy. A detailed study of the database and the algorithm used to extract the features is presented in their paper which can be a useful future direction of research for inter-segment sequential grouping.

Appendices

A.1 Table comparing the performance of the PDAs

PDA	Unvoiced in error (%)	Voiced in error (%)	Gross Errors (%)		Net GE (%)	Absolute deviation (Hz)	
			High	Low		Mean	p.s.d
Male							
CPD	18.11	19.89	4.09	0.64	4.73	2.94	3.60
FBPT	3.73	13.9	1.27	0.64	1.91	1.86	2.89
HPS	14.11	7.07	5.34	28.15	33.49	3.25	3.21
IPTA	9.78	17.45	1.40	0.83	2.23	2.67	3.37
PP	7.69	15.82	0.22	1.74	1.96	2.64	3.01
SRPD	4.05	15.78	0.62	2.01	2.63	1.78	2.46
eSRPD	4.63	12.07	0.90	0.56	1.46	1.40	1.74
mAMDFp	-	-	1.94	2.33	4.27	-	-
SHR	-	-	1.29	0.78	2.07	-	-
ML-AIC (raw)	**8.69**	**7.59**	**0.21**	**0.44**	**0.65**	**1.60**	**1.92**
ML-AIC (filtered)	**5.68**	**6.48**	**0.18**	**0.86**	**1.04**	**1.77**	**2.33**
Female							
CPD	31.53	22.22	0.61	3.97	4.58	6.39	7.61
FBPT	3.61	12.16	0.60	3.55	4.15	5.40	7.03
HPS	19.10	21.06	0.46	1.61	2.07	4.59	5.31
IPTA	5.70	15.93	0.53	3.12	3.65	4.38	5.35
PP	6.15	13.01	0.26	3.20	3.46	6.11	6.45
SRPD	2.35	12.16	0.39	5.56	5.95	4.14	5.51
eSRPD	2.73	9.13	0.43	0.23	0.66	4.17	5.13
mAMDFp	-	-	0.63	2.93	3.56	-	-
SHR	-	-	0.75	1.69	2.44	-	-
ML-AIC (raw)	**4.26**	**14.4**	**0.06**	**2.02**	**2.08**	**3.96**	**4.37**
ML-AIC (filtered)	**2.05**	**13.91**	**0.04**	**1.86**	**1.90**	**4.02**	**4.5**

Table A.1.1: PDA evaluation for male speech (top) and female speech (bottom)

A.2 Performance of intra-segment sequential grouping using true pitch values

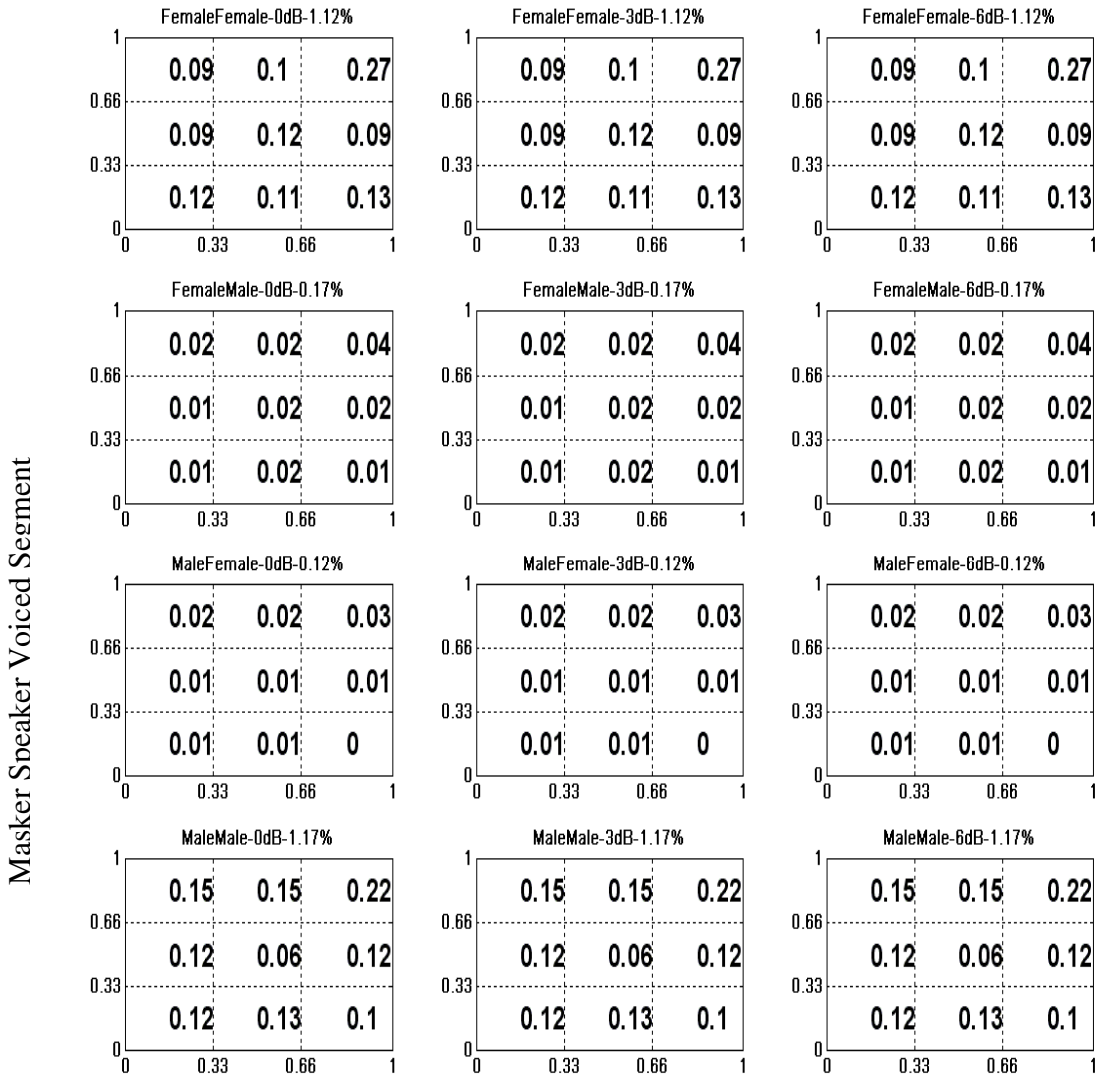


Table A.2.1: Pitch feature used in the analysis of intra-segment sequential grouping using true pitch values

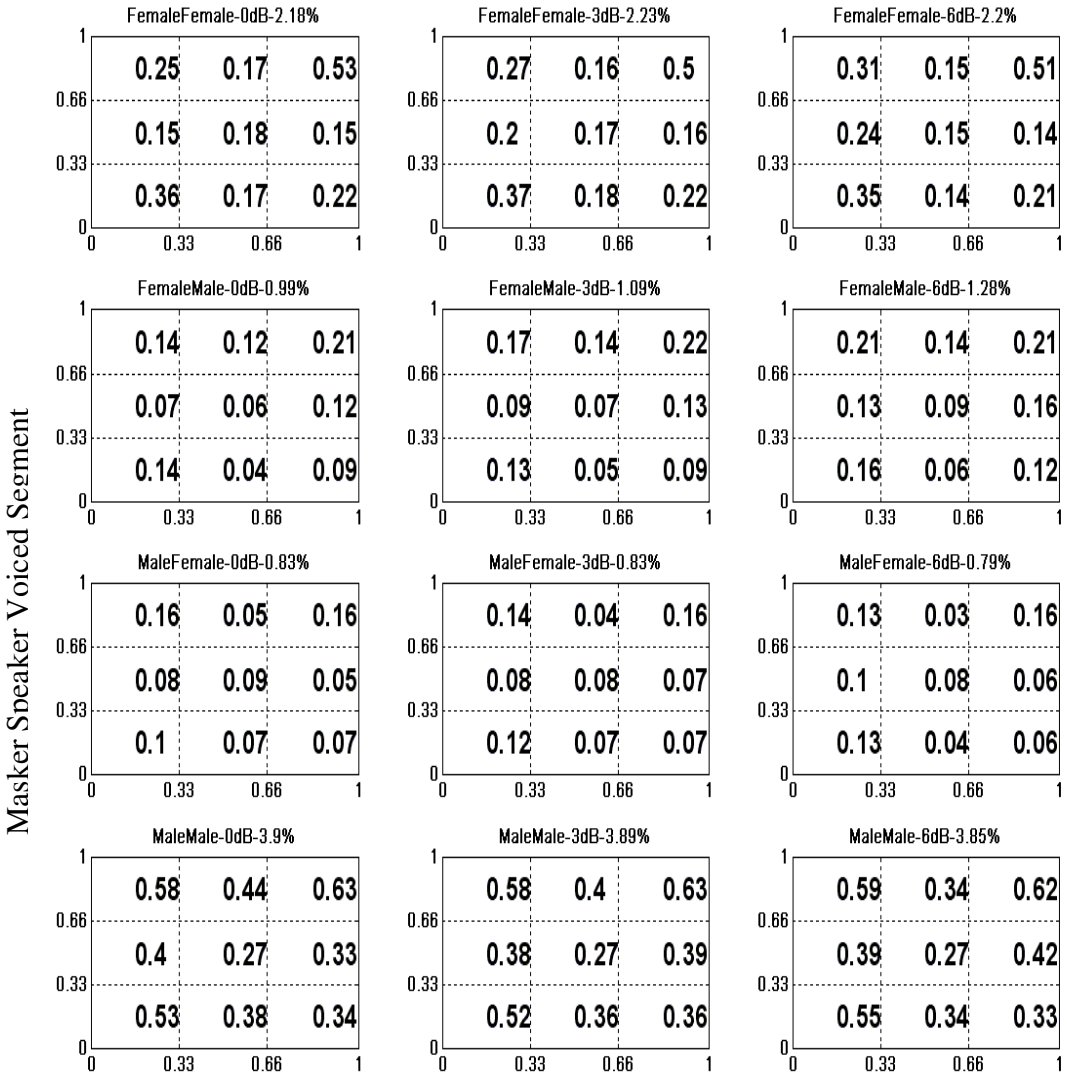


Table A.2.2: MFCC Coefficients (C_0 – C_{12}) used in the analysis of intra-segment sequential grouping using true pitch values

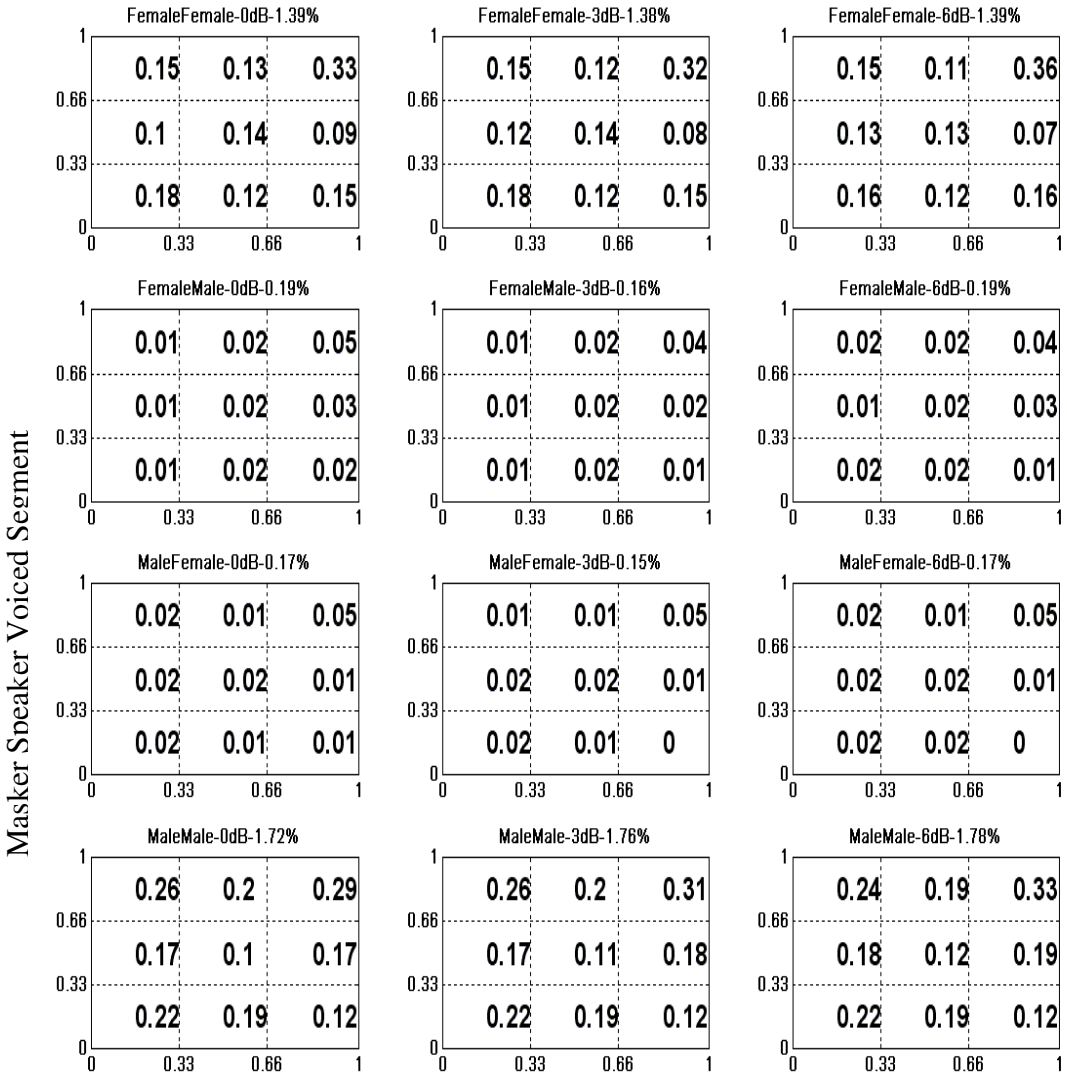


Table A.2.3: MFCC Coefficients and pitch value (C_0 - C_{12} – f_0) used in the analysis of intra-segment sequential grouping using true pitch values

A.3 Performance of intra-segment sequential grouping using estimated pitch values

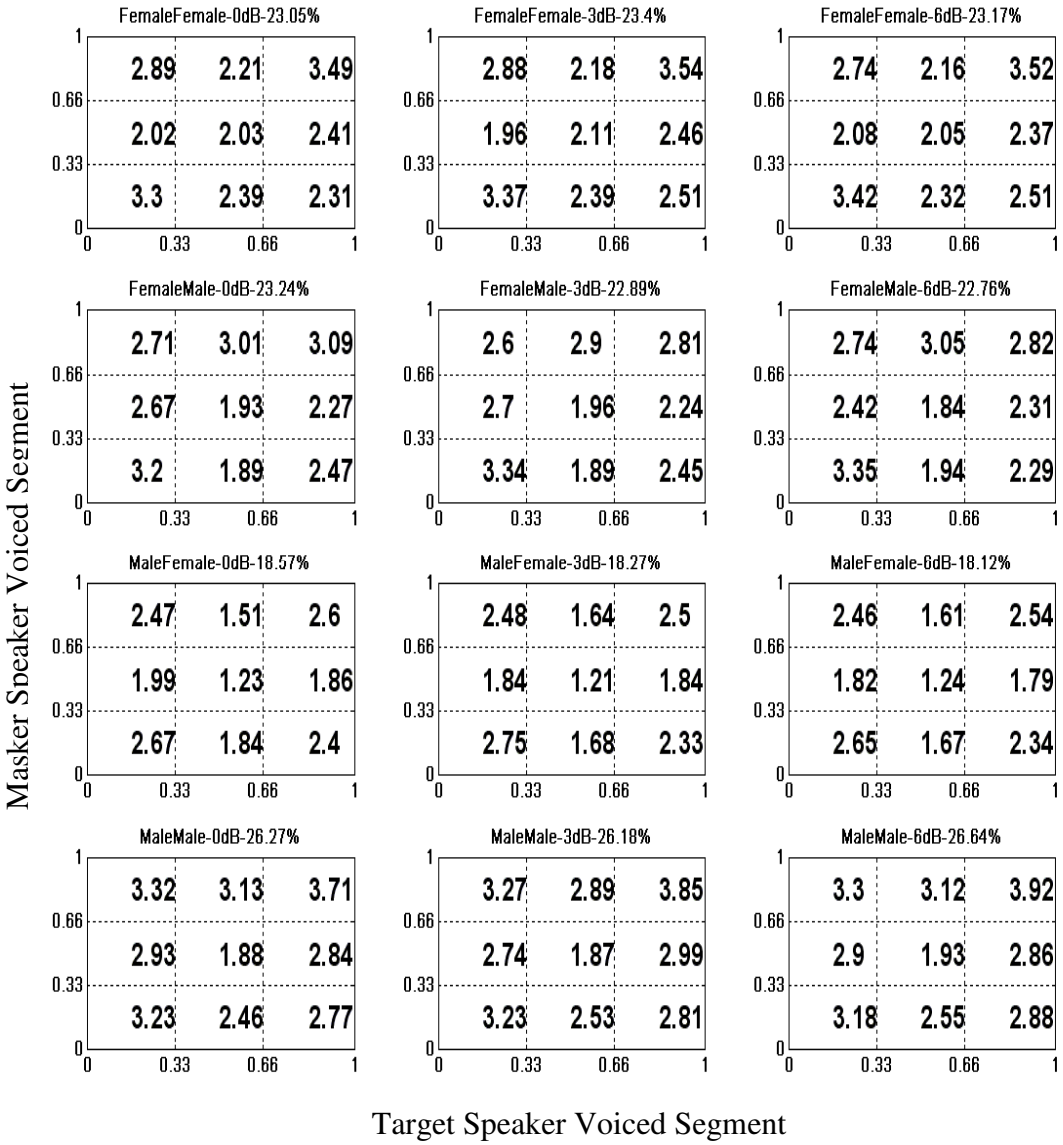


Table A.3.1: Pitch feature used in the analysis of intra-segment sequential grouping using estimated pitch values

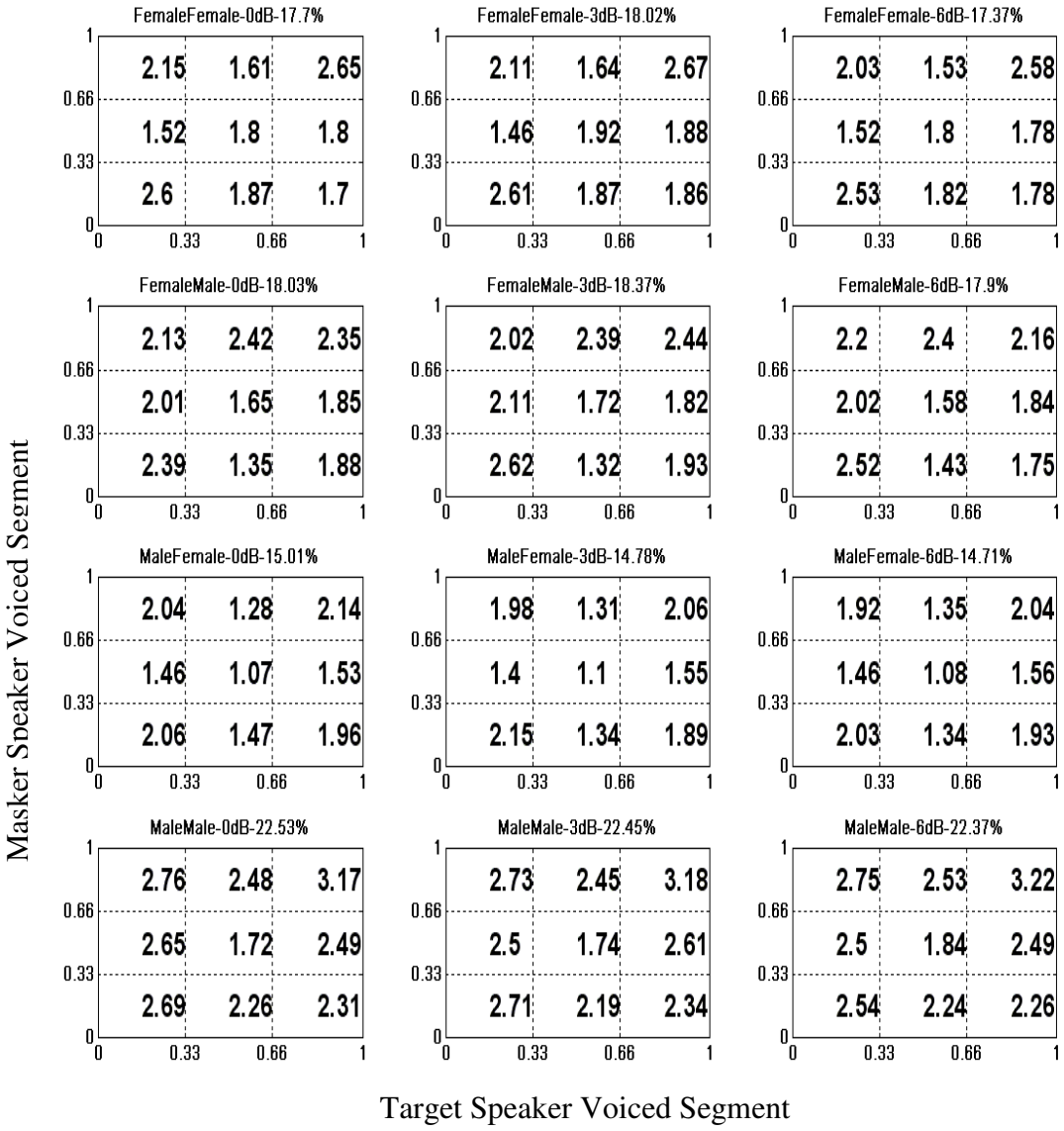


Table A.3.2: MFCC Coefficients (C_0 – C_{12}) used in the analysis of intra-segment sequential grouping using estimated pitch values

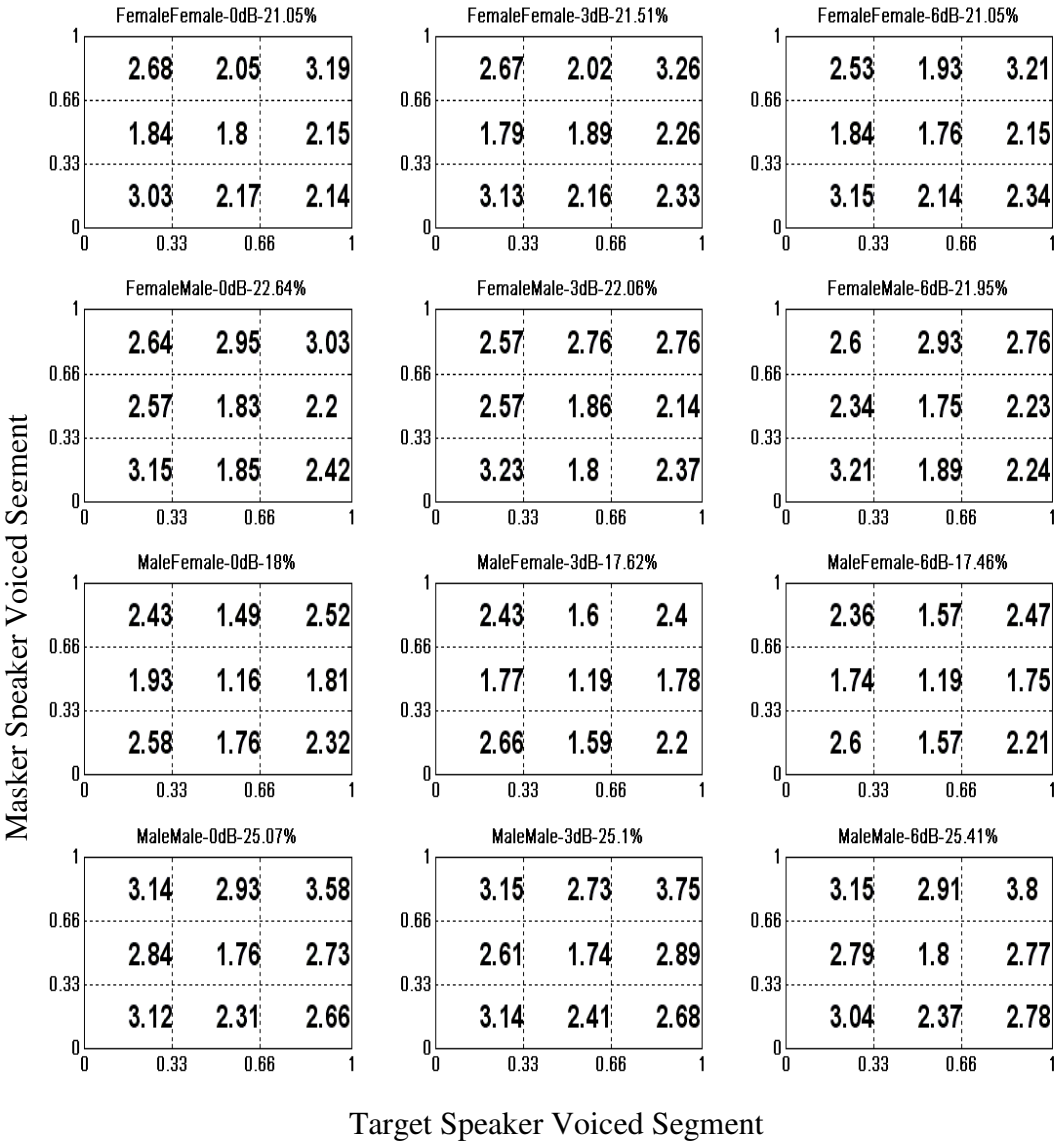


Table A.3.3: MFCC Coefficients and pitch value (C_0 - C_{12} – f_0) used in the analysis of intra-segment sequential grouping using estimated pitch values

A.4 Performance of intra-segment sequential grouping using true pitch values in the critical regions with pitch difference less than 8 Hz

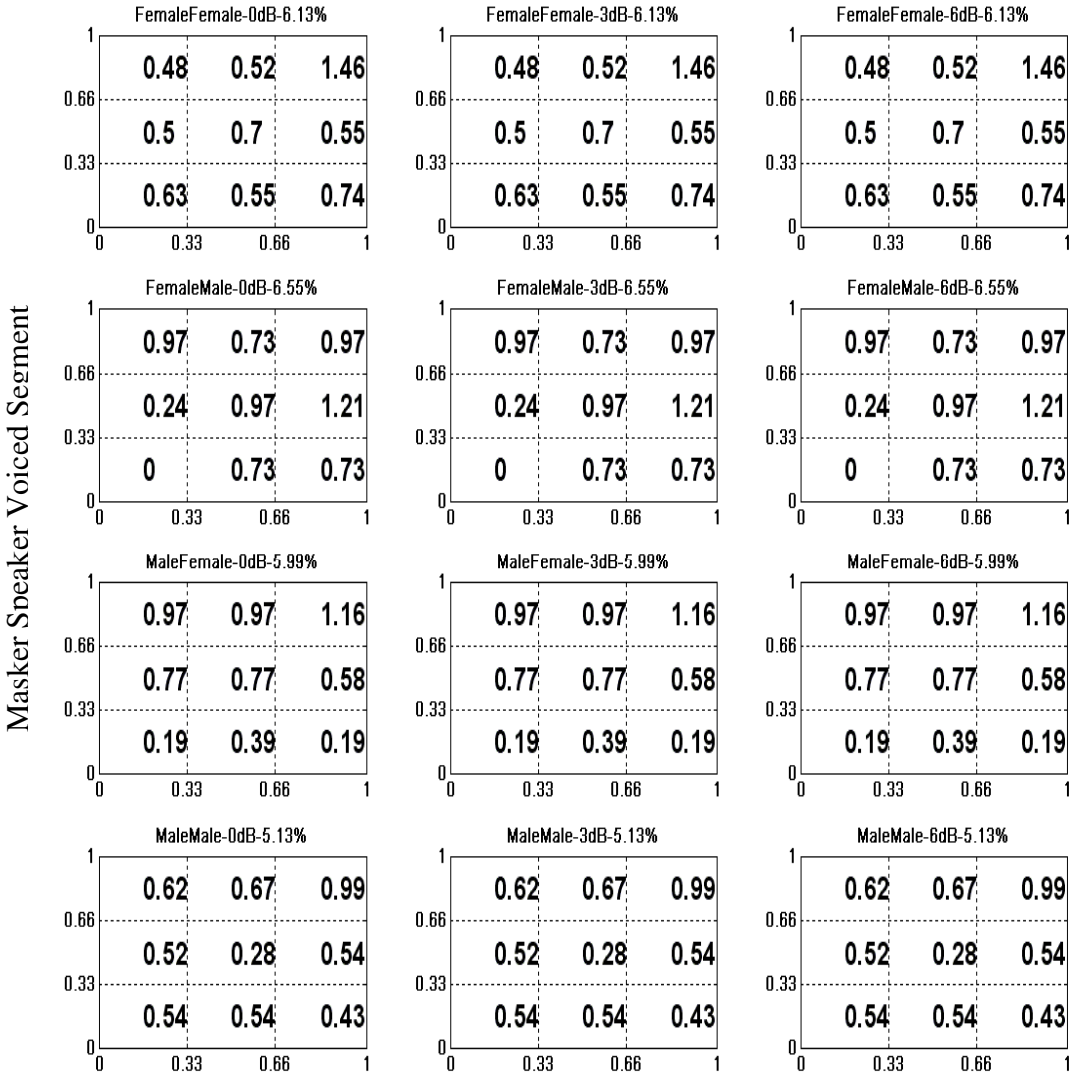


Table A.4.1: Pitch feature used in the analysis of intra-segment sequential grouping using true pitch values on the critical region frames where the pitch difference is less than 8Hz

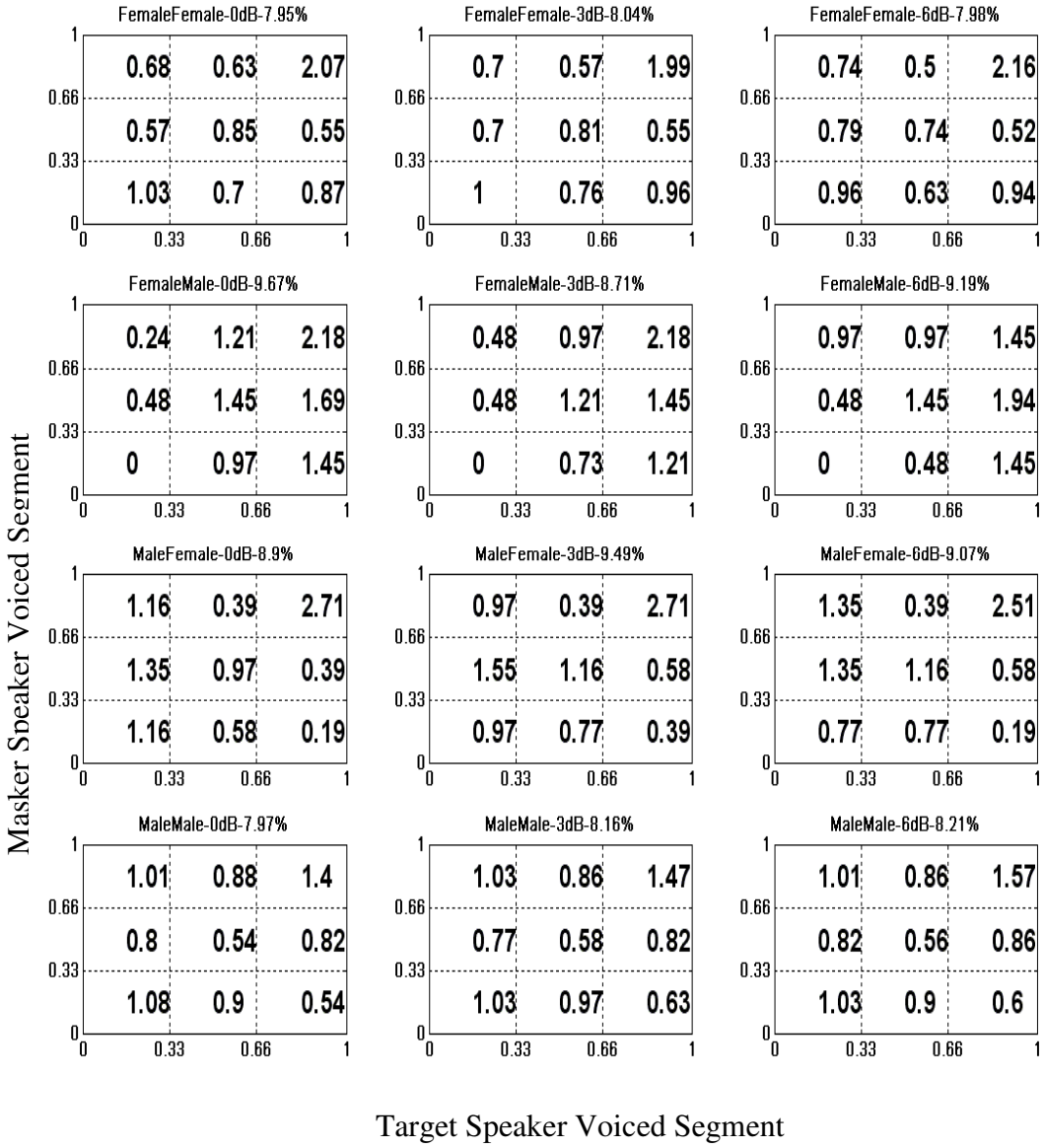


Table A.4.2: MFCC Coefficients (C_0 – C_{12}) used in the analysis of intra-segment sequential grouping using true pitch values on the critical region frames where the pitch difference is less than 8Hz

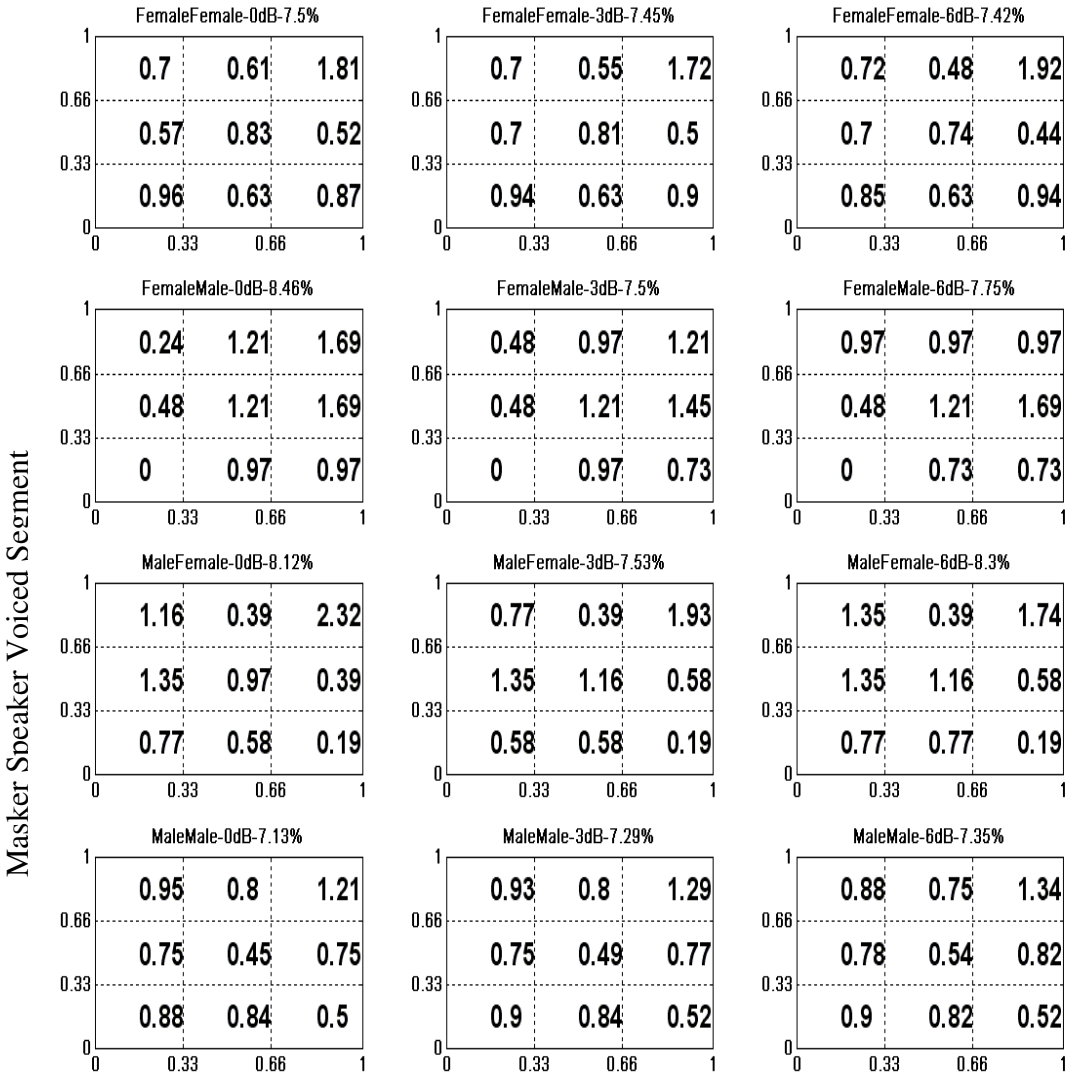


Table A.4.3: MFCC Coefficients and pitch value (C_0 - C_{12} – f_0) used in the analysis of intra-segment sequential grouping using true pitch values on the critical region frames where the pitch difference is less than 8Hz.

Bibliography

[1] Christensen, M. G., Stoica, P., Jakobsson, A. and Jensen, S. H. (2008) "Multipitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr.

[2] Hess, W. J. (1983) "Pitch Determination of Speech Signals – Algorithms and Devices," Berlin, Germany: Springer.

[3] Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., and McGonegal, C. A. (1976) "A comparative study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 399-418, Oct.

[4] Wise, J. D., Caprio, J. R., and Parks, T.W. (1976) "Maximum likelihood pitch estimation," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP -24, no.5, pp. 418-423, Oct.

[5] Steiglitz, K., Winham, G. and Petzinger, J. (1975) "Pitch extraction by trigonometric curve fitting," IEEE Trans. Acoust. Speech, Signal Processing (Special Issue on 1974 Arden House Workshop on Digital Signal Processing) (Corresp), vol ASSP-23, pp. 321-323, June.

[6] Quatieri, T. F. (2002) "Discrete-time speech signal processing – principles and practice," Delhi, India: Pearson Education, Inc.

[7] Tabrikian, J., Dubnov, S. and Dickalov, Y. (2004) “Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model,” IEEE Trans. Acoust., Speech and Audio Processing, vol.12, no. 1, pp 76-87, Jan.

[8] Goldstein, J. L. (1973) “An optimum processor for the central information of pitch of complex tones,” J. Acoust. Soc. Amer., vol. 54, pp. 1496-1516.

[9] Gong, Y. and Haton, J. (1987) “Time domain harmonic matching pitch estimation using time-dependent speech modeling,” IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-35, no. 10, Oct.

[10] McAulay, R. J. and Quatieri, T. F. (1986) “Speech analysis-synthesis based on a sinusoidal representation,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp. 744-754.

[11] Arruda, J. R. F. (2010) “A robust one-dimensional regressive discrete fourier series,” Mechanical Systems and Signal Processing, vol. 24, Issue 3, pp 835-840, Apr.

[12] Rao, C. R., Toutenburg, H., Shalabh, Heumann, C. (2008) “Linear models and generalizations – least squares and alternatives,” Berlin, Germany: Springer-Verlag.

[13] Burnham, K. P. and Anderson, D. (2002) “Model selection and multimodel inference: A practical information theoretic approach,” New York, Springer.

[14] Bagshaw, P. C. (1994) “Automatic prosody analysis,” PhD thesis, University of Edinburg, Scotland, UK.

[15] Noll, A. M. (1967) “Cepstrum pitch determination,” Journal of the Acoustical Society of America, 41(2):293-309.

[16] Phillips, M. S. (1985) “A feature-based time domain pitch tracker,” Journal of the Acoustical Society of America, 77:S9-S10(A).

[17] Schroeder, M. R. (1968) “Period histogram and product spectrum: New methods for fundamental frequency measurement,” Journal of the Acoustical Society of America, 43(4):829-834.

[18] Secrest, B. G. and Doddington, G. R. (1983) “An integrated pitch tracking algorithm for speech systems,” In Proc. IEEE ICASSP-83, pp 1352-1355.

[19] Gold, B. and Rabiner, L. (1969) “Parallel processing technique for estimating pitch period of speech in time domain,” Journal of the Acoustical Society of America, 46(2, part 2):442-448.

[20] Medan, Y., Yair, E. and Chazan, D. (1991) “Super resolution pitch determination of speech signals,” IEEE Trans. Signal Processing, ASSP-39(1):40-48.

[21] Ying, G. S., Jamieson, L. H. and Mitchell, C. D. (1996) “A probabilistic approach to AMDF pitch detection,” Spoken Language, 1996, ICSLP 96.

Proceedings., Fourth International Conference on, vol 2 no., pp. 1201-1204, 3-6 Oct.

[22] Sun, X. (2000) "A pitch determination algorithm based on subharmonic-to-harmonic ratio," the 6th International Conference of Spoken Language Processing, Beijing, China, 4, pp 676-679.

[23] Mahadevan, V. and Espy-Wilson, C. Y. (2011) "Maximum likelihood pitch estimation using sinusoidal modeling," International Conference on Communications and Signal Processing (ICCSP), Feb 10-12 *(accepted).*

[24] Wang, D. L. and Brown, G. J. (2006). Eds., *Computational auditory scene analysis: Principles, algorithms and applications*. IEEE Press/Wiley-Interscience.

[25] Hohmann, V. (2002) "Frequency Analysis and Synthesis Using a Gammatone Filterbank", J. Acta Acoustica, vol. 88, pp. 433-442.

[26] Vishnubhotla, S. and Espy-Wilson, C. Y. (2009) "An algorithm for speech segregation of co-channel speech," in Acoust., Speech & Signal Pro. 2009, IEEE Intl. Conf. on, April, pp. 109–112.

[27] Vishnubhotla, S. and Espy-Wilson, C. Y. (2008) "An algorithm for multipitch tracking in co-channel speech," in Proc. of the Intl. Conf. on Spoken Language Processing (Interspeech 2008), Sep.

[28] Cooke, M., Hershey, J. R. and Rennie, S. J. (2010) “Monaural speech separation and recognition challenge,” Comput. Speech Lang., vol. 24, no. 1, pp. 1–15.

[29] Plack C. J., Fay, A. J. O. R. R. and Popper A. N. (2005) Eds., *Pitch: Neural Coding and Perception*, Springer Handbook of Auditory Research. Springer Science.

[30] Zissman, M. A. (1991) “Cochannel talker interference suppression,” MIT Technical Report 895, July 26.

[31] Danisewicz, R. G. and Quatieri, T. F. (1988) “An approach to co-channel talker interference suppression using a sinusoidal model for speech,” MIT Technical Report 794, Feb 5.

[32] Quatieri, T. F. and Danisewicz, R. G. (1990) "An approach to co-channel talker interference suppression using a sinusoidal model for speech," IEEE Transactions on Acoustics, Speech and Signal Processing, , vol.38, no.1, pp.56-69, Jan.

[33] Zissman, M. A. and Seward, D. C. (1992) “Two-talker pitch tracking for co-channel talker interference suppression,” MIT Technical Report 951, Apr 30.

[34] Loizou, P. C. (2007) *Speech Enhancement: Theory & Practice (Signal Processing and Communications),* CRC Press, June 7.

[35] Bagshaw, P. C., Hiller, S. M. and Jack, M. A. (1993) "Enhanced pitch tracking and the processing of F0 contours for computer and intonation teaching," Proc of European Conference on Speech Communication (EuroSpeech), pp 1003- 1006.

[36] Rix, A. W., Beerends, J. G., Hollier M. P. and Hekstra, A. P. (2001) "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T recommendation P.862, Feb.