Subspace and Hypothesis based Effective Segmentation of Co-articulated Basic-units for Concatenative Speech Synthesis

R. Muralishankar, R. Srikanth and A. G. Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science
Bangalore, INDIA
080-293-2935
sripad, srikanth, ramkiag@ee.iisc.ernet.in

Abstract*— In this paper, we present two new methods for Vowel-Consonant segmentation of a co-articulated basic-units employed in our Thirukkural Tamil Text-to-Speechsynthesis system [1]. The basic-units considered in [1]are CV, VC, VCV, VCCV and VCCC, where C stands for a consonant and V for any vowel. In the first method, we use subspace-based approach for ***vowel-consonant segmentation. It uses ***orientedprincipal component analysis*** (OPCA) where the test feature vectors are projected on to the V and C subspaces. The crossover of the norm-contours obtained by projecting test basic-unit onto the V and C subspaces give the segmentation points which in turn helps in identifying the V and C durations of a test basic-unit. In the second method, we use ***probabilistic principal component analysis*** (PPCA) [2] to get probability models for V and C. We then use Neymen-Pearson (NP) test to segment the basic-unit into V and C. Finally, we show that the hypothesis testing turns out to be an energy detector for V-C segmentation which is similar to the first method.**

1. INTRODUCTION

For the purpose of recognition or synthesis, speech often needs to be segmented into phonetic units. Manual segmentation is tedious, time consuming and error prone. Due to variability both in human visual and acoustic perceptual capability, it is almost impossible to reproduce the manual segmentation results. Hence manual segmentation is inherently inconsistent. Automatic sementation is not faultless, but it is inherently consistent and results are reproducible. Ideally, one likes to have an automatic segmentation which can handle basic-units uttered by different speakers. There are two broad categories of speech segmentation [3] namely, ***implicit*** and ***explicit.*** Implicit methods split up the utterance without explicit information, such as the phonetic transcription, and are based on the definition of a segment as a spectrally stable part of a signal.

Motivation for Subspace bused Segmentation

In our synthesis scheme [1], concatenation is always performed across identical vowels. Changes in duration, pitch and amplitude are obtained by processing the vowel parts of the basic-units only. Thus, the segmentation of basic-units into ***vowel*** and ***consonant*** parts is needed to keep the consonant portion of the waveform intact. Plosives, affricates and fricatives have a common property of low energy when compared with any of the vowels. Figure 1 shows the performance of energy based segmentation for plosive and co-articulated basic-units. As shown in Figs. 1(a) and (b), accurate segmentation can be obtained for non co-articulated units, and not for co-articulated basic-units. The ttue consonant part /y/ in the signal /eyo/ is shown in Fig. 1(b) with the boundaries dotted.

We propose two methods for co-articulated basic-unit segmentation. In the first method, we use subspace approach using ***orientedprincipal component analysis*** (OPCA) for basic-unit segmentation. In the second method, we use ***probabilisticprincipal component analysis*** (PPCA) [2] to obtain probability models for vowel (V) and cconsonants (C). We employ Neymen-Pearson (NP) test using the probability models for basic-unit segmentation.

2. SUBSPACE BASED SEGMENTATION

When we consider an individual vowel or consonant, there exist techniques like LPC to model their statistical properties, While segmenting the vowel part of a basic-unit, we can consider the vowel information (VI) in the feature vectors as the signal and the consonant information (CI) as noise. Similarly when the segmentation of the consonant part is required, we can view CI as signal and VI as noise. We present a linear feature transformation that aims at finding a subspace, of the feature space, in which the Signal-to-Noiseratio (SNR) is maximum. Such a decomposition can be arrived at by representing VI and CI by training vectors obtained using manual segmentation. The directions in the feature space where the SNR is maximum can be obtained by the ***generalized eigenvalue decomposition*** of the covariance matrices of the above vectors. Consider a linear transformation matrix **A'** that maps the original feature vectors x on to $\widehat{x}$.

$$\widehat{x} = W^T x \qquad (1)$$

where x is an n-dimensional vector, $\widehat{x}$ is an m-dimensional vector, $m \leq n$, and W is an n **x** m matrix with m linearly independent columns. Let d_v and d_c represent the training vectors containing VI and CI, respectively, in the original feature space. The covariance martices for these training feature

0-7803-7651-X/03/$17.00 ©2003 IEEE

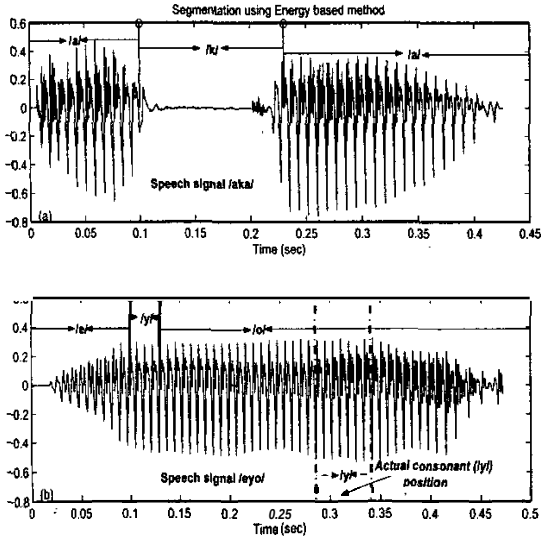

Figure **1.** Basic-unit segmentation using energy based method. (a) Speech signal /aka/. (b) Co-articulated signal feyol (continuous vertical line: segmentaion using energy based method).

vectors can be written as

$$C_v = E[(d_v - \overline{d_v})(d_v - \overline{d_v})^T] \qquad (2)$$
$$C_c = E[(d_c - \overline{d_c})(d_c - \overline{d_c})^T]$$

where $\overline{d_v}$ and $\overline{d_c}$ represent the means of d_v and d_c respectively. We collect an ensemble of feature vectors of length N correspondingto different vowels to estimate the N **x** N vowel covariance matrix C_v. Similarly we estimate the consonant covariance matrix C_c. We wish to find a W that **max**imizes the ratio of the variance of VI to that of CI after the transformation. If the density functions of d_v and d_c are assumed to be normally distributed, then their covariance matrices after transformation **are** given by

$$\widehat{C_v} = W^T C_v W \qquad (3)$$
$$\widehat{C_c} = W^T C_c W$$

A simple measure of the variance or the 'scatter' is the determinant of the covariance matrix [4]. Thus, the criterion function to be maximized is given by

$$J(W) = \frac{|\widehat{C_v}|}{|\widehat{C_c}|} = \frac{|W^T C_v W|}{|W^T C_c W|} \qquad (4)$$

The columns of the optimum W are obtained **as** generalized eigenvectorsfor vowels (GEVV), corresponding to the largest eigenvalues in

$$C_v w_i^{(v)} = \lambda_i C_c w_i^{(v)} \qquad (5)$$

Similarly, we obtain generalized eigenvectors for consonants (GEVC) as

$$C_c w_i^{(c)} = \lambda_i C_v w_i^{(c)} \qquad (6)$$

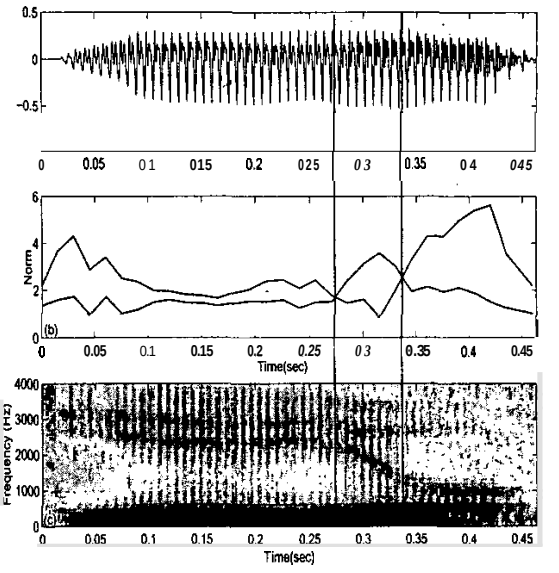

Figure **2.** (a) Speech signal /eyo/. (b) Its segmentation into vowel (/e/ and /o/) and consonant (/y/) regions, using both the vowel and consonant norm-contours. (c) Spectrogram of /eyo/.

Evaluation **of** Norm-Contours

The GEVV and GEVCs are obtained by solving equations **5** and **6.** The test signal is divided into overlapping frames and the feature vector x_k corresponding to the k^{th} frame is obtained using LP-Cepstral coefficients (LPCC) or Mel-Cepstral coefficients (MCC). We evaluate the norm-contours as follows.

$$N_v(k) = \sum_{i=1}^{M} (w_i^{(v)})^T x_k \& N_c(k) = \sum_{i=1}^{M} (w_i^{(c)})^T x_k \qquad (7)$$

N_v and N_c give the norm-contours from V and C subspaces. Norm of the projection of the feature vectors derived from the test basic-unit **on** GEVV and GEVC givethenorm-contours. One of them represents the vowel information and the other, the consonant information. The resulting norm contours obtained for a test signal cross each other at the beginning and at the end of consonant region of a given test basic-unit. The segmentation points are the ones where $N_v(k) \simeq N_c(k)$. We found that optimum results were obtained when M = 3.

3. STATISTICAL TESTING FOR SEGMENTATION'

Here, we view the basic-unit segmentation as classifying frames of a given test basic-unit into one of two classes: vowels and consonants. This indeed is a hypothesis testing (two-class) problem and it requires probability models for vowel & consonant anda robust thresholdas well. Tipping and Bishop [2] proposed PPCA, to emphasize the advantages of associating a probability model with principal *component analysis* (PCA), rather than considering the algorithmic perspective of

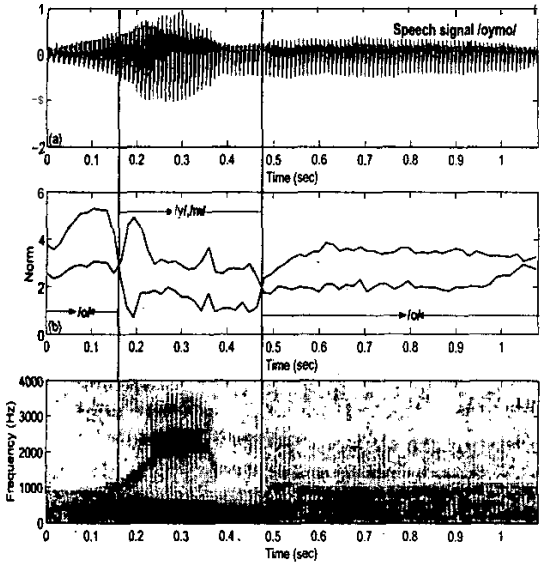

Figure 3. (a) Speech signal /oymo/. (b) Its segmentation into vowel (/o/) and consonant (/y/ and /m/) regions, using both the vowel and consonant norm-contours. (c) Spectrogram of /oymo/.

determining principal components. Using PPCA approach [2], we generate probability model for V and C.

Motivation for Hypothesis based Segmentation

In the section 2, we saw the performance of *oriented principal component analysis* based basic-unit segmentation. Although the performance of OPCA is good, the crossovers (threshold) are not robust. So, segmentation becomes difficult when there are multiple crossovers. Figure 4(a) shows a basic-unit /aumai/, and Fig. 4(b) shows the norm-contours obtained after projecting the basic-unit in Fig. 4(a) on V and C subspaces. One can see from Fig. 4(b) that there are multiple crossovers, making it difficult to choose an appropriate threshold for segmentation. Here, classical decision theory [5] can be used to find the robust threshold. Along with the crossover information, decision regarding vowel or consonant class is used to obtain correct segmentation points by eliminating the false crossovers. The class decision is shown in Fig. 4(d).

Probability Models using PPCA

Unlike in PCA, PPCA gives a probability density for the data through a latent variable model. A latent variable model relates a d-dimensional observed data vector $\mathbf{t}$ to a q-dimensional ($q < d$) latent vector $\mathbf{x}$ by defining a noise model and a prior on the distribution on the latent variables in the form,

$$t = Ux + \mu + \epsilon, \tag{8}$$

where ϵ is an $\mathbf{x}$-independent noise process and $\mathbf{U}$ is the $(d \times q)$ generative matrix. It is common that the prior distribution of the latent variables is a sample Gaussian distribution $\mathbf{x} \sim$

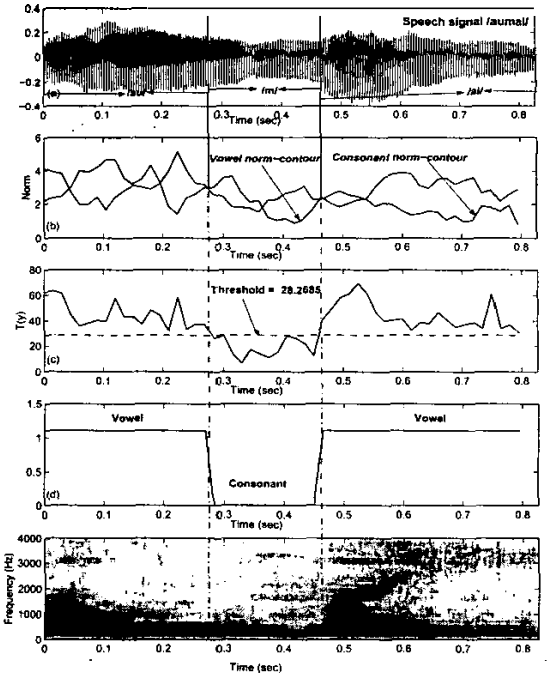

Figure 4. Segmentation of basic-unit using OPCA and PPCA. (a) Test basic-unit /aumai/. (b)Vowel and consonant-norm contours after projecting the test basic-unit into V & C subspaces. (c) Test statistics $T(y)$, along with the threshold for a given P_{FA} 1 $^{-4}$. (d) V & C decision after the class testing. (e) Spectrogram of /aumai/.

$N(\mathbf{0}, \mathbf{I})$ over the latent space. The noise model may also be Gaussian with $\epsilon \sim N(\mathbf{0}, \psi)$. The model is restricted to having a non-zero mean by the parameter μ. These aspects mean that the observed data vectors are normally distributed by $\mathbf{t} \sim N(\mu, \mathbf{C})$, where the model covariance is given by

$$C = \psi + UU^T \tag{9}$$

If $\psi = \sigma^2\mathbf{I}$, the latent variable model is called PPCA [2]. In this terminology, conventional PCA is recovered when $\sigma^2 \rightarrow 0$. One of the ways of estimating the parameters of these latent variable models is by the Expectation-Maximization algorithm. Tipping and Bishop [6] have recently formulated PCA within a maximum-likelihood framework based on a specific form of Gaussian latent variable model. They also showed that with $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{U}^T$, the only stationary points occur when

$$U = W_q(\Lambda_q - \sigma^2 I)^{\frac{1}{2}} R. \tag{10}$$

Here $\mathbf{W}_q$ has q eigenvectors of $\mathbf{S}$ (sample covariance matrix, given by $\mathbf{S} = \frac{1}{N}\sum_n^N (t_n - \mu)(t_n - \mu)^T$) as column vectors, Λ_q is a diagonal matrix with the corresponding eigenvalues, and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix. Tipping and Bishop prove that when $\mathbf{W}_q$ contains the *principal* eigenvectors of $\mathbf{S}$, global maximum of the likelihood occurs. So with $\mathbf{U}$ in Equation 10, the latent variable model defines

a mapping from the latent space into the ***principal subspace*** of the observed data. The maximum likelihood estimator for noise variance σ^2 is given by the average variance lost for each discarded dimension, and can be formulated as

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j. \tag{11}$$

To sum up, a Probabilistic PCA is obtained by finding the q principal eigenvectors and eigenvalues of the sample covariance matrix **S**, while Gaussian distribution with sample mean μ and covariance matrix $\sigma^2\mathbf{I}+\mathbf{U}\mathbf{U}^T$, gives the density model.

Vowel-Consonant Segmentation as Hypothesis Testing

To bring the segmentation problem into a Hypothesis testing framework, we need to have a probability model for ***vowels*** and ***consonants***. This can be accomplished by calculating **U**, using equation 10 and plugging **U** into the covariance model given by $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{U}^T$. Let us call the probability model for ***vowels*** and ***consonants*** as C_{pv} and C_{pc}, respectively. We have,

$$\begin{aligned} C_{pv} &= \sigma_v^2 I + U_v U_v^T \\ C_{pc} &= \sigma_c^2 I + U_c U_c^T \end{aligned} \tag{12}$$

We use GEVV ($\mathbf{W}_q^{(v)}$), **GEVC** ($\mathbf{W}_q^{(c)}$) and the corresponding eigenvalue matrices ($\Lambda_q^{(v)}$ & $\Lambda_q^{(c)}$) to obtain the parameters U_v, U_c, σ_v and σ_c using equations 10 and 11. These parameters are in turn used to generate probability model C_{pv} and C_{pc} using equation 12. Assuming the **V** and *C* features to be zero mean Gaussian random processes with covariance matrices C_{pv} and C_{pc}, the segmentation problem is to distinguish between the hypotheses

$$x \sim \begin{cases} N(0, C_{pc}) & \textit{under } H_0 \\ N(0, C_{pv}) & \textit{under } H_1 \end{cases}$$

Here, x is a zero mean ***test basic-unit***. A NP detector decides H_1 if the likelihood ratio exceeds a threshold or if

$$Lh(x) = \frac{p(x;H_1)}{p(x;H_0)} > \gamma$$

where,

$$p(x;H_1) = \frac{1}{\sqrt{(2\pi)^d\,|C_{pv}|}}\exp[-\frac{1}{2}x^T C_{pv}^{-1}x],$$

$$p(x;H_0) = \frac{1}{\sqrt{(2\pi)^d\,|C_{pc}|}}\exp[-\frac{1}{2}x^T C_{pc}^{-1}x],$$

$$Lh(x) = \frac{\frac{1}{\sqrt{(2\pi)^d|C_{pv}|}}\exp[-\frac{1}{2}x^T C_{pv}^{-1}x]}{\frac{1}{\sqrt{(2\pi)^d|C_{pc}|}}\exp[-\frac{1}{2}x^T C_{pc}^{-1}x]}$$

the log-likelihood ratio (LLR) becomes

$$lh(x) = \{(x^T C_{pc}^{-1}x) - (x^T C_{pv}^{-1}x])\} > 2\ln\ [\gamma\frac{\sqrt{|C_{pv}|}}{\sqrt{|C_{pc}|}}].$$

Hence, we decide H_1 if

$$T(x) = x^T(C_{pc}^{-1} - C_{pv}^{-1})x > \gamma' \quad [\ln 2 \quad \gamma\frac{\sqrt{|C_{pv}|}}{\sqrt{|C_{pc}|}}] \tag{13}$$

Evaluation of the Threshold

By simplifying the test statistics $T(x)$ (eq. 13) into a standard distribution form, then the threshold (γ') can be evaluated [5]. Let $A = C_{pc}^{-1}$ and $B = C_{pv}^{-1}$, equation 13 becomes

$$T(x) = x^T(A-B)x > \gamma'. \tag{14}$$

Define $G = A^{-1}B$. Let the eigenvalue decomposition of G, be

$$GV = \Lambda V,$$

such that

$$\begin{aligned} V^T A V &= \Lambda_1 \\ V^T B V &= \Lambda_2 \end{aligned} \tag{15}$$

where, Λ, Λ_1 and Λ_2 are diagonal matrices. Writing equation 15 in terms of A and B, we have

$$\begin{aligned} A &= V\Lambda_1 V^T \\ B &= V\Lambda_2 V^T \end{aligned} \tag{16}$$

Substituting eq. 16 into **14**, we have

$$T(x) = x^T V[\Lambda_1 - \Lambda_2]V^T x. \tag{17}$$

Let $y = V^T x$. Then equation 17 becomes

$$T(y) = y^T\tilde{\Lambda}y \tag{18}$$

where, $\tilde{\Lambda} = \Lambda_1 - \Lambda_2$. We know that **x** is zero mean. Therefore y is also zero mean, i.e., $E\{y\} = 0$ and the covariance of y is

$$E\{yy^T\} = \begin{cases} V^T E\{xx^T\}V \\ V^T C_{xx} V \end{cases}$$

If **x** is a ***vowel***, then $\mathbf{C}_{xx} = C_{pv}$ and $V^T C_{pv} V = \Lambda_1^{-1}$. Similarly, If **x** is a ***consonant***, then $\mathbf{C}_{xx} = C_{pc}$ and $V^T C_{pc} V = \Lambda_2^{-1}$. So, we can write equation 18 as

$$T(y) = \sum_{n=1}^{N}\frac{y^2(n)}{\eta^2(n)} \tag{19}$$

where $\left\{\frac{1}{\eta^2(1)}, \frac{1}{\eta^2(2)}, \ldots, \frac{1}{\eta^2(n)}\right\}$ are diagonal elements of $\tilde{\Lambda}$. Because y is **i.i.d.**, T(y) is χ^2 (Chi-square) distributed and the detector turns out to be an ***energy detector***. The required threshold, γ', for segmentation is computed by using an optimization algorithm for the fixed ***probability of false alarm***, P_{FA} [5].

$$P_{FA} = \int_{\gamma'}^{\infty}\int_{-\infty}^{\infty}\prod_{n=1}^{N}\frac{1}{\sqrt{1-2j\alpha_n\omega}}\exp(-j\omega t)\frac{d\omega}{2\pi}dt \tag{20}$$

$$P_D = \int_{\gamma'}^{\infty}\int_{-\infty}^{\infty}\prod_{n=1}^{N}\exp(-j\omega t)\frac{d\omega}{2\pi}dt. \tag{21}$$

where,

$$\alpha = \frac{1}{\eta^2(n)}.$$

Now, the decision (segmentation) rule is

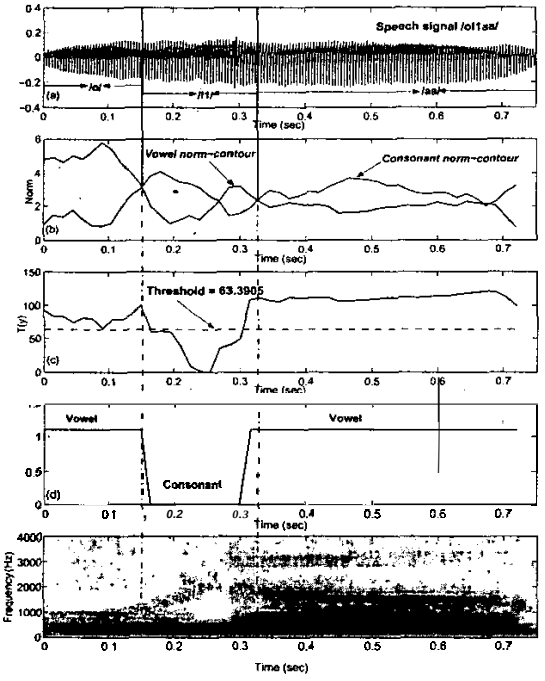

Figure 5. Segmentation of basic-unit using OPCA and PPCA. (a) Test basic-unit /ollaa/. (b) Vowel and consonant-norm contours after the projection of test basic-unit into V & C subspaces. (c) Test statistics $T(y)$, along with the threshold for a given $P_{FA} = 1^{-4}$. (d) V & C decision after the class testing. (e) Spectrogram of ollad.

Decide **Vowel** ($\mathbf{H}_1$) if

$$T(y) = \sum_{n=1}^{N} \frac{y^2(n)}{\eta^2(n)} > \gamma^{\text{II}}$$

otherwise decide Consonant ($\mathbf{H}_0$).

4. Results and Discussion

Speech segmentation experiments were conducted on a Kannada speech database spoken by a female volunteer. GEVV's and GEVC's were obtained from a Tamil speech database spoken by a male volunteer using the method discussed in section 2. Feature vectors were obtained for each frame of a test basic-unit. Duration of each frame of speech was 30 ms, with an overlap of 20 ms between successive frames. Each frame of speech was Hamming windowed and processed to yield a 13-dimensional feature vector. The feature vectors used were MCC and LPCC. We have seen in Fig. 1(b) that energy based segmentation fails to identify the co-articulated consonant region. On the other hand, in Fig. 2, the same consonant region has been correctly identified using subspace-based method. Figure 2(c) also displays the spectrogram of the basic-unit /eyo/. Here, spectrogram shows clear second formant transition from the frontal vowel /e/ to the back vowel /o/ and the transition region corresponds to the consonant /y/. Basic-unit considered in Fig. 3 is VCCV and both C's are *glide* and a *nasal* (lyl and /m/).

Probability models for V and C were obtained using equations 10, 11 and 12. Here, the observed feature vector **t**, and its dimension is **13**, the, latent variable **x** has dimension of 3. ie., $d = 13$ and $q = 3$. Dimension of the latent variable **x** has been chosen based on the dimensions of GEVV and GEVC. The multiple crossovers present some uncertainty in classifying the basic-units into vowel and consonant parts. By employing classical decision theory, we can remove this uncertainty by choosing a threshold on a statistical basis. In plot 4(d), the basic-unit is seen to he clearly segmented. In Fig. 4(b), there are multiple crossovers and using the class decision approach shown in Fig. 4(d), we can eliminate false crossovers. We used fixed $P_{FA} = 10^{-4}$ to calculate the threshold, γ^{II}. Figure 4(e) shows the spectrogram of /aumai/. Figure 5(a) shows the co-articulated basic-unit /ollad. In Fig. 5(b) we can see the multiple cross overs of the norm-contours. Fig. 5(c) shows the test statistics along with the threshold and V & C decision is shown in Fig. 5(d). Figure 5(e) shows the spectrogram of /ollaa/.

5. Conclusion

We have presented subspace and hypothesis based methods for co-articulated basic-unit segmentation. The first method uses crossovers of norm-contours for segmentation. Norm-contours give a measure of energy in the projected subspaces. There may be an ambiguity when there are multiple crossovers. In the second method this ambiguity is resolved by finding a statistical threshold. The test statistic using NP criterion turns out to be an energy detector as in the first method but without any ambiguity in segmentation. PPCA method can be further extended to classify within the vowel or consonant segments in a basic-unit using mixture-PPCA [6].

References

[1] G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and P. Prathibha, "A complete text-to-speech synthesis system in Tamil," ***IEEE workshop on Speech Synthesis,*** 2002.

[2] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysers," ***Journal of the Royal Statistical Society,*** vol. 61, no. 3, pp. 611-622, 1999.

[3] Jan P. van Hemert. "Automatic segmentation of speech:' *IEEE Trans. Signal Proc.*, vol. 39. no. 4. pp. 1008-1012. 1991.

[4] R. Duda and P. Hart ***Pattern Classification and Scene Analysis,*** New York: Wiley, 1973.

[5] Steven M. Kay, ***Fundamentals of Statistical Signal Processing : Detection Theory***, Prentice-Hall, Inc. A Simon & Schuster, New Jersy 07458, 1998.

[6] M.E. Tipping and C. M. Bishop, "Mixture of probabilistie principal component analysers," ***Neural Computation,*** vol. 11, no. 2, pp. 443-482, 1999.