LIP SHAPE ESTIMATION FROM SPEECH WAVEFORM

A dissertation

submitted in the partial fulfillment of the requirements

for the degree of

Master of Technology

by

Vikash Sethia

(Roll No. 02307023)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering

Indian Institute of Technology, Bombay

June, 2004

Indian Institute of Technology, Bombay

Dissertation Approval

Dissertation entitled : **Lip Shape Estimation from Speech Waveform**

by **Vikash Sethia**

is approved for the award of the degree of **MASTER OF TECHNOLOGY** in **Electrical Engineering** with specialization of **Electronic Systems**.

Supervisor	_____	(Prof. P. C. Pandey)
Internal Examiner	_____	(Prof. Preeti Rao)
External Examiner	_____	(Dr. V. K. Madan)
Chairperson	_____	(Prof. B. Bandyopadhyay)

Date: 19-06-2004

Vikash Sethia / Prof. P. C. Pandey (supervisor): "Lip Shape Estimation from Speech Waveform", *M.Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology, Bombay, June 2004.

Abstract

Display of lip shape can be employed for providing visual information to improve speech reception by the hearing impaired persons during telephonic conversation. The objective of this project is to investigate estimation of lip-shape parameters from speech waveform without the knowledge of speech content. Estimation of lip-shape parameters is carried out from the spectral moments of the pitch synchronously computed speech spectra using bivariate least squares approximation, and 2D and 3D Delaunay triangulation methods. Lip-shape parameters extracted from these methods are first analyzed for the training sounds. Segments of vowels /*a*/, /e/, /i/, /o/ and /u/ are used as the training sounds. The spectral moments of the training sounds are then used as the reference data for synthesizing lip shapes for other segments like vowel-vowel and vowel-semivowel-vowel syllables. Analysis results have been consistent for speech signal with SNR greater than 20 dB. Also, lip-shape parameters estimated using 3D Delaunay triangulation method have smooth contour than those using the other two methods.

Acknowledgement

I would like to express my deep sense of gratitude to Prof. P. C. Pandey for his invaluable help and guidance during the course of the project. I am highly indebted to him for constantly encouraging me by providing his critics on my work. I am thankful to Mr. Milind Shah for his invaluable help and guidance on this project. I am also thankful to my friends, Alice Cheeran, Vidhyadhar Kamble, Santosh S. Pratapwar, Vinod Pandey, Anil Luthra and all other colleagues of SPI Lab for their support and valuable suggestions during this project.

Vikash Sethia
June, 2004

Contents

List of figures

Chapter 1
Introduction

1.1 Problem overview

For persons with normal hearing, the process of learning to speak is aided by auditory feedback. Hearing impaired persons lack this auditory feedback, and therefore, experience difficulty in acquiring normal speech characteristics. Thus, in spite of proper speech production mechanism, these persons may not be able to produce intelligible speech. Visual feedback can be provided by display of certain speech parameters which provide necessary cues for uttering a particular speech segment. The parameters can be vocal tract shape, energy and pitch contours, and lip shape, all related to speech segment. Hearing impaired persons can be provided speech training using display of such parameters on a screen. Also, the hearing impaired persons retain the ability to speak clearly, if their hearing loss occurred at a later stage in their life. For them, the ability to speak in spite of hearing loss provides easier adaptability to training for speech reading [1].

Apart from speech assisted devices like hearing aids and communication techniques like sign language, the hearing impaired persons rely on lip reading and audio visual speech perception. Hearing impaired persons, who have the ability to speak and lip-read, can make substantial use of visual information from speaker's face on a videophone during telephonic conversations. For them, the visual information from the speaker's face can efficiently integrate or even substitute audio information for understanding speech [1] [2]. Reported results showed that the speech recognition score in noise can be improved upto 43% by speech reading and 31% by audio visual speech perception when compared to auditory alone. However, bandwidth limitations and storage constraints do not allow the videophones to send every visual frame with the audio [1]. It thus necessitates the extraction

of certain speech parameters which provide good correlation between speech and visual information, and which can be transmitted over the channel with minimum capacity.

1.2 Project objective

In the words of Lavagetto [2], "Lip reading represents the highest synthesis of human expertise in converting visual inputs into words and then into meanings. It consists of a personal database of knowledge and skills constructed and refined by training, capable of associating virtual sounds to specific mouth shapes, generally called visemes, and therefore infer the underlying acoustic message. The lip-reader attention is basically focused on the mouth, including all its components like lips, teeth, and tongue, but significant help in comprehension comes also from the facial expression." There is an ongoing project on the estimation of vocal tract shape at IIT Bombay [3] [4] [5]. Its objective is to estimate the vocal tract shape and pitch from the speech signal, and display these in real time. However, displaying only the internal view of the vocal tract shape and the side view of the lip shape cannot provide adequate information to help the hearing impaired persons. An additional display of the front view of the speaker's mouth, mainly the lip shape would provide additional visual information to assist in speech perception by the hearing impaired persons.

The goal of the project is to synthesize lip movements to visualize speech. If such lip motions can be synthesized, hearing impaired persons may be able to recover auditory information by reading lip movements. This also has wide applications in animation and video telephony.

1.3 Organization of the report

Next chapter provides a brief overview of the various methods proposed and implemented for estimating the lip-shapes. Chapter 2 also discusses the use of Delaunay triangulation for connecting a set of points to the desired parameters. In Chapter 3, the method of finding the pitch period of the speech signal and calculation of spectral moments from the signal have been described. The use of least squares approximation, and the surface fitting and the volumetric fitting using Delaunay

triangulation for calculating the lip-shape parameters have been described in the same chapter. Chapter 4 discusses the analysis of the results obtained for the methods described in Chapter 3. Chapter 5 gives the summary of the work done and some suggestions for future work.

Chapter 2
Speech-driven lip-synchronization

Speech driven lip-synchronization (or lip-sync) involves an accurate portrayal of how the lips, tongue, mouth, and jaw of the speaker appear to move during the utterance of the speech deduced from the speech signal without the necessity of speech recognition or previous knowledge of speech [6]. Lip-sync can also be done with the help of content or text of speech. Text-driven lip-sync systems identify the phonemes in the speech. Speech driven lip-sync systems use audio information to identify the mouth position during the speech production [7]. In the next few sections, some of the methods proposed and implemented to synthesize lip shapes have been described. Last section explains the theory of Delaunay triangulation which will be used for mapping a set of points to generate lip-shape parameters.

2.1 Real time DECface

Waters and Levergood [8] developed DECface as a novel form of real time face-to-face communication. The process involves the ability to generate speech and graphics at real time rates where the audio and graphics are tightly coupled to generate expressive synthetic facial characters. The process employs three major algorithms - (1) the letter to sound system, (2) phonemic synthesizer and (3) vocal tract model. The letter to sound system accepts ASCII text as input and produces phonemic transcription as output. The phonemic synthesizer accepts this phonemic transcription output and produces parameter control records for the vocal tract model. The synthesizer applies intonation, duration and stress rules to modify the phonemic representation based on phrase level context. The resulting phonetic sequence is provided to the synchronization component. The vocal tract model accepts the control records from the phonemic synthesizer and updates its internal state in order to produce the next frame of synthesized speech samples.

The vocal tract model is a formant synthesizer based on Klatt's model [9]. Using the prototype mouth shapes for the known phonemes, the model then computes the current mouth shape for the phoneme being uttered, using interpolation. The synchronization component is then used to synchronize the synthesized speech samples with the graphical display of the mouth shape in real time.

2.2 HMM based approach

Tamura et al [10] proposed a technique for synthesizing synchronized lip movements from auditory input speech signals. The technique is based on an algorithm for parameter generation from HMM with dynamic features. In the analysis phase, they extracted lip contours from the audio-visual speech database as the static features. At the same time, Mel-cepstral coefficients were extracted from the speech signal. They trained the syllable HMMs using the parameters extracted from the database. In the synthesis stage, the speech recognition of the auditory input speech was performed based on the syllable HMMs using the maximum likelihood calculation of HMMs. As a result, they obtained a sequence of syllables and state durations for the input speech. The syllable HMMs corresponding to the obtained syllable sequence were then concatenated to obtain a sentence HMM. A sequence of lip contours were generated from sentence HMM using an ML-based parameter generation algorithm.

There is another HMM based speech to lip movement synthesis model with speech as input, which incorporates a forward co-articulation effect, as proposed by Yamamoto et al [11]. In the analysis phase, the acoustic phoneme HMMs were trained from an audio visual speech database. Next the speech parameters were aligned into HMM state sequences using the forced Viterbi alignment. Frames were classified into "viseme" classes corresponding to succeeding phonemes by looking ahead to context independent HMM state sequence. Synchronized lip parameters associated with the same HMM state and same viseme class of the succeeding phoneme context were then extracted to form a visual database. In the synthesis stage, an input speech signal was aligned into an HMM state sequence using Viterbi alignment. Viseme classes corresponding to each frame were then determined and the lip parameters associated with the HMM state and the viseme class were then

extracted and concatenated into a lip movement sequence.

Williams and Katsaggelos [1] proposed an HMM based speech to video speech synthesizer based on a novel correlation HMM model. The unique feature of the model is the ability to integrate independently trained acoustic and visual HMMs for speech-to-visual synthesis. In this approach, the acoustic speech signal is first preprocessed and then the acoustic feature vectors are extracted to construct an acoustic observation sequence. Next the acoustic HMMs are used to realize an acoustic state sequence that best realizes the input speech. The state sequence is then mapped by the correlation model into a visual state sequence. The heart of the correlation model is the integration technique which is divided into three parts - early integration, intermediate integration and late integration. These integration techniques try to integrate the audio and visual HMMs depending on the state of their occurrence. Finally the visual state sequence and the corresponding visual HMMs are used to produce a visual observation sequence for speech reading.

2.3 Synthesis using RGB information

Liéven and Luthon [12] proposed the use of colour video sequence of the speaker's face for segmenting the lip shapes without any reference to audio information. First the logarithmic colour transform from RGB to HI (hue, intensity) colour space is performed on the video sequence to estimate the sequence dependent parameters, and also to estimate the noise on the motion information. Next, the mouth shape is segmented statistically using Markov random field model, and simultaneously, the Region of Interest (ROI) covering the lip shape is estimated using the red-hue predominant regions.

Most of these methods require video information of the speaker's face for generating lip motions. Another requirement is the use of audio-visual speech database containing visemes related to all known phonemes, which in turn requires a lot of storage capacities, thereby increasing the cost. The major drawbacks are the additional hardware required for tracking the speaker's facial parameters as well as limiting facial motion of the speaker.

2.4 Synthesis using neural network

In synthesizing lip movements using audio information only, the most important part is the identification and generation of lip-shape parameters sufficient enough to support lip-reading. Massaro et al [13] developed an audio to visual speech synthesis which synthesized visual speech directly from the acoustic speech waveform. From the acoustic waveform, 13 cepstral coefficients were generated using 21 Mel-scaled filters. At every frame, the cepstral coefficients were fed to a feed-forward artificial neural network with three layers. At any instant, the cepstral coefficients related to 11 consecutive frames, viz 5 previous frames, current frame, and 5 forward frames, were taken as input, yielding a total of 143 inputs. At the network outputs were 37 control parameters of the animated face. Smoothing algorithm was then used to remove the fluctuations in the control parameters. The smoothed parameters were then fed to a face renderer to drive the articulation of the synthetic face. Results showed that the synthesizer detected the middle phonemes in a word more correctly than the phonemes at the extreme positions, but visemes were generated more correctly at the extremes than in the middle.

Lavagetto et al [14] [2] proposed a similar speech to lip movement conversion method based on the time delay neural network. The major differences are the extraction of cepstral coefficients from LPC speech analysis and the architecture of the neural network.

2.5 Use of spectral moments

One method reported by McAllister et al [15] [7] is based on the observation that the basic shape of the Fourier transform for a given speech signal is relatively static and independent of pitch, and can reliably be correlated with the mouth shape of that sound. Hence rather than locating the formants, which is more difficult and error-prone, the spectral moments of the signal can be used to derive its associated viseme.

First the fundamental frequency of the speech signal at successive intervals or equivalently the glottal pulse duration P_g is determined by optimizing the sum of the magnitude of the first four odd samples of the DFTs at these intervals.

Once the P_g sequence is being tracked accurately, the first two central moments, i.e., mean and variance are calculated from the magnitude spectrum of the speech signal with the window length N equal to P_g, after normalizing it like a probability density function. For every frame, the spectral moments are then mapped to mouth parameters using 2D Delaunay triangulation. Next, using Hermite cubic polynomial and mouth parameters as its control parameters, the lip contours were developed for each frame. The method works well for vowels, semivowels and fricatives, but not so well for affricates and stops.

The advantage of this method is that only two spectral moments and the pitch period sequence of the speech signal were used as the control parameters for generating lip shapes from the acoustic waveform. Also, low amount of redundancy is involved in this case, as the spectral moments can not be used to regenerate the speech waveform. Audio-visual speech database has not been used in this method. Because of these features, this method has been implemented and analyzed with certain modifications in this project as described in the next two chapters.

2.6 Delaunay triangulation

Triangulation algorithm can be used to connect a set of data points to generate a geometric figure. Triangulation is a subdivision of a surface area (volume) into triangles (tetrahedrons). Moreover, a triangulation of a set of points consists of vertices, edges connecting two vertices and faces connecting three vertices. One of the most common methods of achieving triangulation of points is the Delaunay triangulation. Various algorithms have been proposed to obtain Delaunay triangulation like Bowyer-Watson algorithm, Guibas-Stolfi Algorithm, plane-sweep algorithm, divide-and-conquer algorithm etc. [16] [17] [18] [19] [20].

The two properties of Delaunay triangulation are [21]:

1. No point is contained in the circumcircle of any triangle. This empty circle property is used in several Delaunay triangulation algorithms.

2. In 2D only, of all the possible triangulations of a given set of points, the Delaunay triangulation maximizes the minimum angle for all triangular elements which is the requirement for good quality finite elements. Unfortu-

nately, this maximizing the minimum angle property is lost in 3D and higher dimensions where poor quality elements are formed.

When a set of points uniformally distributed in space are connected using 2D Delaunay triangulation, the generated surface takes the shape of a grid with a flexibility of controlling grid resolution. The use of Delaunay triangulation is particularly suited when one does not want to force any constraints on the set of points to be connected. Since the Delaunay triangulation is done over a set of points and does not necessarily conform to imposed boundary (fixed edges), the method can be forced to include the boundary properly. This new forced method is called Constrained Delaunay Triangulation. In this method, the pre-defined edges are in the triangulation and the empty circle property is modified to apply only to points that can be seen from at least one edge of the triangle where the pre-defined edges are treated as opaque. In 3D, there appears to be no concept of constrained Delaunay triangulation.

An example of 2D Delaunay triangulation is shown in Figure 2.1. Another example of 2D Delaunay triangulation, when projected in 3D space and where a set of points in X-Y axes is mapped to an equal number of desired parameters in Z-axis, is shown in Figure 2.2. Using the surface in Figure 2.2, the parameter value for any other point inside the boundary range can be found. This feature would be used for finding lip-shape parameters from the spectral moments of the speech signal as part of the project.

2D Delaunay triangulation can be used for unstructured meshs and structured surface fitting of two variables. Similarly, 3D Delaunay triangulation can be used for volumetric fitting of three variables. In general, both 2D and 3D Delaunay triangulation methods are the fastest methods for unstructured meshes, but the boundary conformance needs to be checked or maintained.

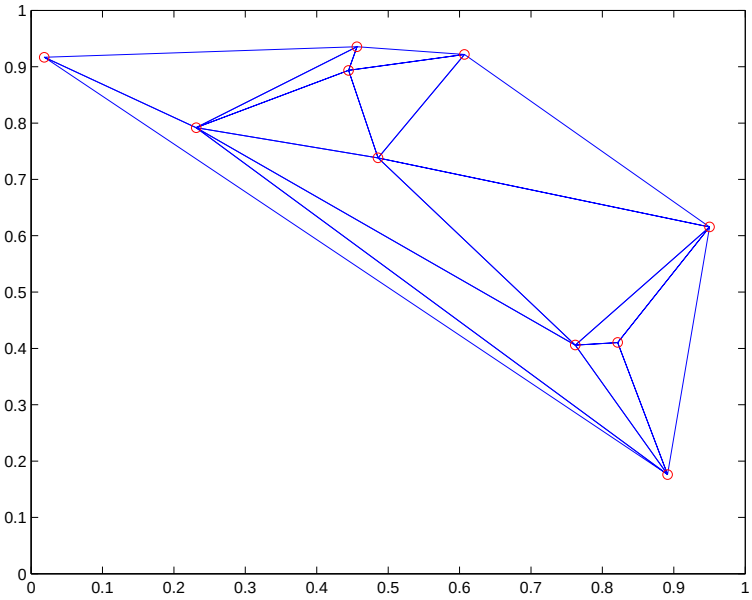

Figure 2.1: 2D Delaunay triangulation of a set of points

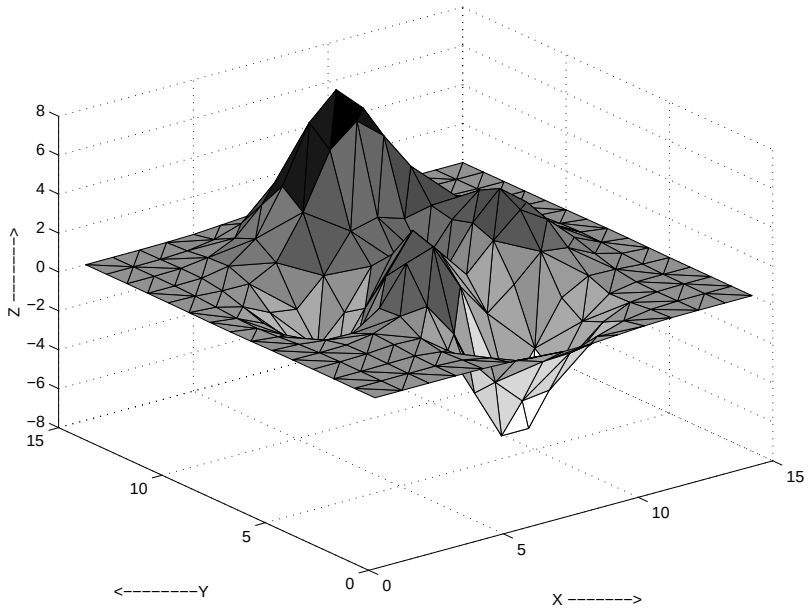

Figure 2.2: 2D Delaunay triangulation of a set of points in 3D space

Chapter 3
Implementation

A Matlab [22] based speech analysis and display package, **VLip**, has been developed for implementation and evaluation of various analysis techniques for lip shape estimation. The natural speech signals were recorded using GoldWave Editor [23] and the synthesized speech signals were generated using Klatt's cascade/parallel formant synthesizer [9]. All the speech signals have been digitized at 11.025 kSa/s. Analysis techniques employed include:

1. Display of speech waveform with an estimate of its pitch period.
2. Display of voiced, and unvoiced segments, which is useful for determining the voice onset.
3. Analysis of the energy of the waveform, which is calculated for a window length equal to thrice the average pitch period of the signal with 30% overlapping.
4. Spectrographic analysis for time-frequency plot of the speech signal.
5. Display of spectral moments of the pitch synchronously computed spectrum.
6. Estimation of lip-shape parameters using least squares approximation method and Delaunay triangulation method.

At first, the pitch period P_g of the speech signal is estimated based on the algorithm proposed by Rodman et al [6] [15] [24] [25]. Segmentation into these periods is then used for estimating lip-shape parameters. Spectral moments, namely, mass, weighted mean, and weighted variance of the average magnitude spectrum of the signal estimated with a window length equal to the pitch period P_g and actual

frequencies of the samples as the weights have been calculated. It will be observed in the next chapter that spectral moments of the vowels occupy distinct positions in the mean-variance space. Finally, these spectral moments are mapped to the lip-shape parameters. Lip-shape parameters are first analyzed for the training sounds. Then, the spectral moments of the training sounds are used as the reference data for synthesizing lip shapes for other speech segments like vowel-vowel and vowel-semivowel-vowel syllables.

3.1 Determining the pitch period

In implementation of the pitch period determination method proposed by Rodman et al [6] [15] [24] [25], we have assumed that a rough estimate of the pitch period P_g can be calculated using any other pitch determination method, and this will determine the pitch of the speech signal as that of a male or female or child. Using these results, we can define a range of valid P_gs. A method of calculating the glottal pulse position is reported by Rodman et al [6] [15] [24]. This method is based on the fact that the magnitudes of the odd harmonics of a periodic function with period T_p is zero when the function is expanded in Fourier series whose coefficients are determined by integrating over the interval $T = 2T_p$. If the periodic waveform is mixed with small amount of random noise, the sum of magnitudes of the odd harmonics is minimum, if the integration interval used for the calculation of the harmonics is equal to $2T_p$. In a narrow band time-dependent Fourier representation, the excitation for voiced speech is manifested in sharp peaks at integer multiples of the fundamental frequency and the contribution of noise is negligible when viewed as a function of frequency. In practice, the speech is digitized and we can take DFT samples as harmonics. Rodman et al have used the sum of the first four odd samples in the DFT. The algorithm is as follows:

Let $s_1, s_2, s_3, \ldots$ be some arbitrary sample points of a discrete time sequence $x(n)$ of voiced speech. A window sequence with window length in the range $[i, j]$ (an estimate of the pitch period $2P_g$) is defined such that

$$i \leq 2P_g \leq j$$

We then compute the magnitudes of the DFTs over the intervals $[s_1, s_i]$, $[s_1, s_{i+1}]$,

$[s_1, s_{i+2}]$, ... producing $|X_i(k)|$, $|X_{i+1}(k)|$, $|X_{i+2}(k)|$, ..., where

$$X_N(k) = \mathrm{DFT}[x(n)] \qquad 0 \leq k \leq N, \quad 0 \leq n \leq N \tag{3.1}$$

where $i \leq N \leq j$. Sum of the magnitudes of first four odd samples of the spectrum $X_N(k)$ is now obtained as

$$Y_N = |X_N(1)| + |X_N(3)| + |X_N(5)| + |X_N(7)| \tag{3.2}$$

Next, we find

$$Y_l = \max[Y_N, i \leq N \leq j]. \tag{3.3}$$

We check to see whether Y_l is below a certain threshold - determined by the maximum of the Y_N obtained when the segment is noise or silence and if so, P_g is set to zero for this segment. The introduction of Equation 3.3 is a slight modification in Rodman et al's approach. This is because a sequence of noise samples do not have a consistent pitch period sequence and that the identification of the noise segments among the voiced speech is a must for the correct working of the glottal pitch tracker. Otherwise, we seek the point m, where

$$Y_m = \min[Y_N, i \leq N \leq j]. \tag{3.4}$$

The value of m is the number of samples in the estimate of twice the pitch period P_g. The analysis window is then shifted forward by P_g samples and we search for the next minimum. This produces a sequence of P_g estimates, P_{g1}, P_{g2},.... To locate the P_{gi}, the estimated window size is

$$i = 1.5P_{g(i-1)}, \qquad j = 2.5P_{g(i-1)}$$

This window size is chosen because of the fact that the window size estimate should be large enough to handle the rapid excursions of the P_g when it encounters a fricative or similar class of sounds, but not so large as to allow the alternate extrema, which would result in the alternate tracking of a multiple of pitch period. Whenever a window segment is found to be unvoiced, P_g is set to zero, and the

analysis window is shifted by $(i+j)/4$, and the new values of i and j are respectively set to $2P_{g,min}$ and $2P_{g,max}$, which are the minimum and maximum values respectively in the predefined pitch period range.

$$\text{If} \quad \left|\frac{Y_m - Y_{m/2}}{Y_m}\right| < 0.01 \quad \text{then} \quad P_g = \frac{m}{4} \tag{3.5}$$

In a deviation with Rodman et al's approach, it is also checked whether any minimum of Y_N occurs at half the current estimate m and the relative difference in the two minima should be less than 0.01 and if so, the new estimate of the pitch period is set to $m/2$. Such extreme minima at window length equal to twice the pitch period and its integer multiples are observed to occur for vowels and semivowels. This is because, if Y_N is minimum for $N = 2P_g$, then it is minimum for $N = 2kP_g$ as well where k is an integer. On the other hand, the estimate should not be small enough such that the glottal pulse tracker converges to zero and crashes.

The fundamental frequency Fo_i of the speech segment with pitch period P_{gi} and the sampling rate 11.025 kHz is then calculated as

$$Fo_i = \frac{11025}{P_{gi}} \tag{3.6}$$

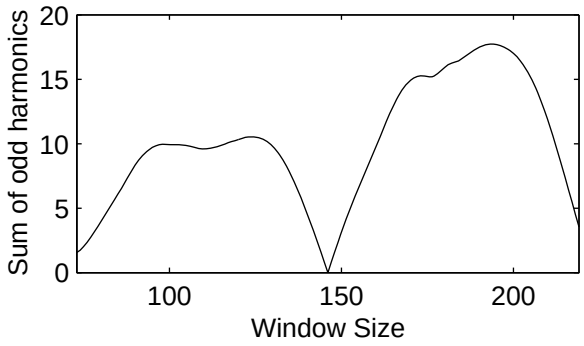

Figure 3.1: An example of sum of the magnitudes of first four odd samples of DFT for synthesized /a/

Figure 3.1 shows the plot of sum of the magnitudes of the odd samples of DFT of a window sequence for synthesized /a/. Since the window in this case is [73, 237], and the minimum is at 147th glottal pulse number, the pitch period P_g is $147/2 \approx 73$ samples.

3.2 Spectrographic analysis

The spectrogram is a two-dimensional representation of the time-dependent spectrum in which the vertical axis represents frequency and the horizontal axis represents time. The spectrum magnitude is represented by the darkness of the marking on the paper [26]. The analysis window determines the spectral bandwidth of the display. For a 3-dB bandwidth of 300 Hz, the spectrogram displays good temporal resolution and poor spectral resolution. For a 3-dB bandwidth of 50 Hz, the spectrogram displays good spectral resolution and poor temporal resolution. In this project, the analysis window used is a Gaussian window. The 3-dB bandwidth is calculated as [27]

$$\text{3-dB bandwidth} = \frac{1.2982804}{\text{window length}} f_s$$

where f_s is the sampling rate. Thus, for wide band and narrow band spectrograms, the window lengths of approximately 47 and 286 samples respectively have been used.

In a spectrogram, it is difficult to view the pitch harmonics and the formant structure clearly, as they are mixed up. However, it is possible to suppress the pitch harmonics by using harmonic or LPC smoothened spectrogram.

Harmonic smoothened spectrogram

A spectral envelope of the speech signal can be obtained with a window length equal to the pitch period P_g. If this spectral envelope is displayed as a spectrogram, the result is Harmonic Smoothened Spectrogram. This spectrogram gives a good display of the formant structure of the speech signal. Also, the formant analysis is dependent on the pitch period.

Linear predictive coefficient (LPC) model

The basis for linear predictive models is the source-filter model of speech production which roughly approximates the human vocal-tract filter to a time-invariant (for a short time window) with an all-pole transfer function. An important feature of the source-filter model is that it separates intonation from phonetic information. Intonation characteristics such as pitch, amplitude and voice quality are features of the sound source, while the vocal tract filtering affects the features of the phoneme being produced. LPC can be used to separate the sound source properties from those of the vocal tract filter, which governs the phoneme articulation and it can provide good parameter estimates for lip-syncing models generated by speech. The detailed discussion on LPC model can be obtained from [26].

The linear predictive coefficients of the speech signal with a fixed window length greater than or equal to $2P_g$ are computed for successive intervals. The magnitude response of an all-pole transfer function, with its coefficients equal to the linear predictive coefficients for each interval, are then computed. This magnitude response is then plotted as a spectrogram with the frequency along Y-axis and the corresponding time interval along X-axis and the response displayed as intensity.

3.3 Voiced-unvoiced segmentation

Once the P_g sequence of the speech signal is determined, it is necessary to know whether a segment of speech under analysis is voiced or not. If the segment is voiced, the fundamental frequency F_0 is an important parameter which needs to be correctly estimated. In presence of noise, it becomes significantly complicated to extract pitch and take the voicing decision. The purpose of a voicing decision is to correctly partition speech into voiced, unvoiced, and silence segments. In case of noise contaminated speech, only voiced-unvoiced segmentation can be made. As the noise is added, the silence segments may appear unvoiced and the weak voiced segments may even appear to be unvoiced [28]. So the voicing decision has to be robust in noisy conditions.

Energy can be an important parameter in finding the voiced segments of the speech signal. However, the energy parameter cannot be used to discriminate be-

tween unvoiced and silence segments. Wang et al [29] observed that the properties similar to that of cepstrum can help identify the voiced portions of the speech signal. The algorithm, as proposed by Wang et al, is implemented for a window length $2f_s/50$ (where f_s is the sampling rate) equivalent to the fundamental frequency 50 Hz in the audio range. The procedure is defined below for a sampling rate of 11.025 kSa/s and window size of 440 samples:

For each window of 440 samples, we compute $|X(k)|$, absolute value of the DFT of $x(n)$ and keep the first half as $R(k), 0 \leq k \leq 219$. Next we compute $T(l)$, the absolute value of the DFT of $R(k)$ and again keep the first half as $Y(l), 0 \leq l \leq 109$.

$$X(k) = \text{DFT}[x(n)] \qquad 0 \leq k \leq 439, \quad 0 \leq n \leq 439 \tag{3.7}$$

$$R(k) = |X(k)| \qquad 0 \leq k \leq 219 \tag{3.8}$$

$$T(l) = \text{DFT}[R(k)] \qquad 0 \leq k \leq 219, \quad 0 \leq l \leq 219 \tag{3.9}$$

$$Y(l) = |T(l)| \qquad 0 \leq l \leq 109 \tag{3.10}$$

Let $M = Y(0)$, i.e. the DC coefficient of $Y(l)$. It is observed that the value M for the voiced segment far exceeds the value M for the silence or weak noise segment. So, we define a threshold B, (30 in our case), to separate these segments. The introduction of threshold B is in deviation to Wang's approach. If the maximum value of $Y(l)$ falls below B, then the segment is termed as silence.

Let P be the maximum value of $Y(l)$ between sample 22 and sample 55, this range corresponds to the possible pitch period range of all speakers. The ratio of P to M is compared with a threshold t, which depends on the window size. If this ratio is greater than t (0.05 in our case), the utterance segment is voiced. It is to be noted that if at most four consecutive segments are termed as silence, then the segments are termed as unvoiced, provided the values M for these consecutive segments exceed the threshold B. Comparison with threshold B is needed, because

for silence segments, M has a low value, and its presence in the denominator of the ratio of P to M may give results comparable to those for voiced segments, thus wrongly terming some silence segments as voiced. The procedure is then repeated with a window shift of 110 samples.

3.4 Calculating spectral moments

Once the glottal pulse duration P_g sequence is determined, the spectral moments are calculated using the shape of the magnitude spectrum produced by using DFT on P_g-sample windows. The magnitude spectrum is normalized like a probability density function and statistical moments - mass, mean and variance are then computed. This technique, as proposed by McAllister et al [15] [25], is as follows:

At first the pitch period P_g of the speech signal is determined by the method given in section 3.1. Once the P_g sequence is being tracked accurately, the magnitude spectrum of the speech signal with the window length N equal to P_g is computed and averaged out, computation being done for the same window length N, with window shifting by at least one sample or 10% of the window length N.

$$X_n(k) = \mathrm{DFT}[x_n(m)] \qquad \begin{aligned} &0 \leq m \leq N-1, \\ &0 \leq k \leq N-1, \\ &0 \leq n \leq N-1 \end{aligned} \tag{3.11}$$

$$S(k) = \frac{1}{N\sqrt{N}} \sum_{n=0}^{N-1} |X_n(k)| \qquad 0 \leq k \leq N-1 \tag{3.12}$$

This does not produce any significant variation compared to the magnitude spectrum of the single speech segment, but it smooths out the spectrum. Next, DC term of $S(k)$ is dropped, and the cube-root of $S(k)$ is taken.

$$S'(k) = \sqrt[3]{S(k)} \tag{3.13}$$

$$M = 4000 * \frac{N}{f_s} \tag{3.14}$$

where f_s is the sampling rate.

Taking the cube-root of $S(k)$ has the effect of lessening the influence of the first formant and emphasizing the differences in the other parts of the spectrum. Next, only the first 4000 Hz part of the spectrum $S'(k)$ is selected using Equation 3.14 and normalized like a probability density function to get $P(k)$ by dividing by its mass m_0. The first central moment, i.e., mean (m_1), and the second central moment about the mean, i.e., variance (m_2), of the spectrum are computed and plotted on a m_1-m_2 space.

$$m_0 = \sum_{k=0}^{M} S'(k) \tag{3.15}$$

$$P(k) = \frac{S'(k)}{m_0} \tag{3.16}$$

$$m_1 = \sum_{k=0}^{M} kP(k) \tag{3.17}$$

$$m_2 = \sum_{k=0}^{M} (k - m_1)^2 P(k) \tag{3.18}$$

It is to be noted that while calculating the moments, the value of k in the computation is scaled by a factor f_s/N to compute the moments in the actual frequency range [0, $f_s/2$]. If the moments are calculated using their sample numbers as their weights, due to differences in the pitch periods even for adjacent segments, moments will then become dependent on the pitch period and hence the fundamental frequency. If the factor f_s/N is taken into account for the weights, then the maximum weight that is possible is $f_s/2$, and much of the above fluctuations diminish. However, some fluctuations do remain because of rectangular windowing as well as computational rounding, and so the spectral moments are smoothened using 5-point median filtering followed by 5-point averaging and then 3-point median filtering.

3.5 Lip-shape parameters from spectral moments

It is observed from the moment space, as shown in the next chapter, that for different speakers, the vowels occupy similar positions but at different points in the moment space, which is in agreement with observations as reported by Taylor

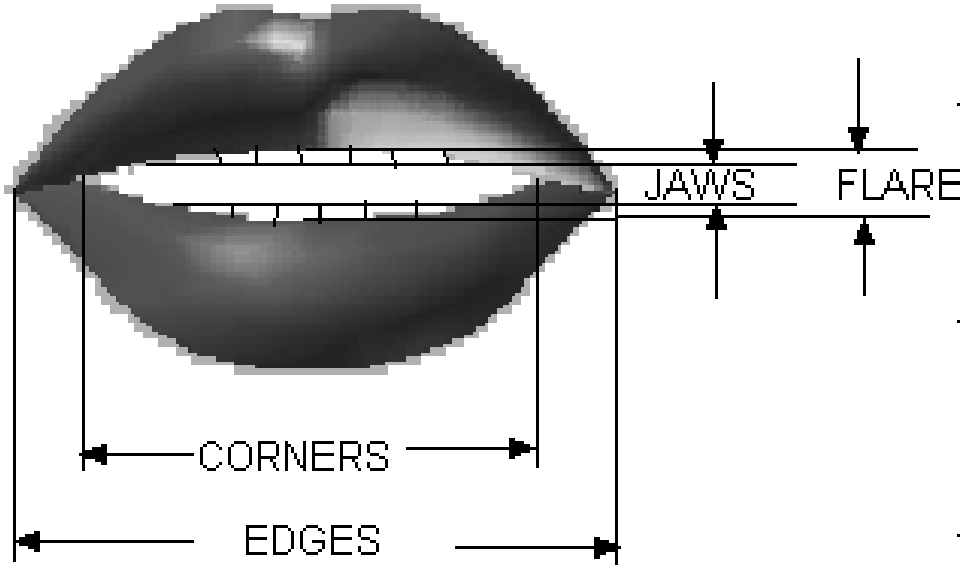

Figure 3.2: Parameters that define the lip shape

Table 3.1: Values of mouth parameters for different mouth positions [7] [30]

Speech Segment	Keyword	Mouth Parameters			
		Jaw	Flare	Edges	Corners
/*a*/	c<u>a</u>lm	1	0.15	0	0
/i/	t<u>ea</u>m	0	-0.75	1	0
/u/	z<u>oo</u>m	0.25	0.75	-1	1
/e/	n<u>a</u>me	0.5	-0.75	0.75	0
/o/	f<u>oa</u>m	0.5	0.4	-0.25	0.5
silence	-	0	0	0	0

et al [31] [32]. This implies that the spectral moments of the vowels can be mapped to the corresponding positions of the lip shape. For experimental investigation,

similar to those of Krothapalli et al [7] [30], four measurable parameters with extreme range like vertical and horizontal openings of the lips, lip-protrusion and actual closing of the lip in the horizontal position are used for developing the mouth shape, as shown in Figure 3.2 taken from an animated cartoon. Also, we have assumed that the parameter values for the silence are zero. This is because any arbitrary lip shape can be used for the silence, and the requirement is for an ideal lip shape for silence, which can be used as reference for calculating lip parameters for other mouth positions. The four parameters are [7] [30] :

1. The parameter **Jaw** is a function of jaw height during articulation, and is measured as the opening between the upper and the lower teeth. As the upper jaw is fixed in position with reference to silence, it is the lower jaw, that determines the parameter value. Jaw parameter takes values in the range [0, 1], because the lower jaw cannot take a higher position than that in silence for any possible mouth shape. The lower the jaw position while uttering the sound, the higher is the jaw value. For sounds like /a/, where the lower jaw moves down to its maximum position, the jaw parameter takes a higher value. For sounds like /i/ and /u/, the lower jaw is nearer to that in silence and so the parameter takes a low value.

2. The parameter **Flare** is a measure of forward or backward motion of the lips. For sounds like /oo/ and /o/, the lips round and protrude forward, and the flare values are high in these cases. For sounds like /i/ and /e/, the lips are stretched, and so have negative flare values.

3. **Edges** is the measure of the distance between the two points where the lips meet internally. For sounds like /u/ and /o/, this distance is less than that in silence and so the edges have negative values. For sounds like /i/ and /e/, where the mouth is wide open, the edges parameter takes a positive value.

4. **Corners** is the measure of the distance between the extreme lip joints in the horizontal direction. The difference between edges and corners parameters is clearly seen in sounds like /u/ and /o/ where the lips are rounded. These sounds take a high value for the corners. Sounds like /i/, /e/, and /a/ have

nearly the same position for corners as that in silence. The three parameters, namely, flare, edges and corners take values in the range [-1, 1].

Relative values of the lip-shape parameters for different speech segments are given in Table 3.1. The relative values are used to represent the mouth positions as it would then become easy to map the values to the absolute values of the actual mouth shapes depending on the language medium used and the size of the mouth. Moreover, the sum of the absolute values of the jaw and the flare gives the actual height of the top of the lower or upper lip from the jaw position while in silence [7].

At present we are concerned with developing mouth shapes for vowels and semivowels. So, speech segments consisting of vowels /*a*/, /e/, /i/, /o/, and /u/ have been used as the training sounds for research. In a deviation from Krothapalli et al's approach, the sound /œ/ has been excluded from the investigation. These vowel sounds are well separated in vowel space and they represent different possible mouth shapes in English and Hindi languages. Because of their longer durations, the middle part of the utterance is more or less static and it therefore becomes easy to define a mouth shape for each of these vowels. Such isolated sounds are chosen for training because obtaining mouth shapes directly from words are more difficult. Because of the co-articulation effect, the mouth positions for semivowels which have shorter durations than vowels, follow that of surrounding vowels at the time of transition. So, there is not a single semivowel as part of training sound. Moreover, the above four parameters are sufficient enough to define the mouth positions for most voiced utterances.

Once we have calculated the spectral moments of the speech signal and have defined the training sounds and their relative mouth positions, the task is now to find an appropriate method for finding the relative mouth position from the spectral moments. When speech is recorded from a speaker, the background noise is also recorded along with it. There might also be some experimental error in calculating the spectral moments of the speech signal. The smoothing techniques used in computing spectral moments remove pitch differences and noise with small variance. However some perturbations in the data still remains. So it is necessary to choose methods that do not suffer from small perturbations in the data. Three

methods, viz. least squares approximation, 2D Delaunay triangulation and 3D Delaunay triangulation methods have been investigated for finding the relative mouth positions.

3.6 Least squares approximation

Various approaches to interpolation have been proposed for finding the relationship between a set of data and the resultant outcomes [33] [34] [35]. The problem with the interpolation method is that it tends to give unexpected results for some inputs. There is another approach, least squares approximation - one that involves fitting a curve (surface) to a set of data without restricting that curve (surface) to coincide with the data points [35]. The least squares approximation method is used in case where the number of equations is greater than the number of unknowns. In many applications, there will be noise or experimental error along with the field data. Even though the field data remains roughly constant, the resultant outcomes vary. Therefore we need to find the control parameters which give minimal deviation of the desired outcomes from all data points. This can be achieved through the method of least squares. The method of least squares assumes that the best fit curve of a given type is the curve that has the minimum sum of least square error from a given set of data.

Given the data set $(x_1, y_1, h_1), (x_2, y_2, h_2), \ldots, (x_m, y_m, h_m)$, a polynomial $p(x, y)$ of degree n which gives a good data fit has to be found. Usually n is 1, 2 or 3 while m is very large. The polynomial $p(x, y)$ of third degree is given as

$$p(x,y) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 y + a_5 y^2 + a_6 y^3 + a_7 xy + a_8 x^2 y + a_9 xy^2 \quad (3.19)$$

The residual vector r at (x_i, y_i) is given by

$$r_i = p(x_i, y_i) - h_i \quad (3.20)$$

The sum R of the residual squares is given by

$$R = \sum_{i=1}^{m} r_i^2 \quad (3.21)$$

The coefficients are chosen to minimize R. The term a_0 is set to zero, since for silence, mean and variance are zero and so $p(x_i = 0, y_i = 0)$ should be zero. Experiments with various terms have shown that R has a much lower value and lower fluctuations in $p(x, y)$ are observed, when the terms a_7, a_8, a_9 are assumed to be zero, than when all the terms are present, due to the mutual independence of the mean and variance.

In matrix form, the equation 3.19, with the terms a_0, a_7, a_8, a_9 assumed to be zero, is written as

$$\mathbf{P} = \mathbf{X}\mathbf{A} \tag{3.22}$$

where

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_6 \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} p(x_1, y_1) \\ p(x_2, y_2) \\ \vdots \\ p(x_m, y_m) \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1 & x_1^2 & x_1^3 & y_1 & y_1^2 & y_1^3 \\ x_2 & x_2^2 & x_2^3 & y_2 & y_2^2 & y_2^3 \\ \vdots & & & & & \\ x_m & x_m^2 & x_m^3 & y_m & y_m^2 & y_m^3 \end{bmatrix}$$

The coefficient vector A is then calculated as

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{P} \tag{3.23}$$

where the matrix $(\mathrm{X}'\mathrm{X})^{-1}$ is known as pseudo inverse of X.

Here, the mean m_1 and variance m_2 are chosen as the two independent variables x and y respectively. The desired outcome h is one of the four mouth parameters, thus requiring four polynomials, each corresponding to one of the four parameters to be solved. We now need to define the speech segments which can best represent the training sounds for investigation. The choice is necessary because such training sounds should give the best possible parameter values for developing the mouth shape. The training sounds can be the middle parts of the speech segments /a/, /e/, /i/, /o/, /u/, which are assumed to be static. Also those sections of the speech segments should be avoided which have wide fluctuations in the contours of the pitch and spectral moments. Using these criteria, a cluster of 15-20 points, each for mean and variance from the selected segments are taken and

concatenated into two lists. These lists represent the data for the mean and the variance of sounds under analysis. We assume that for these values of the mean and variance of the selected segments, the corresponding values of the lip-shape parameters, as given in Table 3.1, are known.

We observed from the plots of spectral moments that the spectral moments of noise segments are not insignificant compared to that of voiced segments and also the moments of the noise segments do not have specific values. However, the mass of the spectrum of the noise or silence segment is comparably lower than that of the voiced segments. So, we set to zero the mean and variance of the spectrum of the noise segments using their mass and the voiced-unvoiced segmentation as the criteria. This is because for silence segment, with zero values for mean and variance, the parameters would be zero by Equation 3.19.

The vector A is calculated for each of the four mouth parameters, giving four vectors $A_{jaw}, A_{corners}, A_{flare}$, and A_{edges} respectively. These vectors are then multiplied by the spectral moments of the speech segments to get the values of the corresponding mouth parameters for the same speaker.

The problem with this method, as observed from the results in the next chapter, is that this method produces large deviation for some of the data points. Moreover, experiments have shown that increasing the degree of the polynomial results in poor curve-fitting. Even smoothing the resultant outcomes does not give good curve fitting. The reason is the localized data variation which the least squares method fails to ignore.

3.7 Surface fitting using 2D Delaunay triangulation

Least squares approximation is a global method, and this is why there are local fluctuations in the outcomes extracted using this method. In accordance with Krothapalli et al's findings [7] [30], we wish to investigate the 2D Delaunay triangulation method which can ignore small fluctuations in the localized region. Matlab provides a package for Delaunay triangulation, and data gridding and surface fitting. The only limitation is that the method uses interpolation for finding

intermediate points and does not employ extrapolation and so any seeking of data outside a given region will return NaNs.

Similar to the least squares approximation method, the mean m_1 and variance m_2 are chosen as the two independent variables. Moreover, the criteria used for selecting the speech segments for training sounds are same as that used for least squares approximation method. The m_1 and m_2 of the sequence are set to zero for the noise segments. We assume that for certain values of the mean and variance of the training sounds, the corresponding values of the mouth parameters, as given in Table 3.1, are known. Using the mean and variance and the values of the mouth parameters of the training sounds as the reference points, and the Matlab function 'griddata()' for 2D Delaunay triangulation, the values of the mouth parameters can be calculated for the utterances of the same speaker.

3.8 Volumetric fitting using 3D Delaunay triangulation

Some fluctuations in the results are still observed while using 2D Delaunay triangulation. This is because, the regions of some vowels for the same speaker overlap in the moment space, and the energy and the mass m_0 of the speech are not even roughly flat for most of the durations. This happens mostly during phonetic transitions and silence-speech transitions. So, we try to investigate whether the use of three independent variables m_0, $m1$ and m_2 will give a better outcome.

The mass m_0, mean m_1 and variance m_2 are taken as the three independent variables. The criteria for selecting the segments for training sounds have been described earlier. Unlike the previous procedure, the m_1 and m_2 of the noise segments are not set to zero. Moreover, we assume that for certain values of the spectral moments of the training sounds, the corresponding values of the mouth parameters, as given in Table 3.1, are known. Using these spectral moments and the respective values of the mouth parameters as the reference points, and the Matlab function 'griddata3()' for 3D Delaunay triangulation, the values of the mouth parameters are calculated for the utterances of the same speaker.

Chapter 4
Analysis and Results

The techniques presented in the previous chapter were evaluated using synthesized vowels and the speech utterances of five adult male speakers. The pitch range for the male speakers has been taken as 65-260 Hz. The range of the pitch period P_g for the male speakers will then be approximately 42-170 samples for a sampling rate of 11.025 kSa/s using Equation 3.6. For the speech utterances of the female and the child, a small modification regarding the pitch range has to be made.

Speech segments /a/, /e/, /i/, /o/, /u/ extending 1-2 secs in durations were recorded from different speakers as part of research for training sounds. The vowel-vowel and vowel-semivowel-vowel segments were also recorded from the same speakers. Firstly, the pitch period sequence and the lip-shape parameters are extracted for the training sounds recorded by a speaker. Next, the spectral moments of the training sounds and their corresponding lip-shape parameters are saved as reference data. Then, using these data, the lip-shape parameters are extracted for various syllables for the same speaker. This method is then repeated for different speakers.

4.1 Determination of pitch period and voiced-unvoiced segmentation

Figure 4.1 shows the waveform and average magnitude spectrum for synthesized vowels /a/, /i/, and /u/ for two fundamental frequencies. This figure also shows a plot of the sum of the magnitudes of first four odd samples of the DFT, as a function of window length. The pitch period estimate is equivalent to half the window length for which the sum of the magnitudes of first four odd samples of DFT, i.e. Y_m of Equation 3.4, is minimum. Last column in the figure shows the

corresponding magnitude spectrum evaluated over one pitch period.

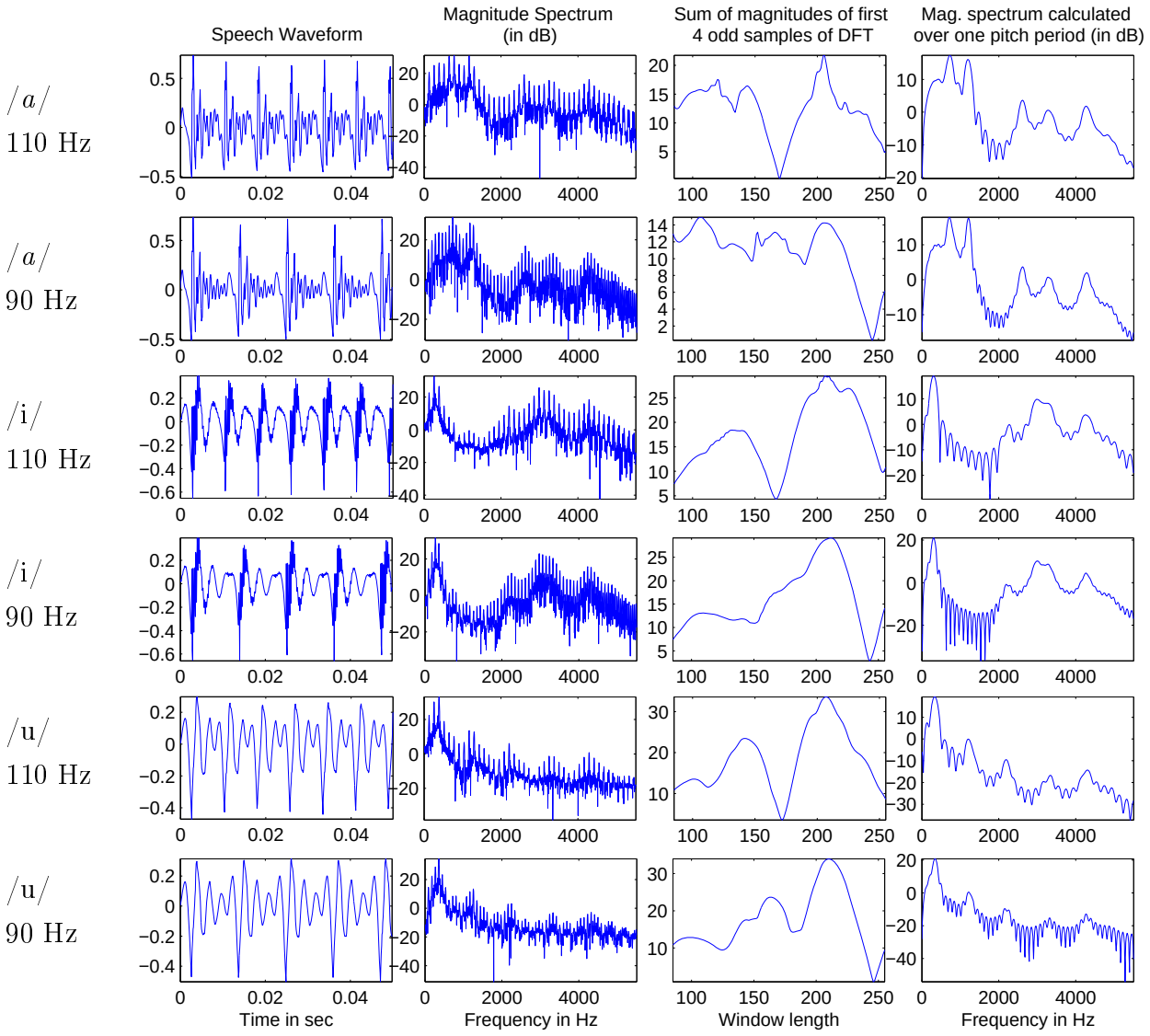

Figure 4.1: Spectral analysis of synthesized vowels /a/, /i/, /u/ with Fo = 90 Hz and 110 Hz

We now analyze the pitch period determination method and the voiced-unvoiced segmentation method with synthesized vowel syllables with varying pitch or amplitude characteristics. All synthesized segments in Fig 4.2 have a constant amplitude of voicing $AV = 60dB$ and a pitch contour in steps of 100, 100, 120, 120, 90, 90 Hz as fed to Klatt synthesizer tool [9]. The segments in Fig. 4.3 have a constant

pitch of 130 Hz and amplitude contour in steps of 50, 50, 65, 65, 55, 55 dB.

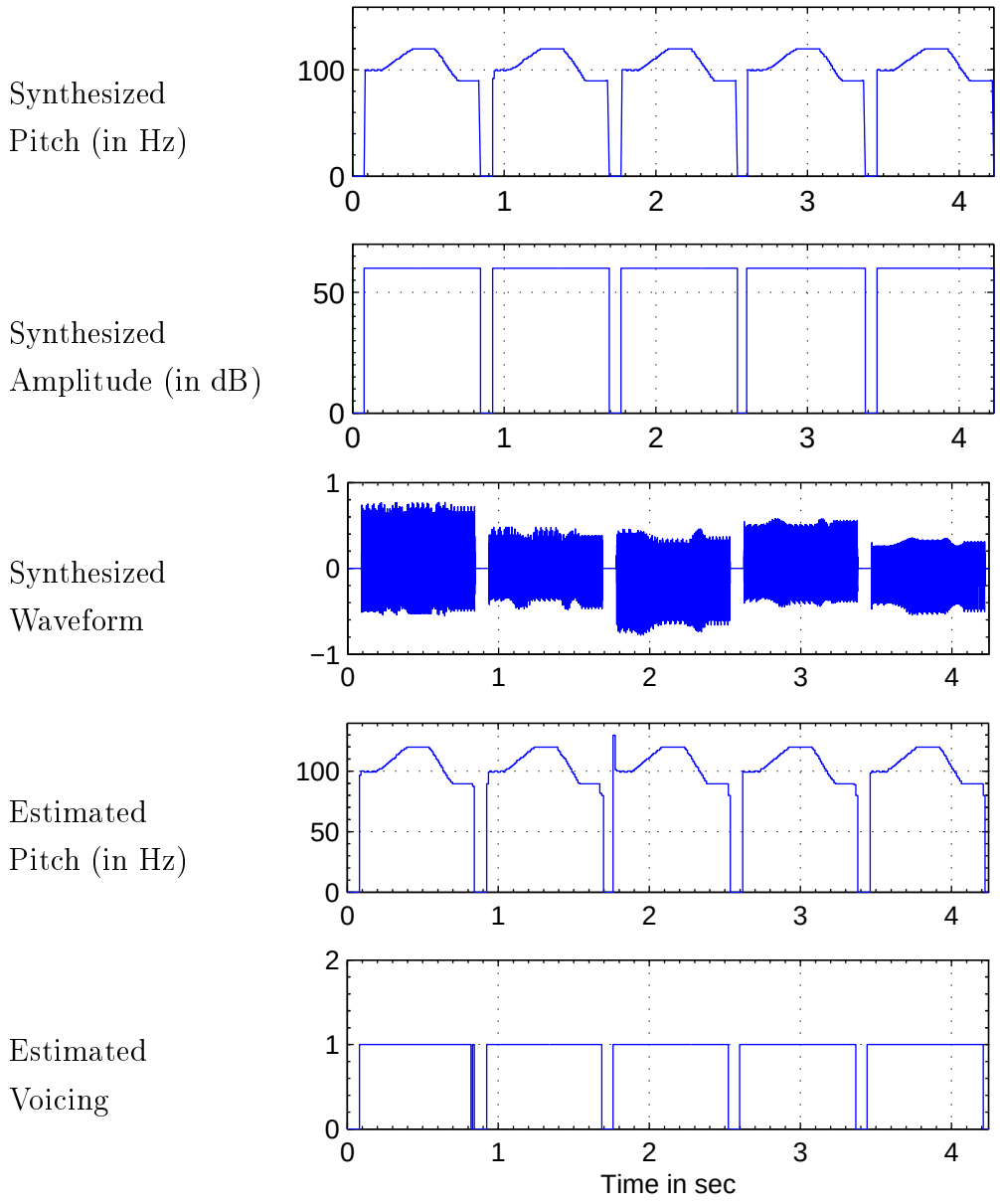

Figure 4.2: Pitch period and voiced-unvoiced segmentation of synthesized syllables /a/, /e/, /i/, /o/, /u/ with varying Fo and constant amplitude

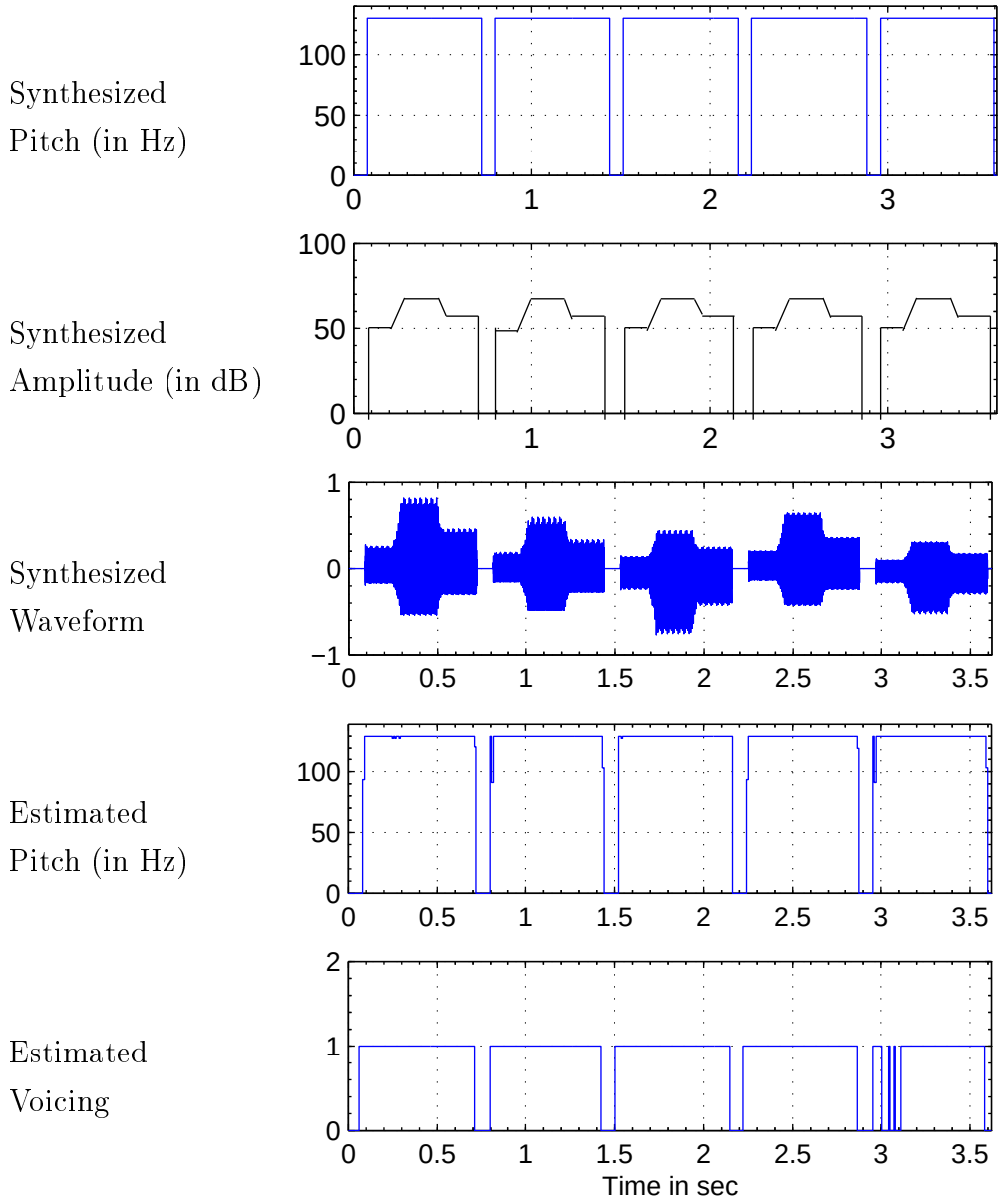

Figure 4.3: Pitch period and voiced-unvoiced segmentation of synthesized syllables /*a*/, /e/, /i/, /o/, /u/ with constant Fo and varying amplitude

We observed that the methods work correctly for all segments as shown in Figures 4.2 and 4.3. However, during voiced-silence transition, the pitch period determination method shows an error. This is because, at the transition, the segment contains a sharp transition between voiced region and silence. For natural segments as in Fig 4.4 and 4.5, the change in the pitch period is very well detected, and unlike fluctuations in pitch period during speech-silence transitions in the synthesized segments, no such case occurs in case of natural segments.

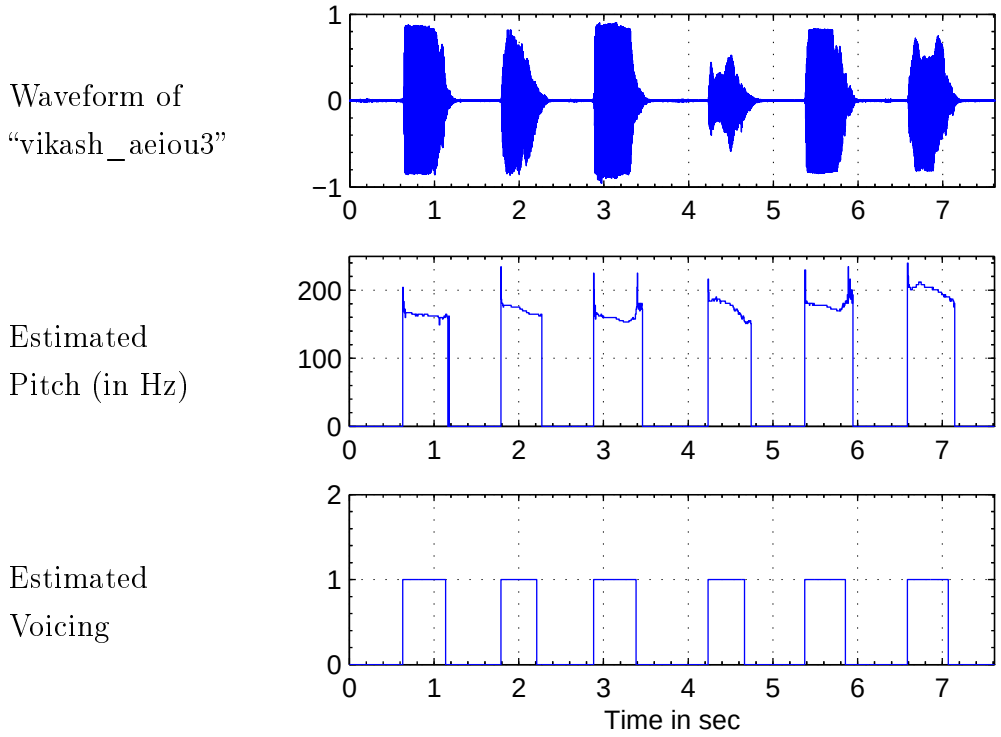

Figure 4.4: Pitch period and voiced-unvoiced segmentation of natural syllables /*a*/, /e/, /œ/, /i/, /o/, /u/ by a male speaker "Vikash"

Similarly, for segments like /*aa*Iye/, /*aya*/ and /*awa*/ in Figures 4.6 and 4.7, the pitch period tracker is able to follow the changes in the pitch period. However, the voiced-unvoiced segmentation method shows some unvoiced segments for syllable /*aa*Iye/ which is actually voiced speech.

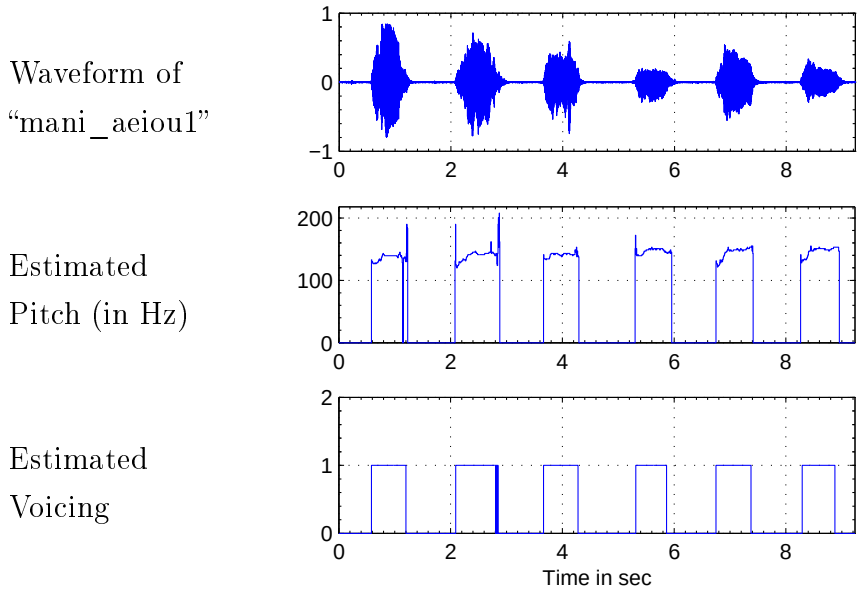

Figure 4.5: Pitch period and voiced-unvoiced segmentation of natural syllables /*a*/, /e/, /œ/, /i/, /o/, /u/ by a male speaker "Mani"

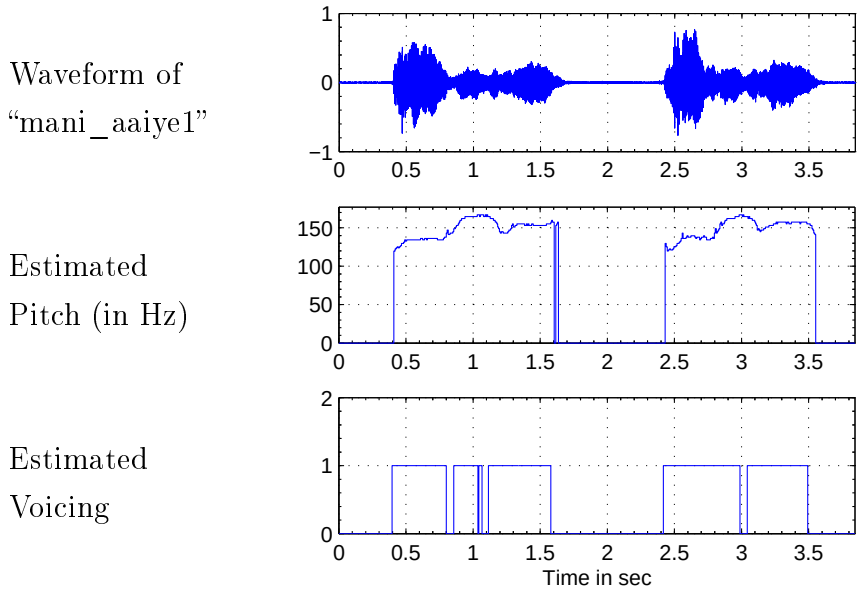

Figure 4.6: Pitch period and voiced-unvoiced segmentation of natural syllables /*aa*Iye/ (uttered twice) by a male speaker "Mani"

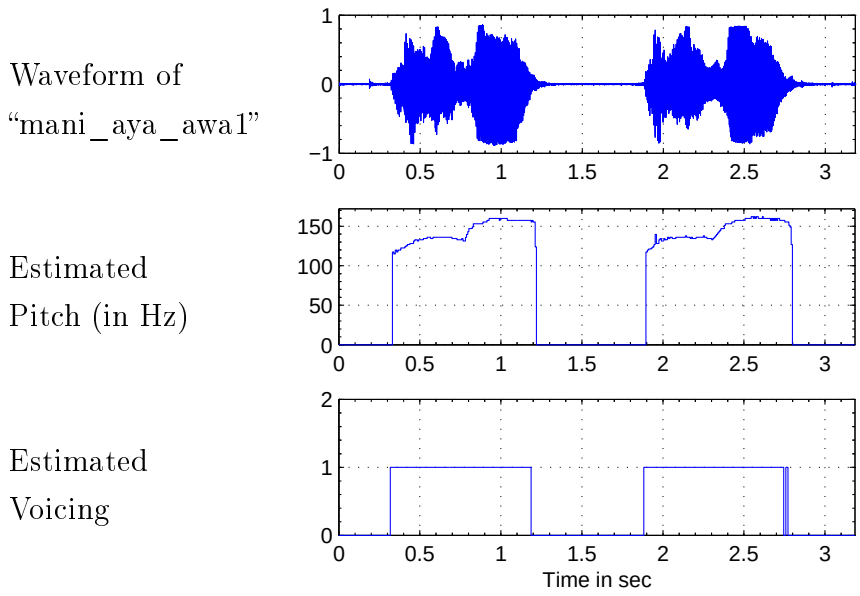

Figure 4.7: Pitch period and voiced-unvoiced segmentation of natural syllables /*aya*/-/*awa*/ by a male speaker "Mani"

We now test the pitch period tracker and the voiced-unvoiced segmentation method with fricatives and stops. Results in Figure 4.8 show that unvoiced segments in the syllable /*apa*/ are well detected. For /*aba*/, it shows unvoiced during closure, and voiced soon after the burst release, indicating an early voiced onset for the voiced stop.

Figure 4.9 shows the results for syllables /*afa*/ and /*ava*/, As seen from the figure, the unvoiced segments in the former syllable is detected only for a short duration, because of the low threshold value of t or else the high M value. Even the pitch period tracker gives zero values around the unvoiced segments. The latter syllable is found totally as voiced which is true. However, there is a dip in the pitch period contour for the latter syllable when the syllable makes a transition from vowel to fricative.

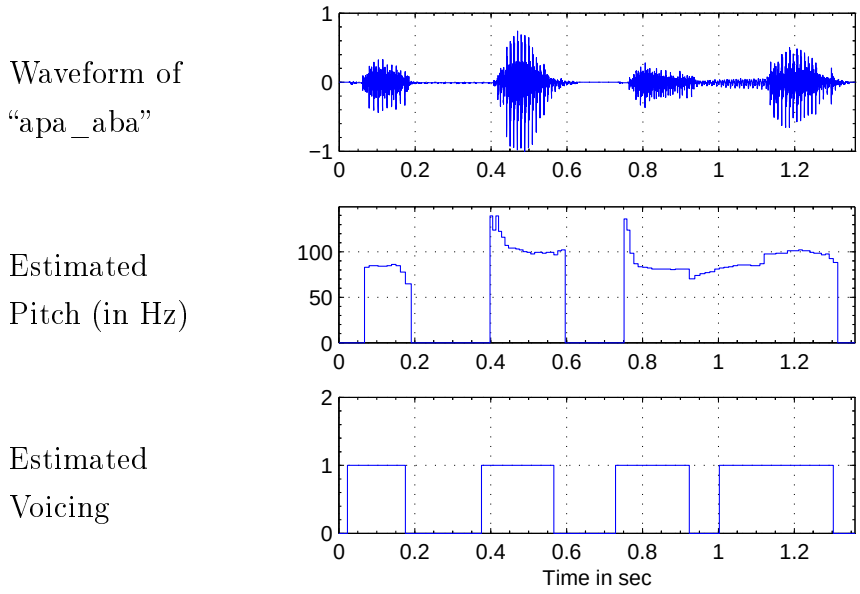

Figure 4.8: Pitch period and voiced-unvoiced segmentation of natural syllables /*apa*/ and /*aba*/ by a male speaker "Peter"

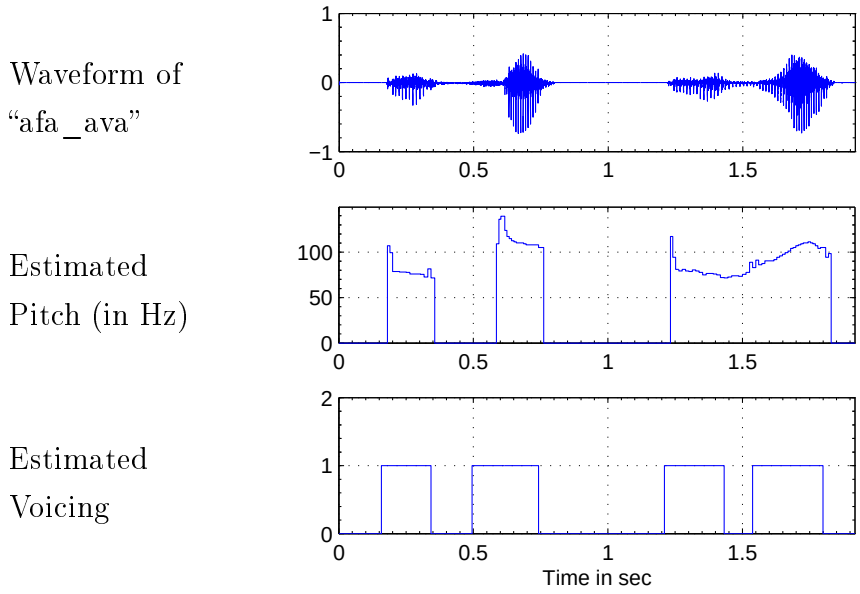

Figure 4.9: Pitch period and voiced-unvoiced segmentation of natural syllables /*afa*/ and /*ava*/ by a male speaker "Peter"

4.2 Analyzing spectral moments

After verifying the pitch period determination method, we now need to analyze the spectral moments of the syllables. As observed from Figures 4.10 and 4.11 for synthesized syllables, fluctuations are observed between silence-speech transitions. Also the mean and standard deviation of the speech are roughly flat for most of the duration in case of speech segments. From the mean vs variance plot, we observed that the vowels occupy distinct, but similar positions independent of amplitude and pitch. Similar results are obtained for natural syllables as shown in Fig. 4.12 and 4.13. However, even though the noise is negligible in amplitude compared to speech segment, their mean and standard deviation are not insignificant, because of the low value of the corresponding mass m_0.

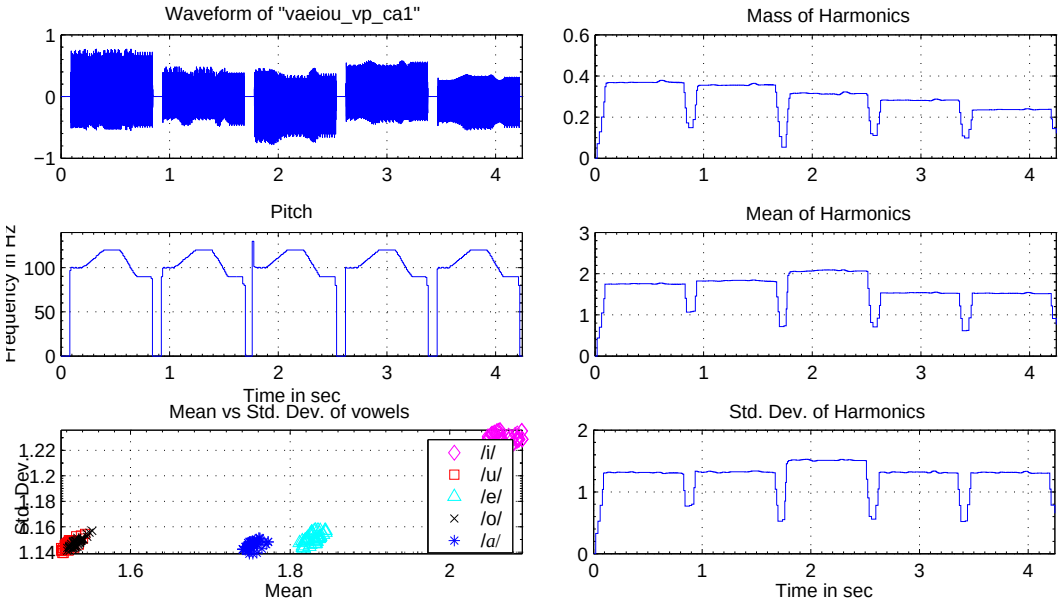

Figure 4.10: Spectral moments of synthesized syllables /a/, /e/, /i/, /o/, /u/ with varying Fo and constant amplitude

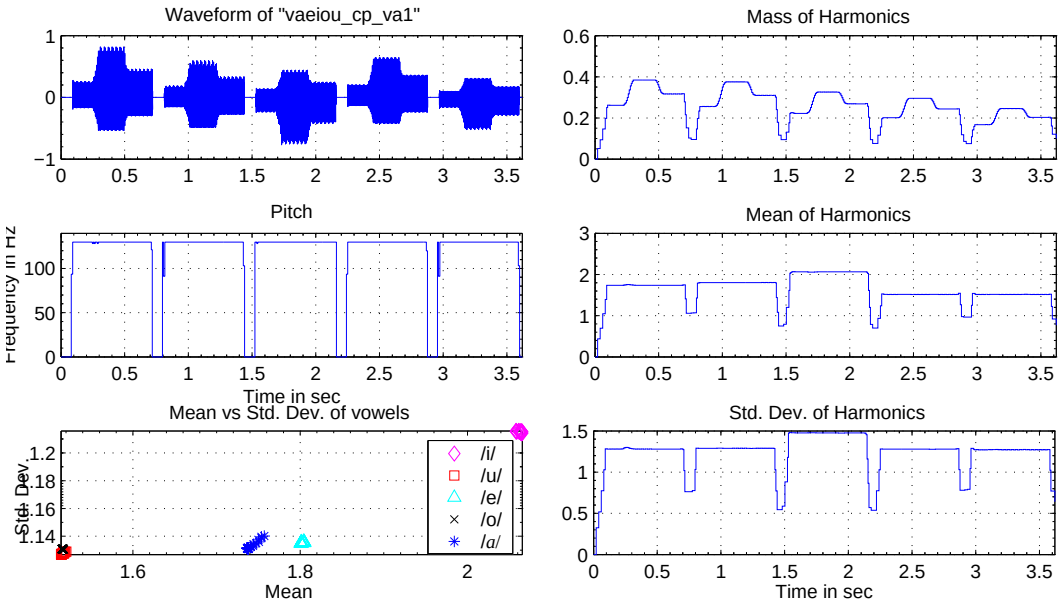

Figure 4.11: Spectral moments of synthesized syllables /a/, /e/, /i/, /o/, /u/ with constant Fo and varying amplitude

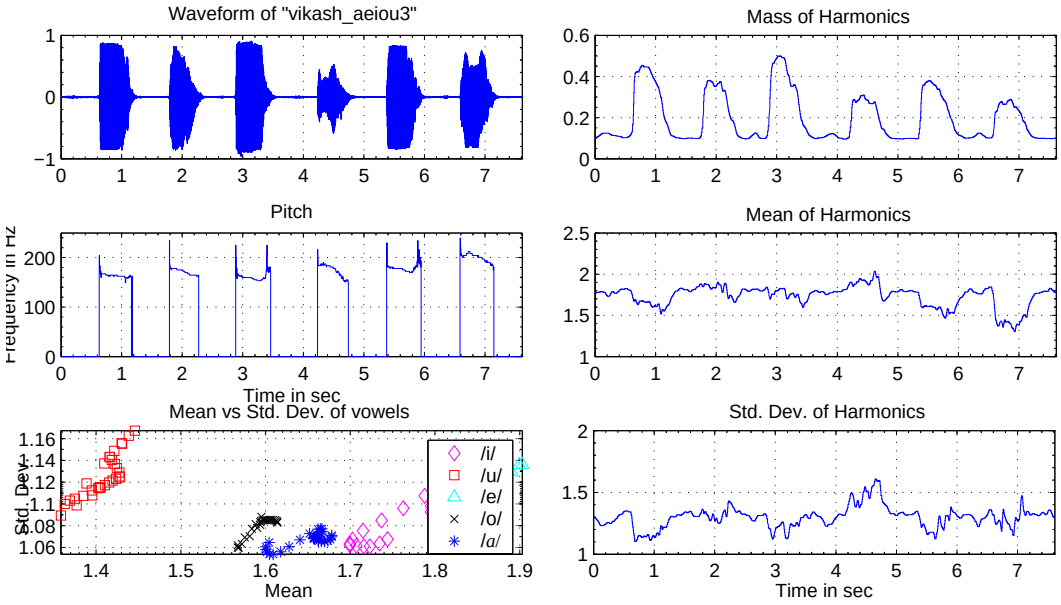

Figure 4.12: Spectral moments of natural syllables /a/, /e/, /œ/, /i/, /o/, /u/ by a male speaker "Vikash".

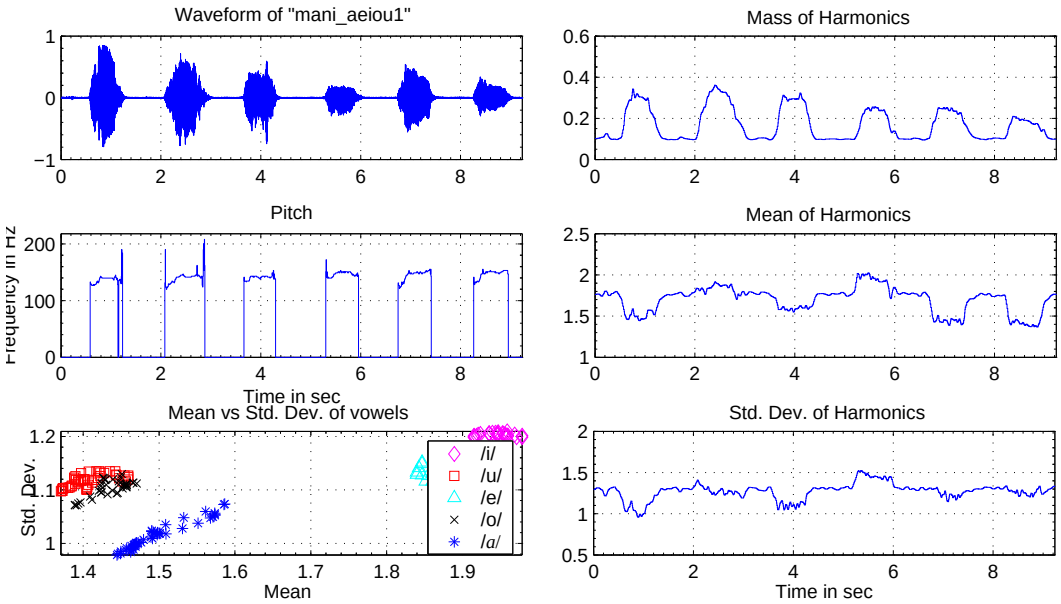

Figure 4.13: Spectral moments of natural syllables /a/, /e/, /œ/, /i/, /o/, /u/ by a male speaker "Mani".

4.3 Extracting lip-shape parameters

As observed in the previous section, the vowels, whether synthesized or natural, occupy similar but different positions in the moment space. Hence, the spectral moments of the speech can be mapped to the relative mouth parameters independent of the speaker. In the previous chapter, we have described about the least squares approximation, grid surface fitting using 2D Delaunay triangulation and grid volumetric fitting using 3D Delaunay triangulation. In the current section, we will show that the lip-shape parameters calculated from these methods are directly related to the positions or tracks of the syllables in the moment space.

As described earlier in the previous chapter, the vowels /a/, /e/, /i/, /o/, /u/ are used as the training sounds. The mean m_1 and variance m_2 are used as the independent variables for the least squares approximation method and 2D Delaunay triangulation method. Apart from mean and variance, the mass m_0 is used as another independent variable for 3D Delaunay triangulation method. From

figures in the previous section for spectral moments, we observe that the contours for mean and variance do not have wide fluctuations in the middle segments for all vowels. Since we need the best possible results with training sounds, we choose these middle segments of the vowels as the best choice for training sounds, and their respective mean and variance as the mean and variance for the training sounds. So, a cluster of 15-20 points each for mass, mean and variance is selected from all vowels as part of research for training sounds.

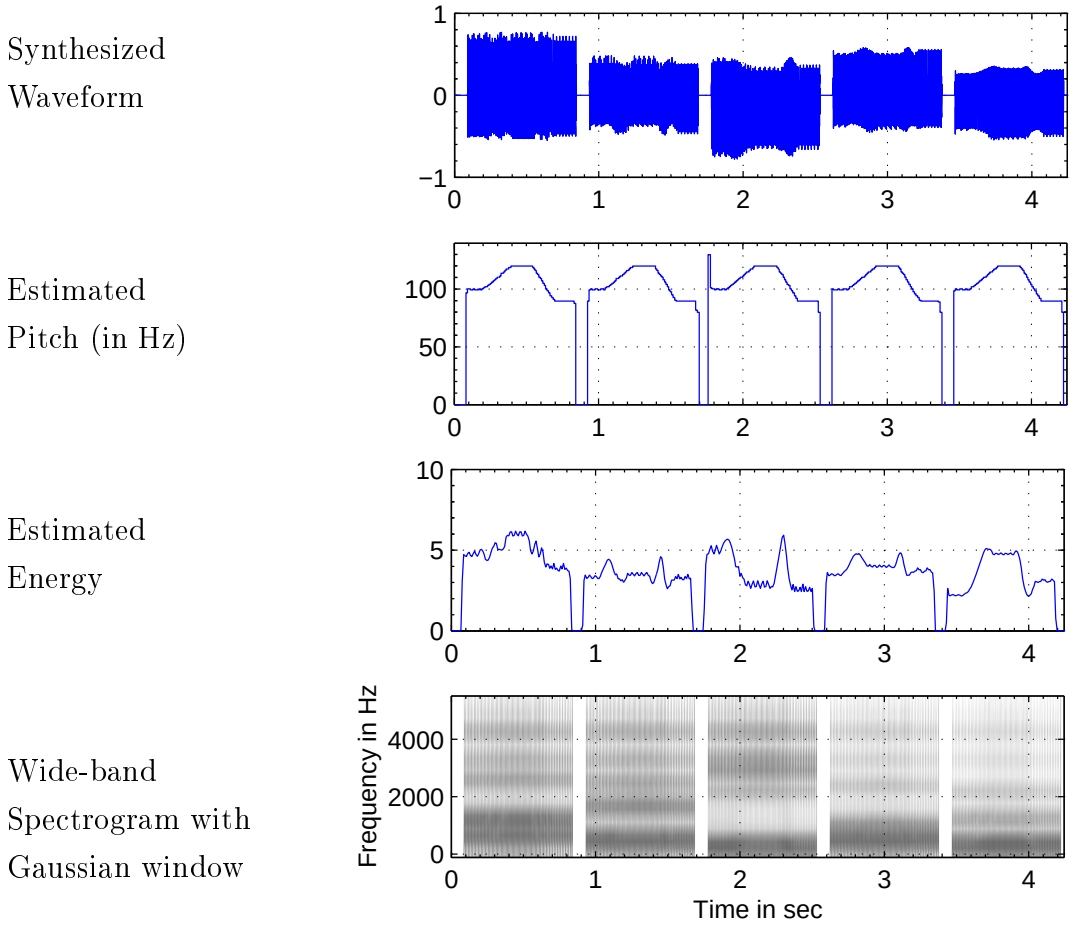

Figure 4.14: Characteristics of synthesized syllables /a/, /e/, /i/, /o/, /u/ with varying Fo and constant amplitude

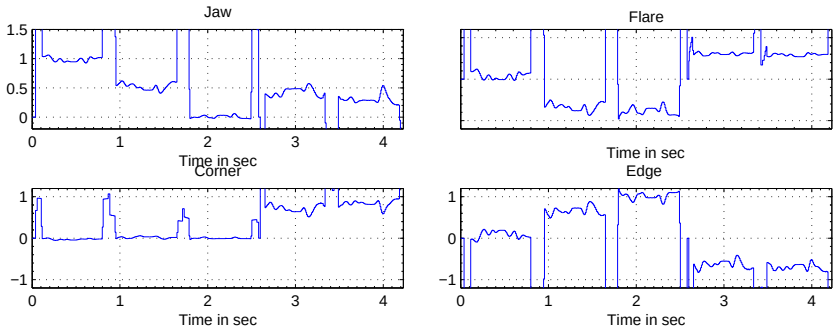

Figure 4.15: Lip-shape parameters of synthesized syllables in Fig. 4.14 calculated using least squares approximation.

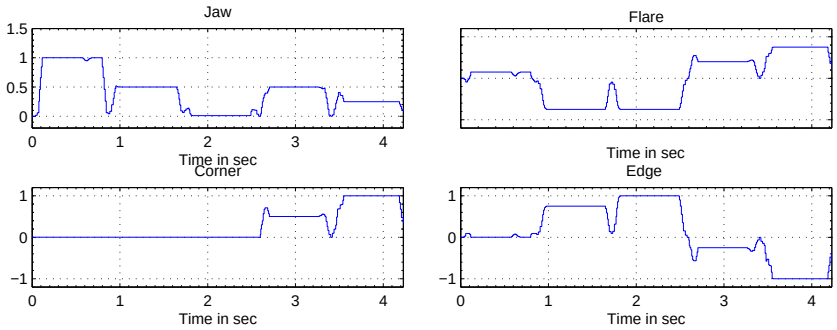

Figure 4.16: Lip-shape parameters of synthesized syllables in Fig. 4.14 calculated using 2D Delaunay triangulation.

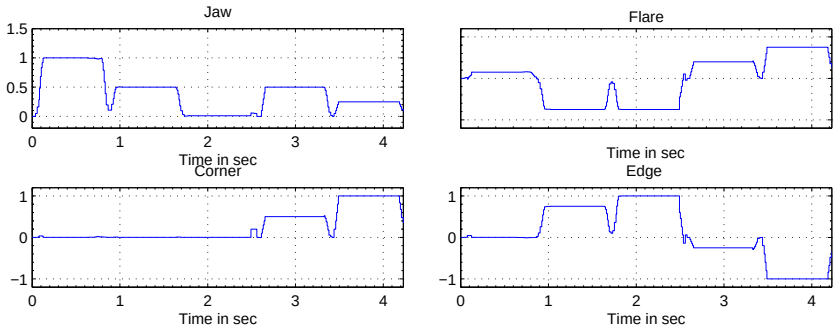

Figure 4.17: Lip-shape parameters of synthesized syllables in Fig. 4.14 calculated using 3D Delaunay triangulation.

From Figures 4.15, 4.16, 4.17 and 4.19, 4.20, 4.21, it is observed that irrespective of the pitch or amplitude variation, the lip-shape parameters are roughly flat for most of the durations. There are some fluctuations observed in case of lip-shape parameters calculated using least squares approximation as shown in Fig 4.15 and 4.19. For both types of synthesized syllables as in Fig 4.14 and 4.18, the mouth parameters comply to the values given in Table 3.1. As for example, the jaw parameter takes a very high value for vowel /a/ and a low value for vowel /i/. Similarly, the mouth shape is rounded and lips move forward in case of vowels /o/ and /u/ as clearly confirmed by the parameters flare and edges.

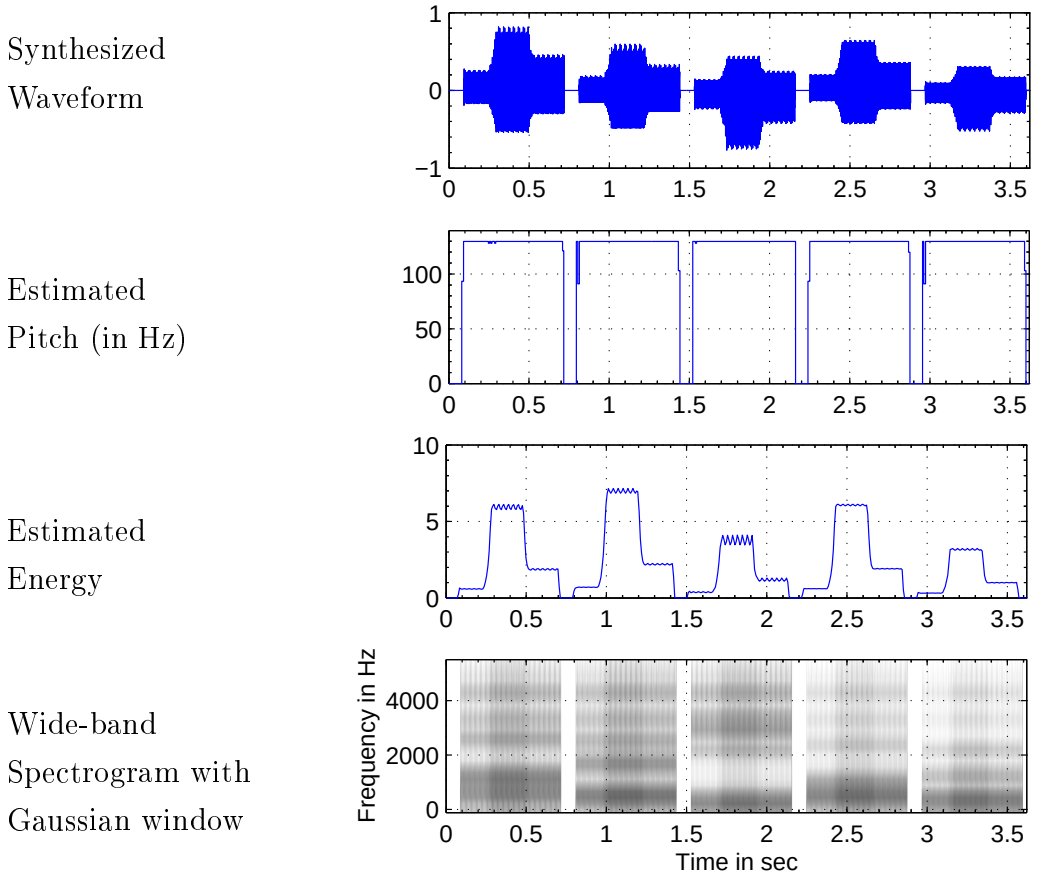

Figure 4.18: Characteristics of synthesized syllables /a/, /e/, /i/, /o/, /u/ with constant Fo and varying amplitude

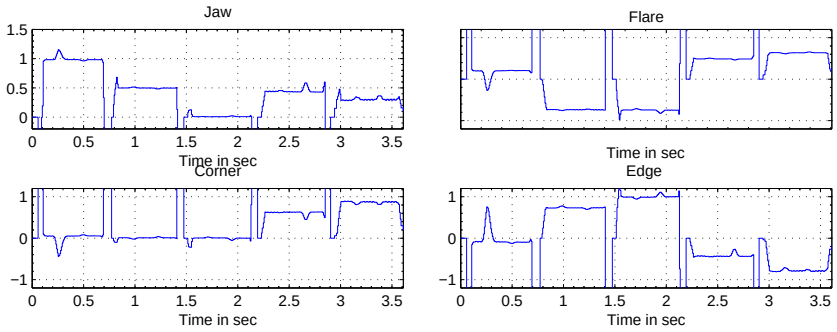

Figure 4.19: Lip-shape parameters of syllables in Fig. 4.18 calculated using least squares approximation.

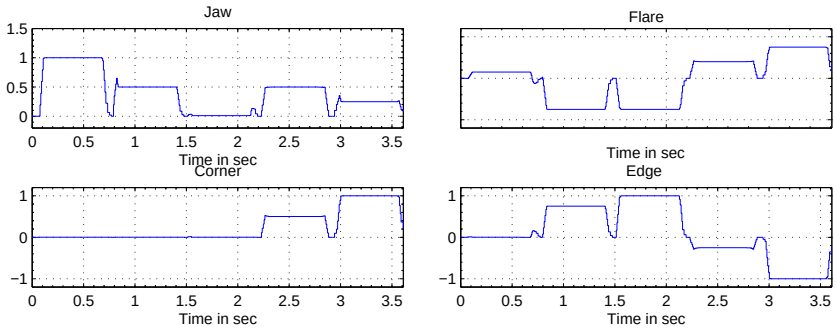

Figure 4.20: Lip-shape parameters of syllables in Fig. 4.18 calculated using 2D Delaunay triangulation.

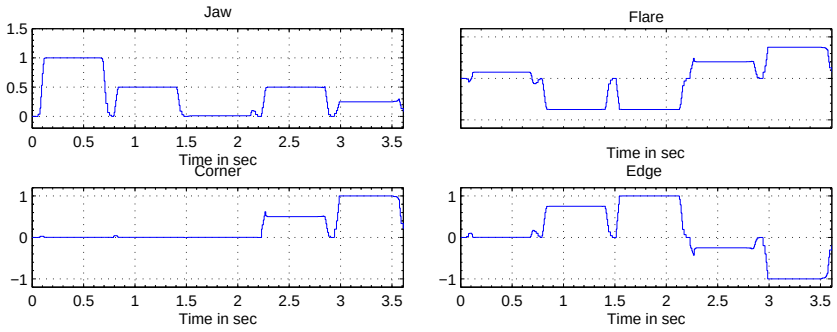

Figure 4.21: Lip-shape parameters of syllables in Fig. 4.18 calculated using 3D Delaunay triangulation.

4.4 Lip shape analysis for speaker "Vikash"

Next, it is to be observed that the mouth parameters for natural vowels also comply to the values given in Table 3.1. As expected they indeed do as observed in Figures 4.23, 4.24 and 4.25 for speaker "Vikash". But for vowel /œ/, the mouth parameters are not flat for most of the durations. A glance at the results for synthesized and natural vowels show that 3D Delaunay triangulation method gives smoother results than the other two. Least squares approximation method does come close to it in results with fluctuations at some points.

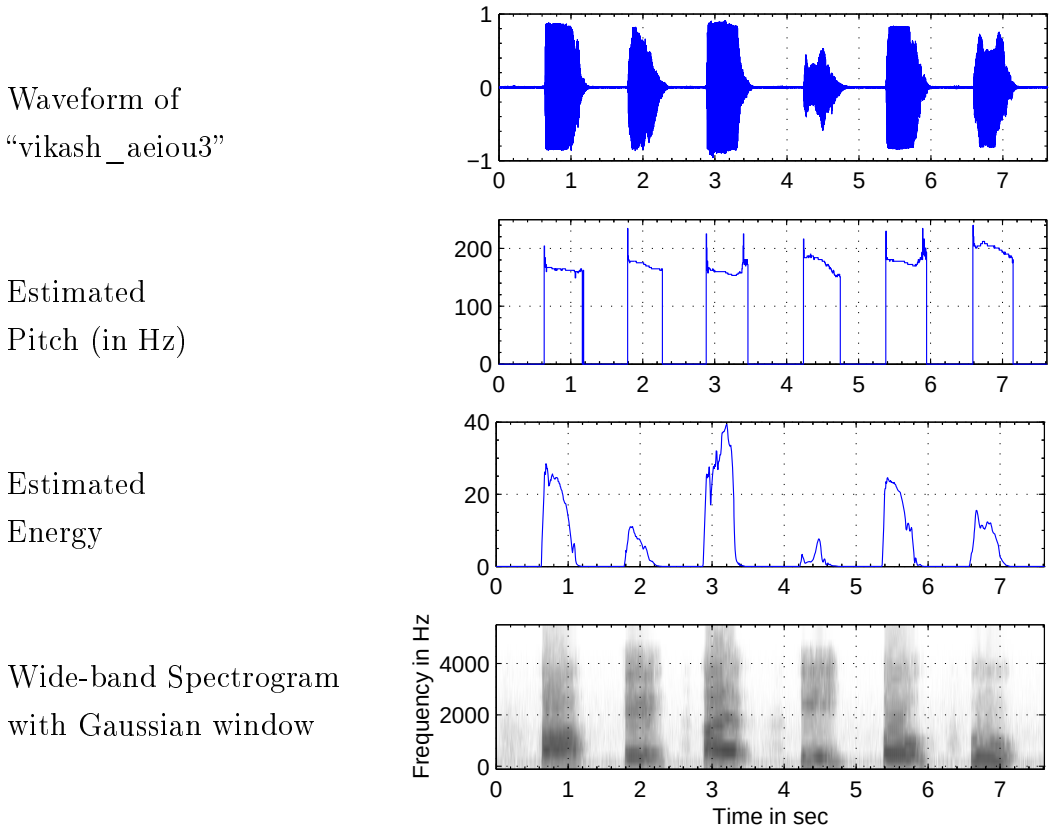

Figure 4.22: Characteristics of natural syllables /a/, /e/, /œ/, /i/, /o/, /u/ by speaker "Vikash"

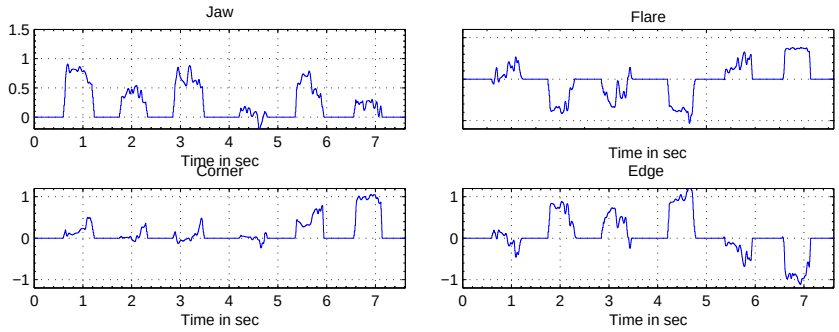

Figure 4.23: Lip-shape parameters of syllables in Fig. 4.22 calculated using least squares approximation.

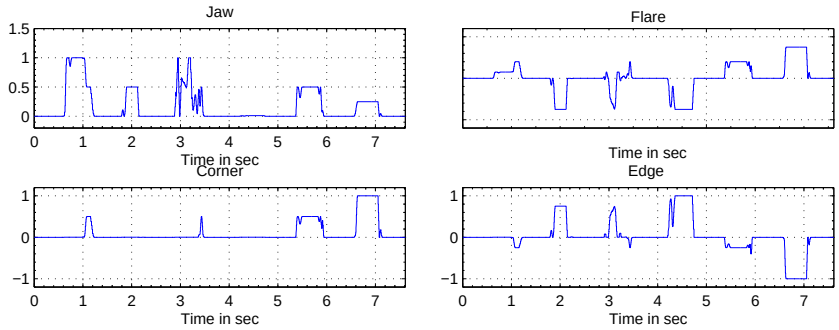

Figure 4.24: Lip-shape parameters of syllables in Fig. 4.22 calculated using 2D Delaunay triangulation.

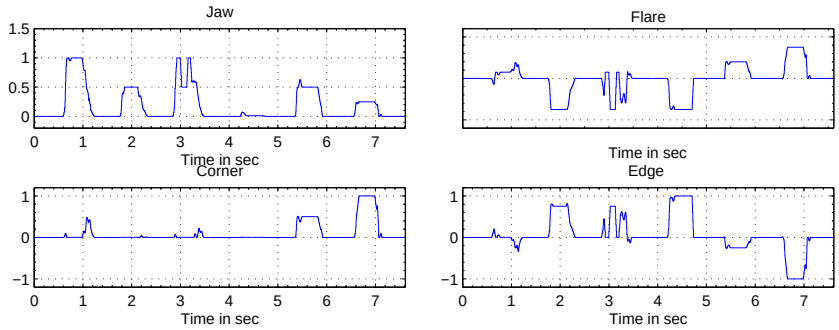

Figure 4.25: Lip-shape parameters of syllables in Fig. 4.22 calculated using 3D Delaunay triangulation.

Using the training moments for Fig 4.22, we wish to analyze the mouth parameters for the composite syllables for the same speaker. Here we have experimented with the syllables /*aa*Iye/ and /*aya*/-/*awa*/ for the same speaker. As in Fig 4.26, the syllable /*aa*Iye/ is uttered twice. It is observed from the Figures 4.27 and 4.29 that the lower jaw moves down for the phoneme /*a*/, then gradually moves to position defined for silence for the phoneme /i/, and then moves down for the phoneme /e/ at the end of the syllable. Moreover, the lips protrude forward and then backward during the utterances of syllable /Iye/, as observed from the flare value. Moreover the horizontal opening of the mouth as indicated by edge parameter also becomes small during the utterance of the syllable /Iye/.

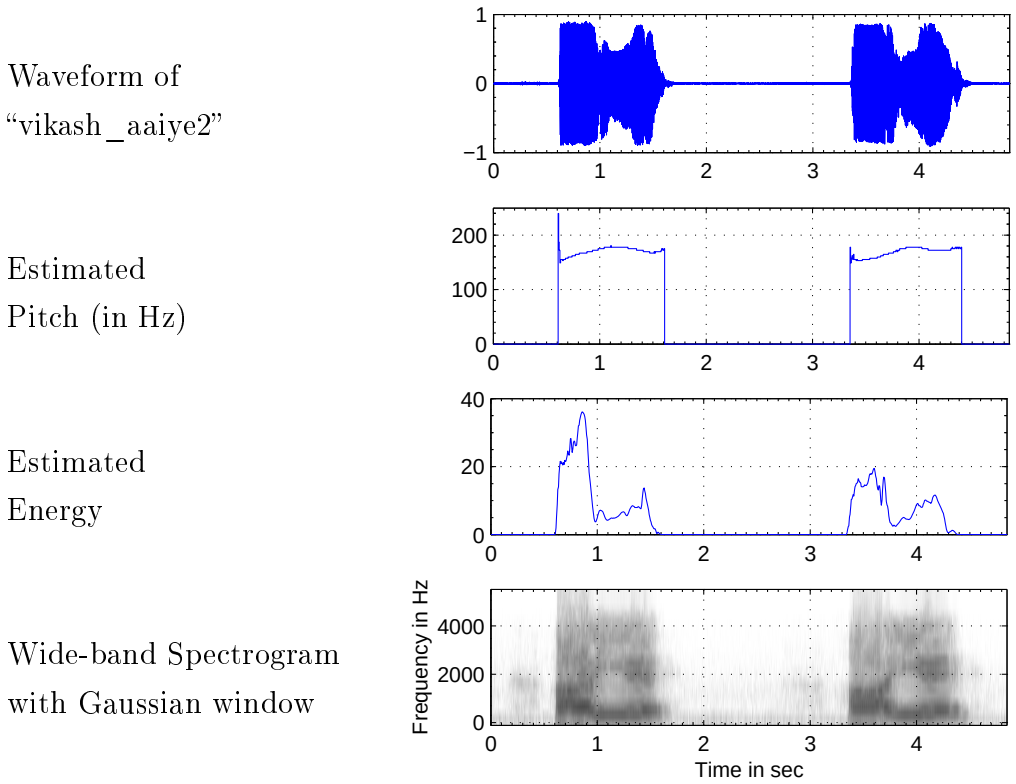

Figure 4.26: Characteristics of natural syllables /*aa*Iye/ by speaker "Vikash"

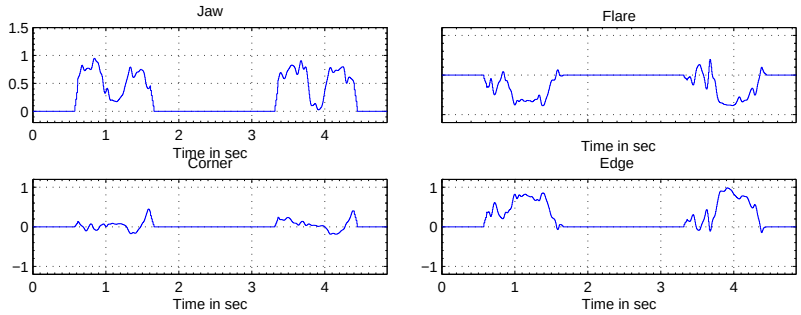

Figure 4.27: Lip-shape parameters of syllables in Fig. 4.26 calculated using least squares approximation.

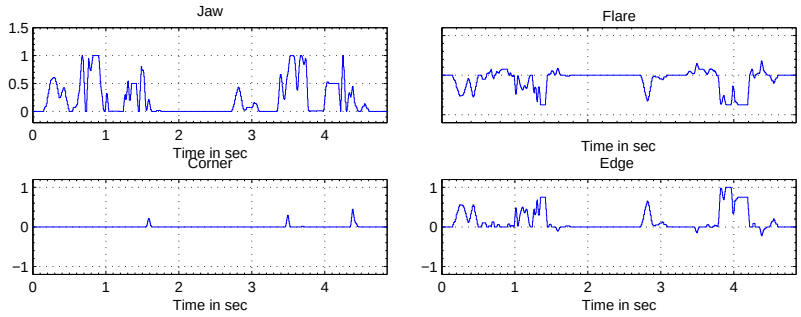

Figure 4.28: Lip-shape parameters of syllables in Fig. 4.26 calculated using 2D Delaunay triangulation.

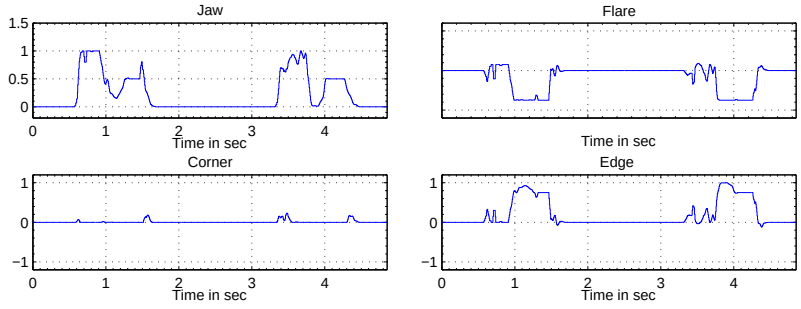

Figure 4.29: Lip-shape parameters of syllables in Fig. 4.26 calculated using 3D Delaunay triangulation.

Similarly from Figures 4.31 and Figures 4.33 for the syllables /*aya*/ and /*awa*/, the jaw parameter takes a very high value during utterance of /*a*/, goes to a low value for phoneme /y/ or /w/ and then again rises to a high value at the end. Moreover, during the utterance of /w/, the lips protrude forward as indicated by flare parameter /w/ and the mouth has a small horizontal opening as observed in the same figures. Moreover the actual meeting of the lips in the horizontal direction moves toward the centre during the utterance of /w/, whereas there is little motion similar to /i/ during the utterance of /y/ as indicated in these figures by the parameter corner. Similar results with random fluctuations at some points are also observed using 2D Delaunay triangulation as in Fig 4.32.

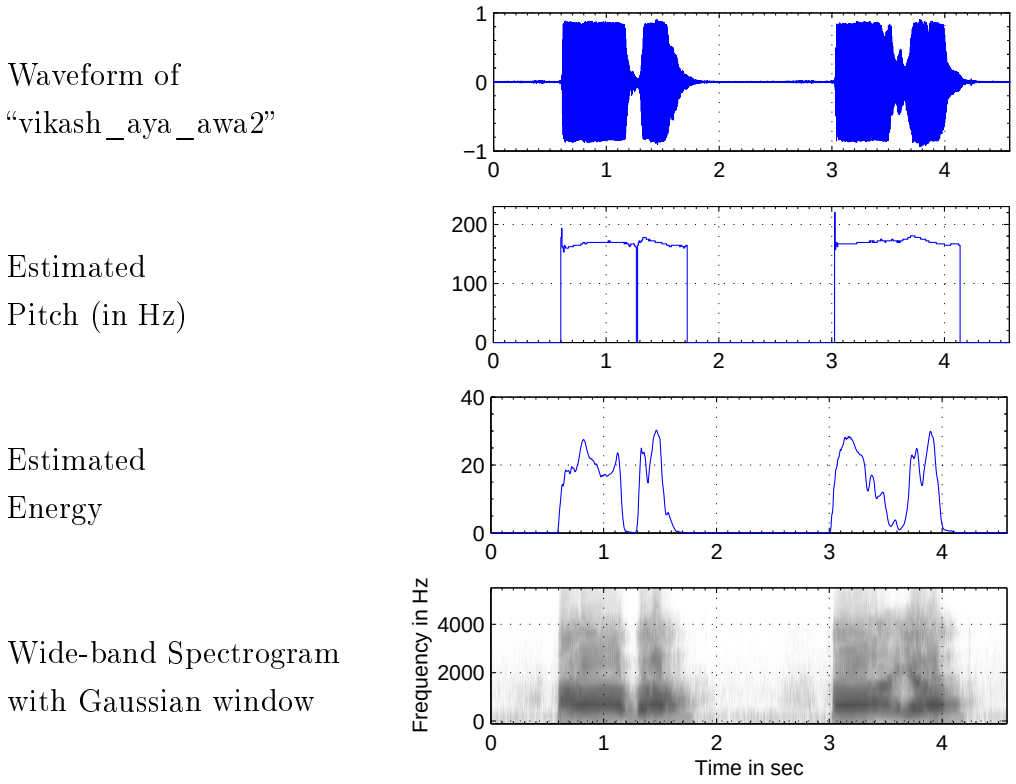

Figure 4.30: Characteristics of natural syllables /*aya*/-/*awa*/ by speaker "Vikash"

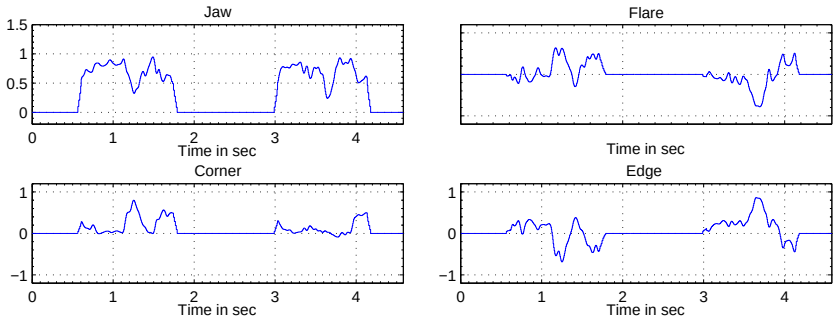

Figure 4.31: Lip-shape parameters of syllables in Fig. 4.30 calculated using least squares approximation.

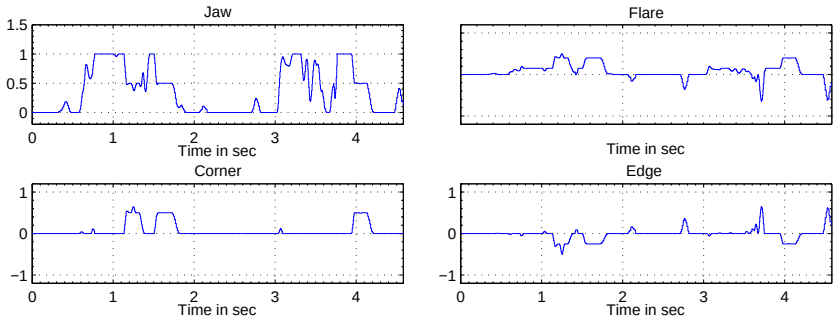

Figure 4.32: Lip-shape parameters of syllables in Fig. 4.30 calculated using 2D Delaunay triangulation.

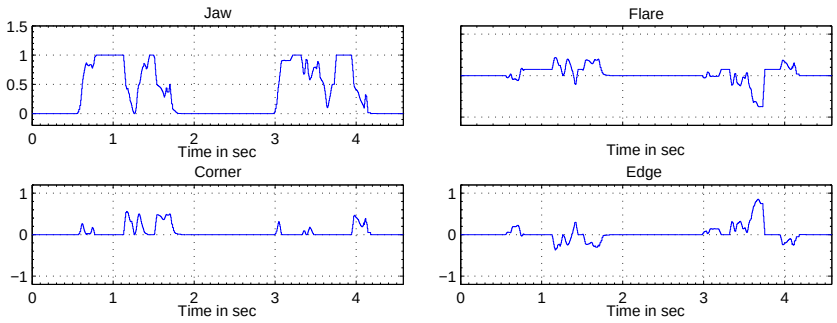

Figure 4.33: Lip-shape parameters of syllables in Fig. 4.30 calculated using 3D Delaunay triangulation.

4.5 Lip shape analysis for speaker "Mani"

We next analyze the whole method for a different speaker "Mani". The selection of moments used for training is the same as described earlier for synthesized vowels. As expected, the results comply with the data given in Table 3.1. Also the 3D Delaunay triangulation method gives smoother results than the other two.

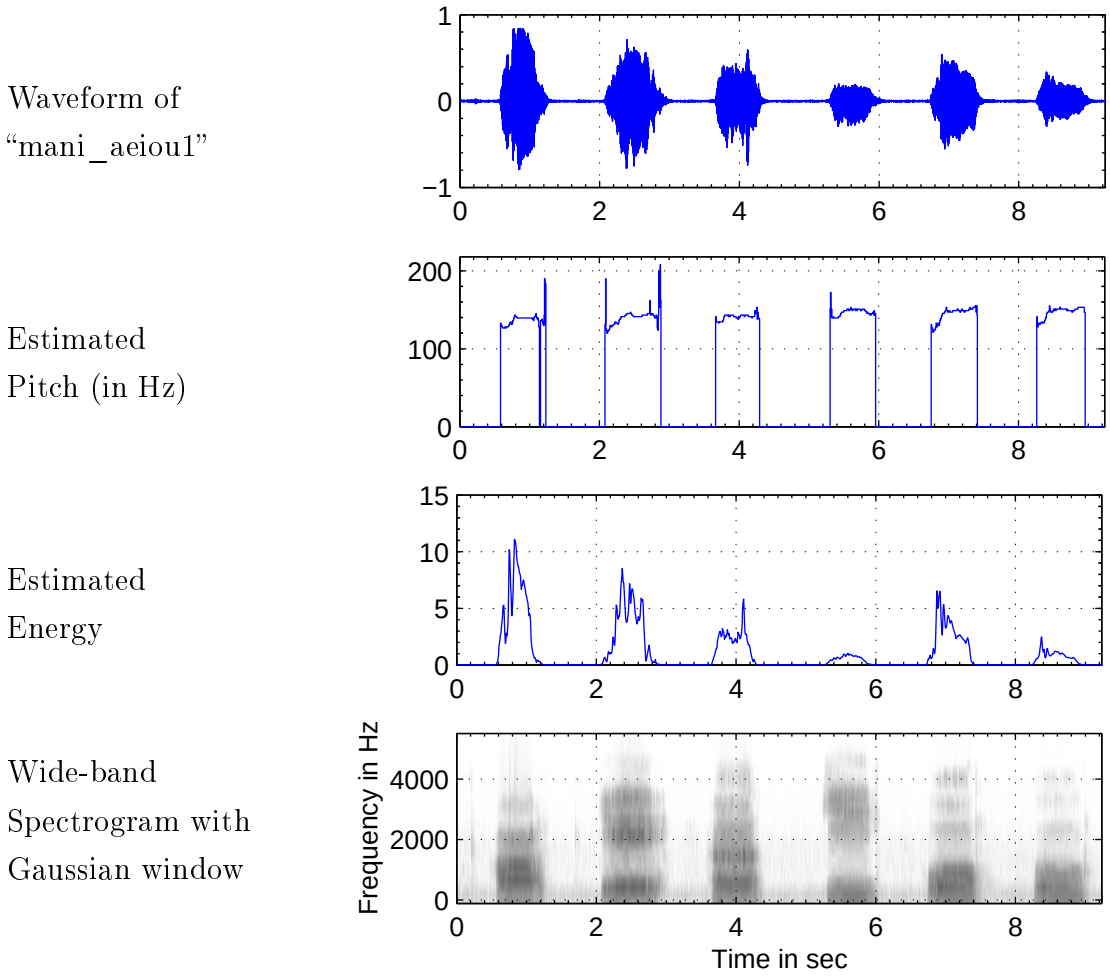

Figure 4.34: Characteristics of natural syllables /a/, /e/, /œ/, /i/, /o/, /u/ by speaker "Mani"

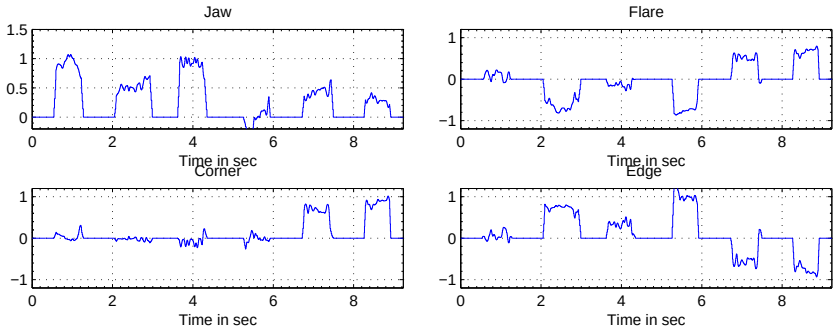

Figure 4.35: Lip-shape parameters of syllables in Fig. 4.34 calculated using least squares approximation.

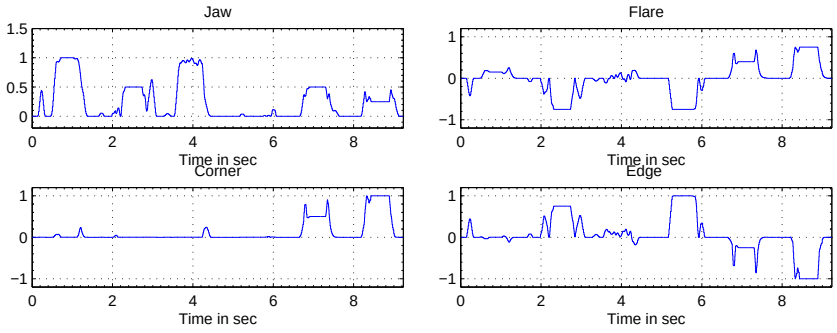

Figure 4.36: Lip-shape parameters of syllables in Fig. 4.34 calculated using 2D Delaunay triangulation.

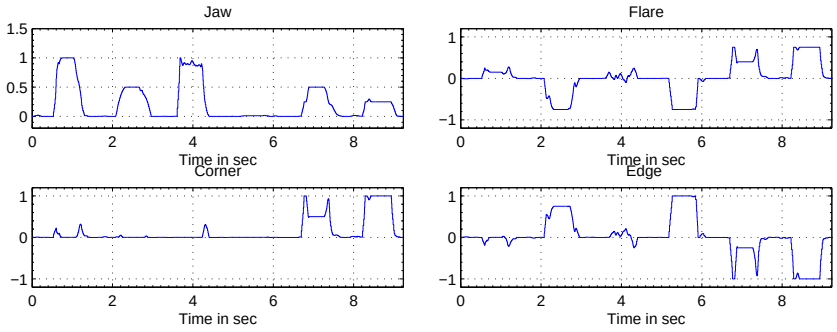

Figure 4.37: Lip-shape parameters of syllables in Fig. 4.34 calculated using 3D Delaunay triangulation.

We now wish to analyze the mouth parameters for the composite syllables for the same speaker "Mani". Here we have experimented with the syllables /*aa*o/, /*aa*Iye/, /*aya*/, /*awa*/, /*ayi*/, /*awi*/, /*ara*/ and /*ala*/ for the same speaker. It is observed from the Figures 4.39 and 4.41 that the jaw parameter takes a very high value for the phoneme /*a*/, then gradually decreases to a very low value for the phoneme /o/. Moreover, the lips protrude forward at the end of the syllable, as observed from the flare value. The horizontal opening of the mouth, as indicated by the parameter 'edges', becomes small during the end of the syllable.

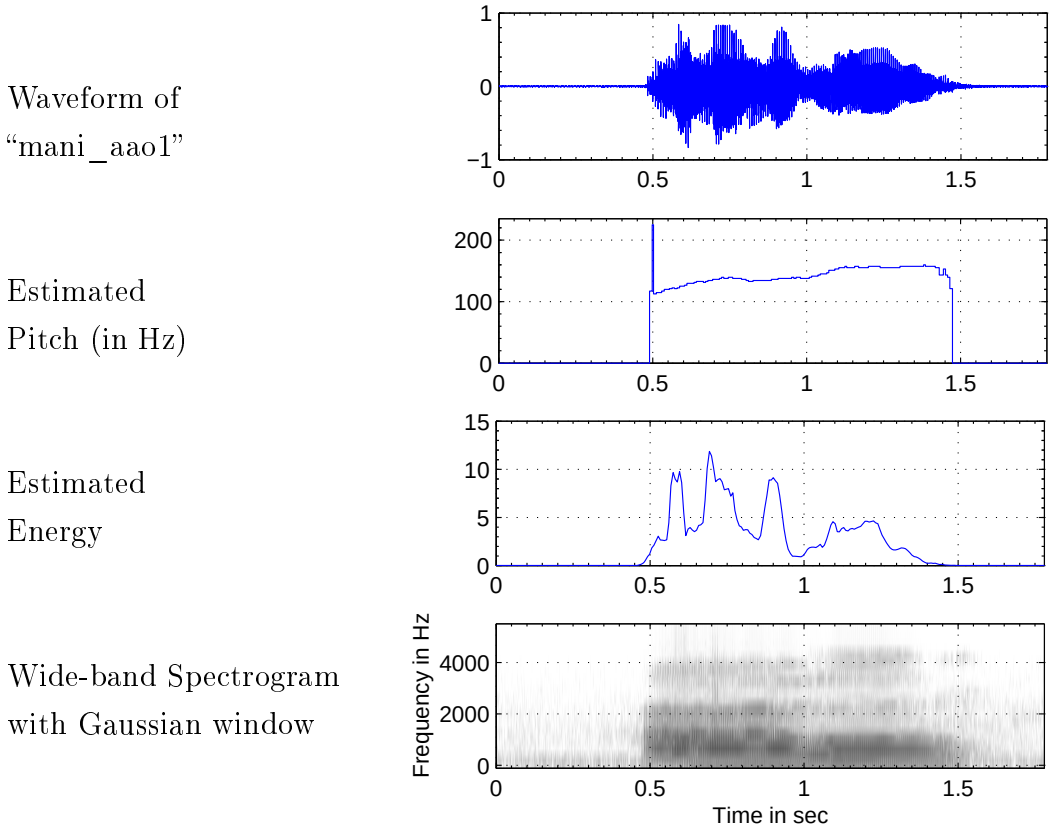

Figure 4.38: Characteristics of natural syllables /*aa*o/ by speaker "Mani"

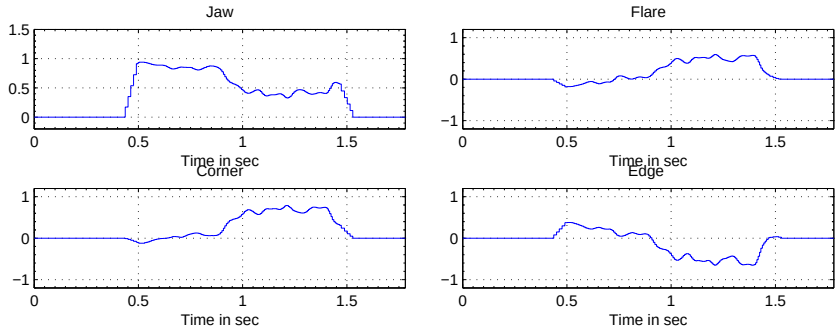

Figure 4.39: Lip-shape parameters of syllables in Fig. 4.38 calculated using least squares approximation.

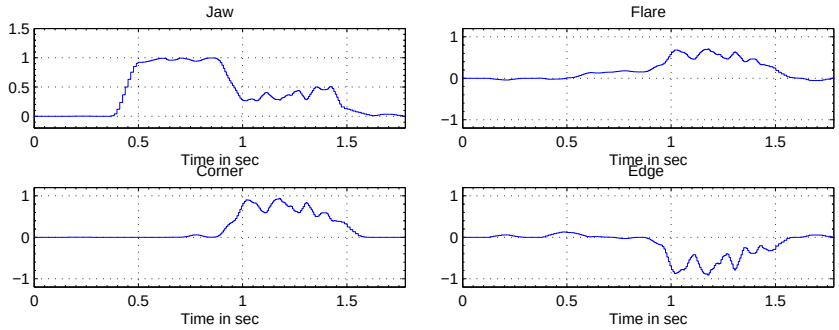

Figure 4.40: Lip-shape parameters of syllables in Fig. 4.38 calculated using 2D Delaunay triangulation.

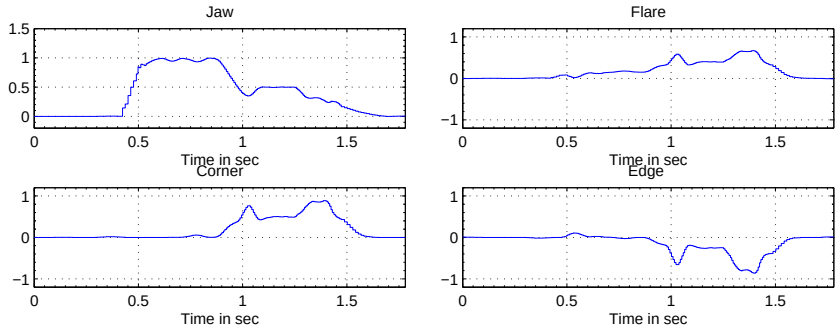

Figure 4.41: Lip-shape parameters of syllables in Fig. 4.38 calculated using 3D Delaunay triangulation.

Similar results are obtained for the mouth parameters for the syllable /*aa*Iye/ as shown in Figures 4.43, 4.44 and 4.45 for speaker "Mani" as for speaker "Vikash". However, the effect of noise while using 2D Delaunay triangulation for calculation of mouth parameters becomes more prominent as observed in Fig 4.44. The reason is because of the exclusion of the noise spectrum from the reference data. As observed in Fig 4.45, the inclusion of the moments of the noise spectrum in the reference data actually improves the results.

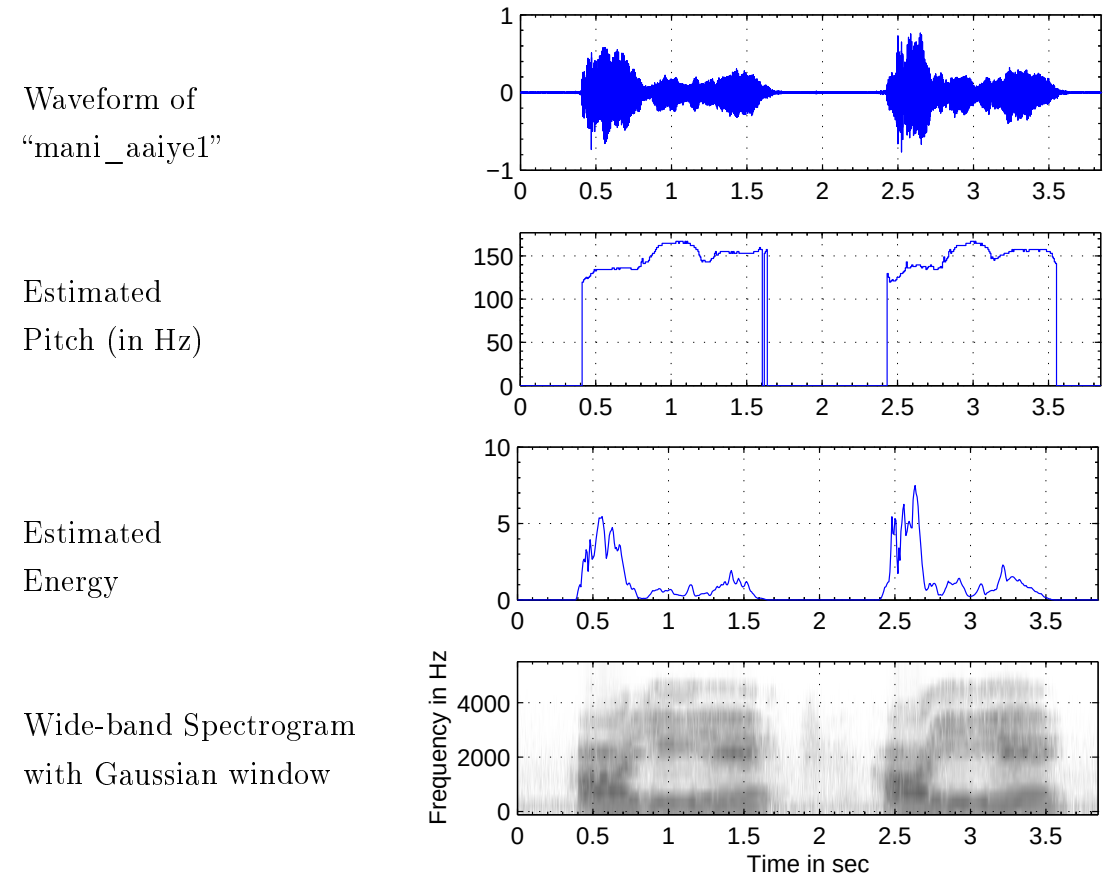

Figure 4.42: Characteristics of natural syllables /*aa*Iye/ (uttered twice) by speaker "Mani"

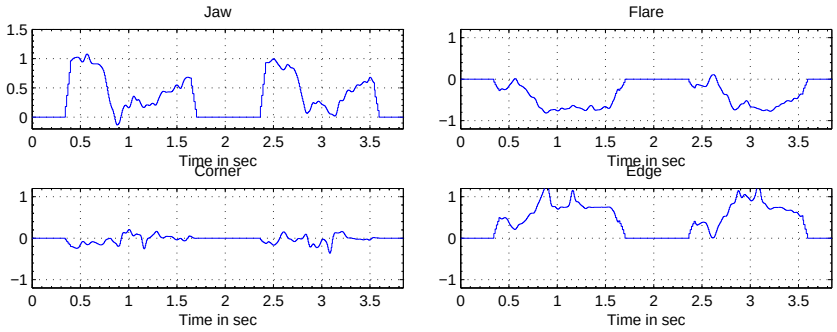

Figure 4.43: Lip-shape parameters of syllables in Fig. 4.42 calculated using least squares approximation.

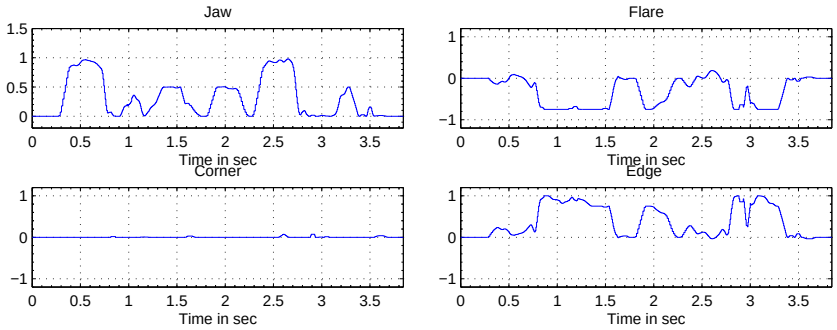

Figure 4.44: Lip-shape parameters of syllables in Fig. 4.42 calculated using 2D Delaunay triangulation.

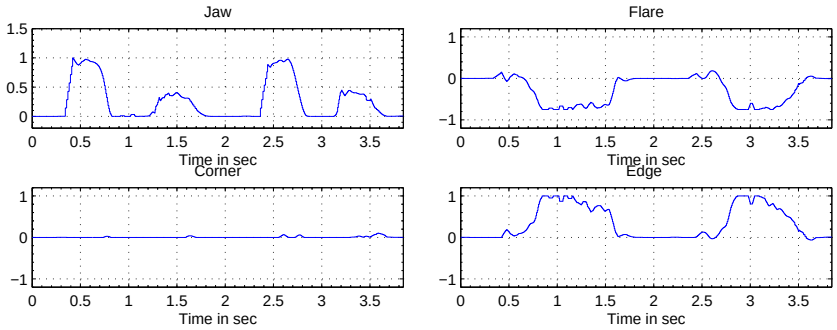

Figure 4.45: Lip-shape parameters of syllables in Fig. 4.42 calculated using 3D Delaunay triangulation.

Similarly from Figures 4.47 and 4.49 for the syllables /*aya*/-/*awa*/, we observe that similar results are obtained for speaker "Mani" as well. However, from Fig 4.47, the flare value during utterance of /w/ goes negative i.e. mouth protruding out wards which is not correct. Also the effect of noise spectrum while using 2D Delaunay triangulation for the calculation of mouth parameters is also significant.

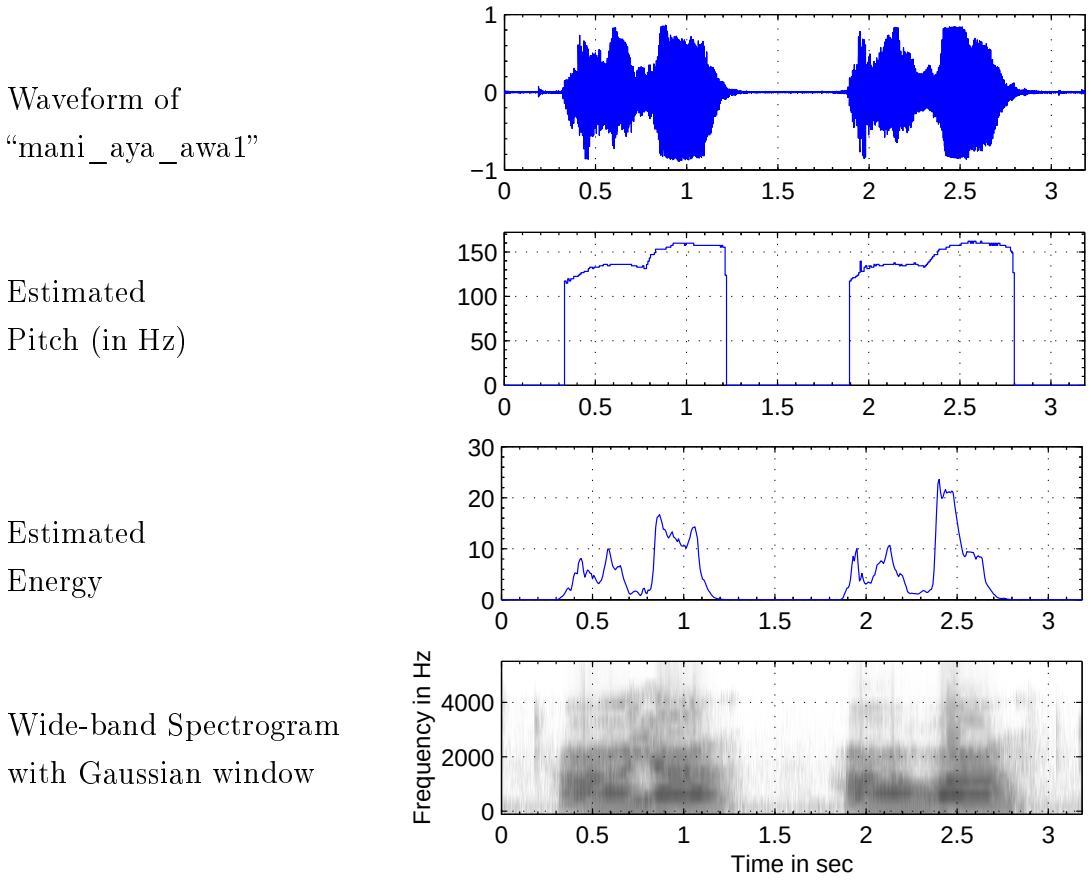

Figure 4.46: Characteristics of natural syllables /*aya*/-/*awa*/ by speaker "Mani"

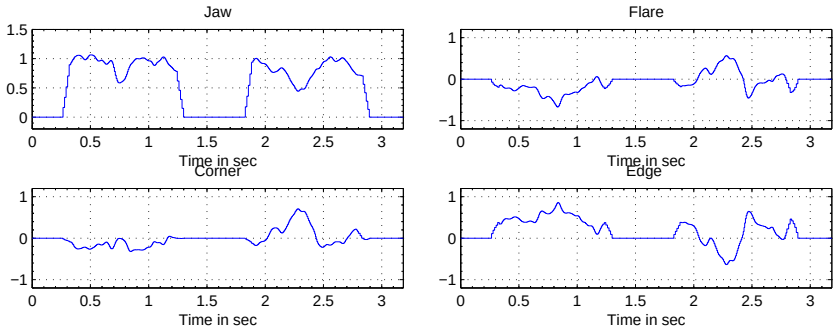

Figure 4.47: Lip-shape parameters of syllables in Fig. 4.46 calculated using least squares approximation.

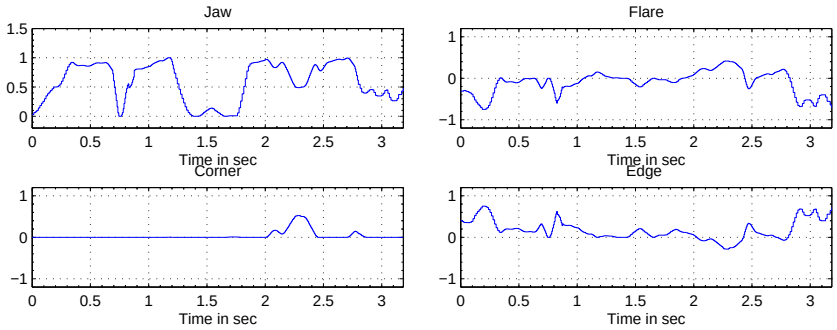

Figure 4.48: Lip-shape parameters of syllables in Fig. 4.46 calculated using 2D Delaunay triangulation.

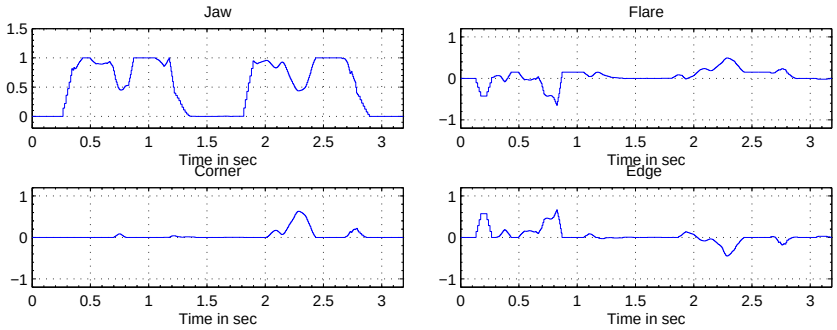

Figure 4.49: Lip-shape parameters of syllables in Fig. 4.46 calculated using 3D Delaunay triangulation.

Figures 4.51, 4.52 and 4.53 show the results for the syllables /*a*yi/ and /*a*wi/. The jaw contour for /*a*yi/ indicates that the lower jaw moves down at the start of the former syllable, remains flat till the phonemic transition to /y/ begins, and then moves up to the position attained by phoneme /i/, Similar is the case with the latter syllable /*a*wi/. Moreover, the lips protrude forward during the utterance of /y/ and /w/ till the end of the utterance of /i/ as indicated by the flare parameter. Also, the contour for corners show that the lip joints try to move to the position attained by /u/ during the utterance of /w/, but soon move back to the position attained by phoneme /i/. It is also confirmed that 3D Delaunay triangulation method gives smoother results than the other two.

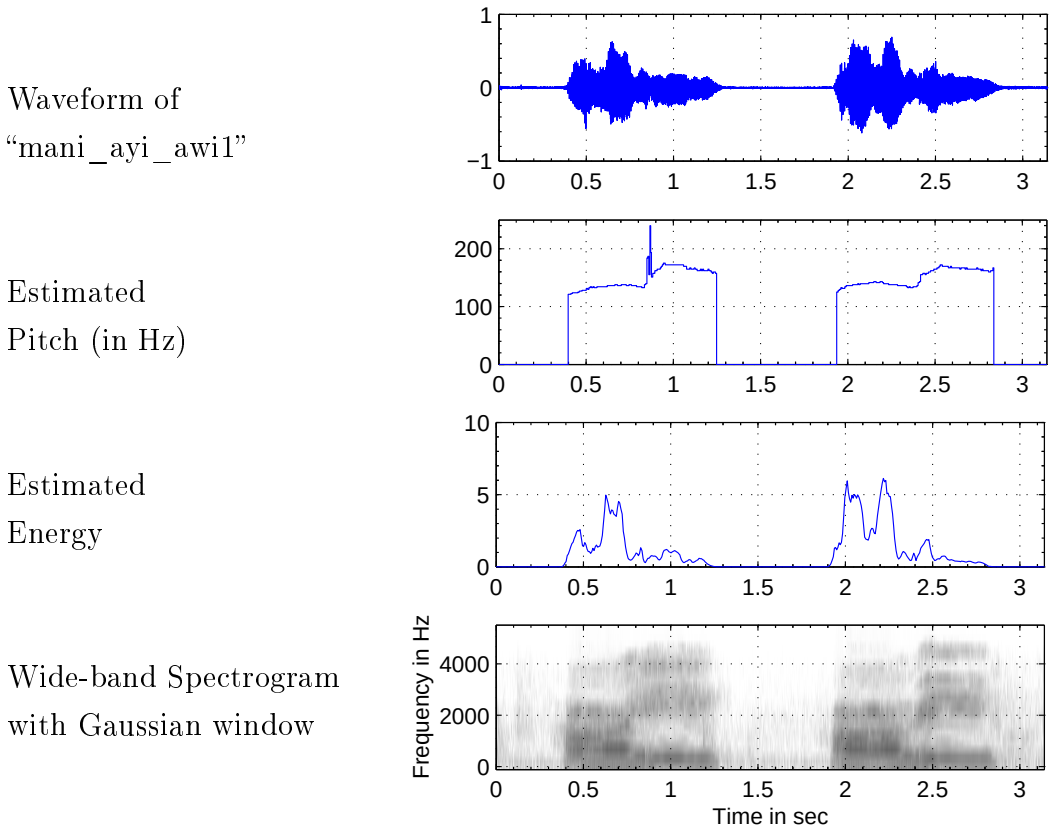

Figure 4.50: Characteristics of natural syllables /*a*yi/-/*a*wi/ by speaker "Mani"

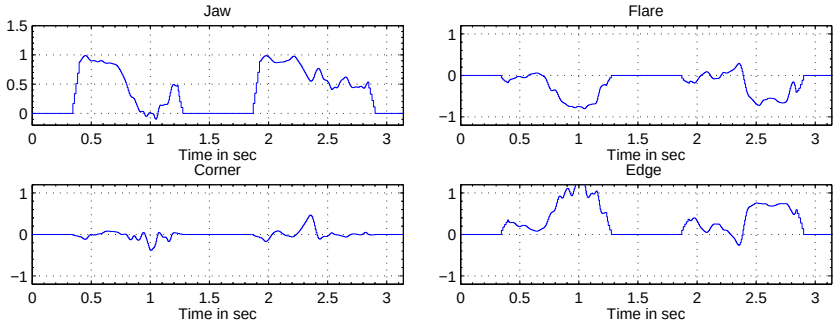

Figure 4.51: Lip-shape parameters of syllables in Fig. 4.50 calculated using least squares approximation.

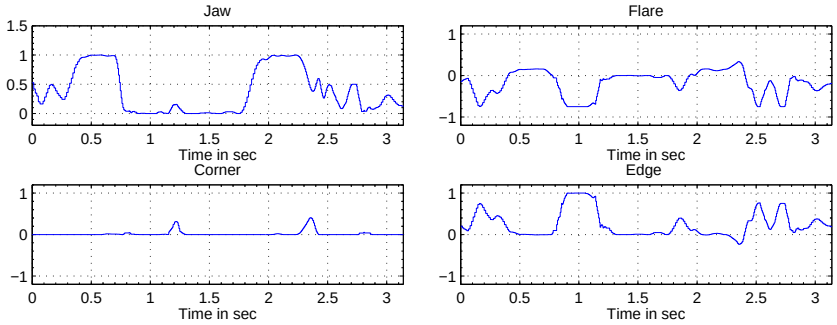

Figure 4.52: Lip-shape parameters of syllables in Fig. 4.50 calculated using 2D Delaunay triangulation.

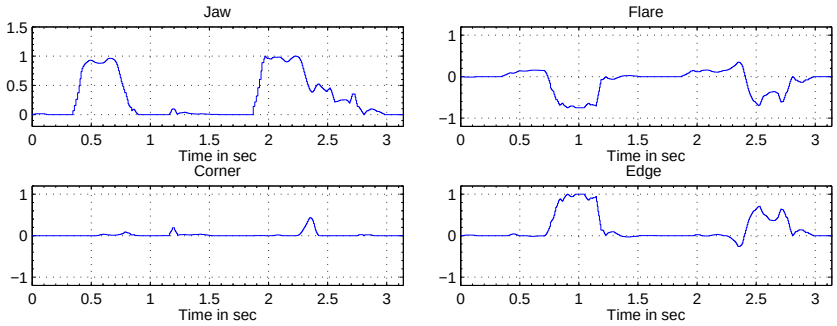

Figure 4.53: Lip-shape parameters of syllables in Fig. 4.50 calculated using 3D Delaunay triangulation.

4.6 Discussion

Analysis of the pitch period determination method was carried out for vowels, semivowels, fricatives and stops. The pitch period tracker was able to identify the pitch changes for vowels, semivowels and voiced fricatives. For syllables containing low energy content phonemes like unvoiced fricatives, voiced, and unvoiced stops, the pitch period tracker was able to detect the pitch changes in the strong voiced portions of the speech, and showed the voice onset time contrast for unvoiced and voiced stops. The tracker worked well for speech segments with SNR as low as 20 dB. For segments with SNR below 20 dB, the effect of noise caused the tracker to give random values. Next, the spectral moments of the speech signal were computed, and when the mean and variance of the vowels were plotted on a mean-vs-variance space, we observed that vowels, whether synthesized or natural, occupied similar but different positions for different speakers. Also, the value of the lip-shape parameters can be ascertained from the mean-variance space. It is observed from Figures 4.10, 4.11, 4.12, and 4.13, that the jaw parameter takes a high value for low values of variance, as indicated by the positions of /a/. Also, the low values of mean results in lip rounding as indicated by the positions of /o/ and /u/. This implies that the spectral moments can be used for finding the lip parameters.

Analysis was then carried out for mapping the spectral moments of the speech segment to the lip-shape parameters for vowels and semivowels. It was observed that least squares approximation method and 3D Delaunay triangulation method produced results in conformity with the predefined values in Table 3.1 for synthesized and natural vowels. For vowel-vowel and vowel-semivowel-vowel segments, co-articulation effect was observed during phonemic transition. In case of 2D Delaunay triangulation method, lip-shape parameters for noise segments were non-zero, thus deviating from the predefined values. Also the 3D Delaunay triangulation method provided smoother results than the other two.

Chapter 5
Conclusion

5.1 Summary

For hearing impaired persons, the visual information from the speaker's face can efficiently integrate or even substitute audio information for understanding speech [1] [2]. However, because of limited channel capacity and storage constraints, it is not feasible for the videophones to send every visual frame over the transmission channel. The solution is to extract certain speech parameters that would help in visualizing speech, which is the objective of this project.

In literature, Waters and Levergood [8] developed DECface as a real time facial animation capable of generating synchronous speech and mouth motion from text. Liéven and Luthon [12] have used RGB information of the speaker's mouth for synthesizing lip-movements without any reference to audio information. Other approaches include HMM-based visual speech synthesis [1] [10] [11]. Some of the requirements of the above methods are the colour video sequence of the speaker's face and the availability of audio-visual speech database containing visemes related to all known phonemes, which in turn requires a lot of storage capacities, thereby increasing the cost. Massaro et al [13] developed an audio to visual speech synthesis using neural network, which synthesized visual speech from the cepstral coefficients extracted directly from acoustic speech waveform. McAllister et al [7] [15] [25] have used spectral moments of the pitch synchronously computed speech spectra for generating the lip shapes.

In this project, a GUI package in Matlab has been developed for implementation and evaluation of the various analysis techniques. Speech segments for vowels /*a*/, /e/, /i/ , /o/, /u/ were recorded from different speakers as part

of research for training sounds. Also segments containing semivowels, fricatives and stops preceded and succeeded by vowels and vowel-vowel combination were recorded from the same speakers. The implementation of the pitch period tracker is based on the algorithm, as proposed by McAllister et al [15], that the sum of the amplitudes of the odd samples of the magnitude spectrum of a speech segment is zero for a window length of twice the pitch period. Modifications in McAllister et al's approach have been made for finding the correct estimate of the pitch period if the pitch period range contains the pitch period and its multiples. The pitch period tracker is able to identify the changes in the pitch for vowels, semivowels and voiced fricatives. For syllables containing low energy content phonemes like unvoiced fricatives, voiced, and unvoiced stops, the pitch period tracker is able to detect the pitch changes in the strong voiced portions of the speech, and shows the voice onset time contrast for unvoiced and voiced stops. The tracker works well for speech segments with SNR as low as 20 dB. For segments with SNR below 20 dB, the effect of noise causes the tracker to give random values.

In accordance with McAllister et al's approach, mass, weighted mean, and weighted variance of the average magnitude spectrum estimated with a window length equal to the pitch period P_g and actual frequencies of the samples as the weights have been calculated. When the mean and variance of the vowels were plotted on a mean-vs-variance space, we observed that vowels, whether synthesized or natural, occupy similar but different positions for different speakers. This implies that the spectral moments can be used for finding the lip parameters.

In accordance with McAllister et al's approach [7], four parameters such as jaw, edges, flare and corners have been used for generating the lip shapes. Certain values have been pre-defined for these parameters for the training sounds. Krothapalli et al [7] [30] have used 2D Delaunay triangulation method for finding the lip shape parameters from the spectral moments. We have used least squares approximation and 3D Delaunay triangulation methods and compared the results with that obtained from 2D Delaunay triangulation method. Speech segments /a/, /e/, /i/, /o/ and /u/ were used in the analysis stage. Spectral moments of these training sounds were used as the reference data for Delaunay triangulation in the synthesis stage. For least squares approximation and 2D Delaunay

triangulation methods, the spectral moments of the noise segments were forced to zero in accordance with the predefined values of the mouth parameters. In the synthesis stage, vowel-vowel and vowel-semivowel-vowel segments were used to verify the correctness and accuracy of the methods. Results have shown that 3D Delaunay triangulation method gives more smoother and correct results than the other two. It was also observed that the parameters obtained using 2D Delaunay triangulation method were non-zero for the noise segments, thus deviating from the predefined values. Moreover, the results obtained using the least squares approximation method were closer to that obtained using 3D Delaunay triangulation method. Our conclusion is that for storage constraints, the least squares approximation method with appropriate filtering can be used for mapping the spectral moments to the lip shape parameters.

5.2 Suggestions for future work

After the extraction of mouth parameters, animation of the mouth shape depicting the lip-movements has to be designed. Such animation can be both in 2D and 3D space. Further, the facial motion tracking can be implemented to add facial expressions to the animated mouth shape. It is to be noted here that during stop closures, analysis cannot give us lip shape. Techniques developed in our lab [5] for estimation of vocal tract shape during stop closure may be extended to lip shape estimation during stop closures.

References

[1] J. J. Williams and A. K. Katsaggelos, "An HMM based speech to video synthesizer," *IEEE Trans. Neural Networks*, vol. 13, pp. 900–915, July 2002.

[2] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. Rehab. Eng.*, vol. 3, Mar. 1995.

[3] S. Kshirsagar, "A speech training aid for hearing impaired," Master's thesis, Supervisor: Prof. P. C. Pandey, Dept. of Electrical Engineering, IIT Bombay, 1998.

[4] M. S. Shah and P. C. Pandey, "Areagram display for investigating the estimation of vocal tract shape for a speech training aid," in *Proc. Symposium on Frontiers of Research on Speech and Music*, (IIT Kanpur), pp. 121–124, Feb. 15-16 2003.

[5] M. S. Shah and P. C. Pandey, "Estimation of vocal tract shape during stop closures," in *Proc. Int. Conf. on Systemics, Cybernetics, Informatics, ICSCI 2004*, (Hyderabad, India), pp. 304–309, Feb. 12-15 2004.

[6] R. Rodman, D. McAllister, D. Bitzer, H. Fu, and B. Xu, "A pitch tracker for identifying voiced consonants," in *Proc. 10th Int. Conf. Signal Processing Applications and Technology, ICSPAT'99*, Nov. 1999.

[7] C. Krothapalli, "Developing predictor surfaces for vowels and voiced fricatives for lip synchronisation," Master's thesis, Supervisors: D. McAllister, R. Rodman, and D. Bitzer, Dept. of Computer Science, North Carolina State University, 2000. http://www.lib.ncsu.edu/etd/public/etd-639118510151581/etd.pdf.

[8] K. Waters and T. M. Levergood, "DECface: An automatic lip synchronization algorithm for synthetic faces," Technical Report Series, Digital Equipment Corporation, Cambridge Research Lab, Aug. 1993. http://www.hpl.hp.com/techreports/CompaqDEC/CRL934.html.

[9] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971–995, Mar. 1980.

[10] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda, "Visual speech synthesis based on parameter generation from HMM," in *Proc. Int. Conf. Auditory Visual Speech Processing*, (Terrigal, Sydney, Australia), pp. 219–224, Dec. 1998.

[11] E. Yamamoto, S. Nakamura, and K. Shikano, "Subjective evaluation for HMM based speech to lip movement synthesis," in *Proc. Audio Visual Speech Processing AVSP98*, (Terrigal, Sydney, Australia), pp. 227–232, Dec. 1998.

[12] M. Liévin and F. Luthon, "Unsupervised lip segmentation under natural conditons," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing, ICASSP'99*, vol. 6, (Phoenix, Ariz., USA), pp. 3065–3068, Mar. 1999.

[13] D. W. Massaro, J. Beskow, M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proc. Audio Visual Speech Processing, AVSP'99*, (Santa Cruz, Cal., USA), pp. 133–138, Aug. 1999.

[14] F. Lavagetto, D. Arzarello, and M. Caranzano, "Lipreadable frame animation driven by speech parameters," in *Proc. Int. Symp. Speech, Image Processing Neural Networks*, (Hong Kong, China), pp. 626–629, Apr.13-16 1994.

[15] D. McAllister, R. Rodman, D. Bitzer, and A. Freeman, "Lip synchronisation as an aid to hearing impaired," in *Proc. American Voice Input/Output Society, AVIOS'97*, (San Jose, CA, USA), pp. 233–248, 1997.

[16] G. Leach, "Improving worst-case optimal Delaunay triangulation algorithms," in *Proc. 4th Canadian Conf. Comput. Geom.*, (St. John's, Newfoundland, Canada), June 1992.

[17] T. Lambert, "An optimal algorithm for realizing a Delaunay triangulation," *Information Processing Letters*, vol. 62, pp. 245–250, June 1997.

[18] Z. Jian-ming, S. Ke-ran, Z. Ko-ding, and Z. Qiong-hua, "Computing constrained triangulation and Delaunay triangulation: A new algorithm," *IEEE Trans. Magnetics*, vol. 26, pp. 694–696, Mar. 1990.

[19] T. P. Fang and L. A. Piegl, "Delaunay triangulation in three dimensions," *IEEE Comput. Graph. Appl.*, vol. 15, pp. 62–69, Sept. 1995.

[20] L. Rognant, J. M. Chassery, S. Goze, and J. G. Planes, "The Delaunay constrained triangulation: The Delaunay stable algorithms," in *Proc. Int. Conf. Information Visualisation*, (London, England), pp. 147–152, Jul 14-16 1999.

[21] O. R. Musin, "Properties of Delaunay triangulation," in *Proc. 13th Annual ACM Symp. Comput. Geom.*, (Nice, France), pp. 424–426, 1997.

[22] Matlab v6.1 R12, "The mathworks worldwide, products & services," 2002. http://www.mathworks.com.

[23] GoldWave Digital Audio Editor, "Goldwave version 5 download," 2002. http://www.goldwave.com/release.php#download.

[24] R. Rodman, D. McAllister, D. Bitzer, and D. Chappell, "A high resolution glottal pulse tracker," in *Proc. Int. Conf. Spoken Language Processing, ICSLP 2000*, vol. 3, pp. 881–884, Oct. 2000.

[25] D. F. McAllister, R. Rodman, D. Bitzer, and A. Freeman, "Toward speaker independence in automated lip-sync," in *Proc. Compugraphics*, (Algarve, Portugal), pp. 10–15, 1997.

[26] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, ch. 6,8, pp. 310–315,396–453. Englewood Cliffs, New Jersey: Prentice Hall, 1978.

[27] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 4.1.4)," 2002. http://www.fon.hum.uva.nl/praat/.

[28] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise corrupted speech," *IEEE Trans. Signal Processing*, vol. 39, pp. 319–329, Feb. 1991.

[29] M. Wang, D. Bitzer, D. McAllister, R. Rodman, and J. Taylor, "An algorithm for V/UV/S segmentation of speech," in *Proc. 2001 Int. Conf. Speech Processing (ICSP 2001)*, (Acoustic Society of Korea, Seoul, Korea), pp. 541–546, Sept. 2001.

[30] C. Krothapalli, D. McAllister, R. Rodman, D. Bitzer, M. Wang, and J. Taylor, "Predictor surfaces for lip synchronization animation of voiced input," in *Proc. 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001)*, vol. 7, (Skokie, Ill., USA), Aug. 2001.

[31] J. Taylor, D. Bitzer, R. Rodman, D. McAllister, and M. Wang, "Speaker independence in lip synchronization of vowels," in *Proc. 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th Int. Conf. Information Systems Analysis and Synthesis (ISAS 2001)*, (Skokie, Ill., USA), Aug. 2001.

[32] J. Taylor, D. Bitzer, R. Rodman, and D. McAllister, "Achieving speaker independence in automatic lip synchronization," in *Proc. the Applied Voice Input/Output Society, AVIOS 2001*, (San Jose, Cal., USA), pp. 163–171, Sept. 2001.

[33] M. Sarfraz, Z. Habib, and M. Hussain, "Piecewise interpolation for designing of parametric curves," in *Proc. Conf. Information Visualisation*, (London, England), pp. 307–313, Jul. 29-31 1998.

[34] G. Farin and N. Sapidis, "Curvature and the fairness of curves and surfaces," *IEEE Comput. Graph. Appl.*, vol. 9, pp. 52–57, Mar. 1989.

[35] S. A. Dyer and X. He, "Least-squares fitting of data by polynomials," *IEEE Instrum. Meas. Mag.*, vol. 4, pp. 46–51, Dec. 2001.