ABSTRACT

In our (knowledge-based) synthesis system [1], we use single instances of basic-units, which are polyphones such as CV, VC, VCV. VCCV and VCCCV, where C stands for consonant and V for vowel. These basic-units are recorded in an isolated manner from a speaker and not from continuous speech or carrier-words. Modification of the pitch, amplitude and duration of basic-units is required in our speech synthesis system [1] to ensure that the overall characteristics of the concatenated units match with the true characteristics of the target word or sentence. Duration modification is carried out on the vowel parts of the basic-unit leaving the consonant portion in the basic-unit intact. Thus, we need to segment these polyphones into consonant and vowel parts. When the consonant present in any basic-unit is a plosive or fricative, the energy-based method is good enough to segment the vowel and consonant parts. However, this method fails when there is a co-articulation between the vowel and the consonant. We propose the use of oriented principal component analysis (OPCA) to segment the co-articulated units. The test feature vectors (LPC-Cepstrum & Mel-Cepstrum) are projected on the consonant and vowel subspaces. Each of these subspaces are represented by generalized eigenvectors obtained by applying OPCA on the training feature vectors. Our approach successfully segments co-articulated basic-units.

1. INTRODUCTION

For the purpose of 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 segmentation 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 [2] 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 the segment as a spectrally stable part of a signal. In [2], a segment is defined as a number of consecutive frames whose spectra are similar. Here, the normalized correlation $C_{ij}$ between the LPC smoothed log-amplitude spectra of the $i^{th}$ and $j^{th}$ frames in the utterance is used as the measure of similarity. If the correlation equals 1, then the $i^{th}$ and $j^{th}$ frames are identical. A heuristically chosen threshold value defines the beginning and end of the significant part of the correlation curve. This threshold defines the extent to which the correlation is allowed to drop within one segment. Explicit segmentation methods split up the utterance into segments that are defined by phonetic transcription. In general, explicit methods have the disadvantage that the reference patterns need to be generated before the method can be used [2]. The results obtained in [2] show that the explicit scheme is inaccurate as compared to the implicit one. Finally, a combination of the two methods is proposed in [2].

2. FEATURE TRANSFORMATION

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-Noise ratio (SNR) is maximum. Such a decomposition can be arrived at by representing VI and CI by training vectors obtained using manual segmentation of the basic-units extracted from the data base collected by us for Tamil synthesis system. 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 $W$ that maps the original feature vectors $x$ on to $Z$.

$$Z = W^T x$$

(1)
where $x$ is an $n$-dimensional vector, $\hat{x}$ is an $m$-dimensional vector, $m \leq n$, and $W$ is an $n \times m$ matrix with $m$ linearly independent columns. Let $d_0$ and $d_c$ represent the training vectors containing VJ and CI, respectively, in the original feature space. The covariance matrices for these training vectors can be written as

$$E_v = E[(d_v - \bar{d}_v)(d_v - \bar{d}_v)^T]$$

$$E_c = E[(d_c - \bar{d}_c)(d_c - \bar{d}_c)^T]$$

where $\bar{d}_v$ and $\bar{d}_c$ represent the means of $d_v$ and $d_c$ respectively. We wish to find a $W$ that maximizes the ratio of the variance of $V$ to that of $C$ after the transformation. If the density functions of $J_v$ and $d_c$ are assumed to be normally distributed, then their covariance matrices after transformation are given by

$$\tilde{E}_v = W^T C_v W$$

$$\tilde{E}_c = W^T C_c W$$

A simple measure of the variance or the "scatter" is the determinant of the covariance matrix [1]. Thus, the criterion function to be maximized is given by

$$J(W) = \frac{\det(\tilde{E}_v)}{\det(\tilde{E}_c)} = \frac{|W^T C_v W|}{|W^T C_c W|}$$

The columns of the optimum $W$ are obtained as generalized eigenvectors for vowels (GEVV), corresponding to the largest eigenvalues in $C_v u^{(v)}_i = \lambda_i C_v u^{(v)}_i$.

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

$$C_c w^{(c)}_i = \lambda_i C_c w^{(c)}_i$$

In [4], Mallawah et al. introduced a SNR measure, defined as the ratio of these variances when the original feature vectors are projected on to $u^{(v)}_i$.

$$\gamma = \frac{u^{(v)}_i u^{(v)}_i^T C_v w^{(v)}_i}{u^{(v)}_i u^{(v)}_i^T C_v u^{(v)}_i}$$

If the first $m$ eigenvectors are used, the SNR becomes

$$\gamma = \frac{\text{trace}(W^T C_v W)}{\text{trace}(W^T C_c W)}$$

The SNR of the original feature vectors can be calculated from Eq. 8 by making $W$ an identity matrix. Figs. 2(a) and (b) show the SNR before and after transformation for LPC-Cepstrum (LPCC) and Mel-Cepstral Coefficients (MCC), respectively. From the figures, it can be seen that the SNR of the transformed feature vectors is substantially higher than that of the original feature vectors. Since the eigenvalues are ordered as a decreasing sequence, the SNR after transformation $\gamma = \sum_{i=1}^{m} \lambda_i$ decreases with increase in the dimension of the feature vectors.

3. FEATURE TRANSFORMATION AS FILTER BANK

It has been shown in [6] that block orthogonal transforms, such as DFT, DCT and KL, can be interpreted as uniform perfect reconstruction filter banks. We will use this relationship to arrive at an understanding of the feature transformation process in the log spectral domain. In a uniform $N$-channel filter bank, the input signal is decomposed through a set of filters $H_0(z), ..., H_{N-1}(z)$. Each subband signal is then decimated by $N$. The reverse process at the synthesis bank reconstructs the signal. Now consider a $N \times N$ transformation matrix $W$ such that

$$\hat{x} = W^T x$$

Let $w_k$ be the $k^{th}$ column of $W$. The $k^{th}$ component of $\hat{x}$ is the inner product of $x$ with $w_k$. That is

$$\hat{x}_k = w_k^T x = \sum_{i=0}^{N-1} x(i) w_{ik}$$

where $w_{ik}$ is the $i^{th}$ component of $w_k$. This summation can be interpreted as filtering of $x(n)$ advanced by $N - 1$ samples [6]:

$$\hat{x}_k(n) = \sum_{i=0}^{N-1} x(n + N - 1 - i) h_k(i)$$

where

$$h_k(i) = w_{N-1-i-k}$$

are the impulse response coefficients of the filter $H_k(z)$. The sequences $\hat{x}_k(n)$ are obtained by downsampling the sequences $\hat{x}_k(n)$. 

Fig. 3. Variation of SNR for GEVV as a function of feature dimension. (a) for LPC-cepstrum. (b) for Mel-cepstrum.
by \( N \). In the transform domain, the convolution in Eq.11 can be written as
\[
\hat{X}_k(e^{j\omega}) = X(e^{j\omega})H_k(e^{j\omega})
\] (13)
We know that LPC-Cepstrum is the inverse Fourier transform of the logarithm of the all-pole LPC spectrum. Thus, when the input features \( x(n) \) are the LPC-cepstral vectors, \( X(e^{j\omega}) \) represents the log spectrum. Therefore, according to Eq.13, the transformation process can be seen as a multiplication of the log spectrum by the frequency response of the filters \( H_k(z) \). Since the transformation matrix is derived from the data (through generalized eigenvalue decomposition), the frequency response of the filters corresponding to the largest eigenvalues indicates the relative importance of different frequency bands for basic unit segmentation. The first four principal component filters for vowels and consonants are shown in Figs 3 and 4, respectively. From Figs 3 and 4, we see that the frequency response of the first four principal filters indicate the relative importance of mid-frequency region of the speech spectrum for vowels and low and high frequency regions for consonants. In [5], it is demonstrated that the low and high frequency regions of the speech spectrum convey more speaker information than the mid-frequency regions. Combining our results with those of [5], we can say that the speech information in the mid-frequency region of the speech spectrum corresponds to vowel information. Similarly, speaker information in the low and high frequency regions corresponds to consonant information. From this, we can conclude that the consonants have more speaker information.

4. OPCA METHOD FOR V-C SEGMENTATION
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 LPCC or MCC. We evaluate the norm-contours as follows.
\[
N_v(k) = \sum_{i=1}^{M} (w_i^{(v)})^T x_k \quad \& \quad N_c(k) = \sum_{i=1}^{M} (w_i^{(c)})^T x_k
\] (14)
\( N_v \) and \( N_c \) give the norm-contours from V and C subspaces. Norms of the projections of the feature vectors (derived from the test basic-unit) on GEVV and GEVC give the norm-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 end of consonant region of a given test basic-unit. The segmentation points are the ones where \( N_v(k) = N_c(k) \). We found that optimum results were obtained when \( M=2 \).

5. RESULTS AND DISCUSSION
Basic-unit segmentation experiments were conducted on database collected from a female volunteer for our Kannada synthesis system. The database has isolated utterances of VC, CV, VCV, VCCV and VCCCV. GEVV and GEVC were obtained from our Tamil database recorded from a male volunteer. In the case of all the Indian languages, diphthongs have distinct symbols and are grouped with and called as vowels. Our notations and analysis follow this system. 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. For obtaining MCC, the Mel-scale was simulated using a set of 24 triangular filters and for LPCC, a 12th order LPC analysis was performed after preemphasis with \( a = 0.95 \). We have seen in Fig 1(b) that energy based segmentation fails to identify consonant regions in co-articulated basic-units. Results of our
segmentation algorithm are shown in Figs. 5 to 8, along with the respective spectrograms. These test basic-units cover most of the classes of speech. For the same basic-unit shown in Fig. 1(b), consonant region has been correctly identified (see Fig. 5) using our algorithm. Here, the consonant is a glide, ‘/ys/’. In the basic-unit shown in Fig. 6, vowel ‘/ai/’ is diphtong and the consonant is nasal. In Fig. 7, consonant is a liquid, ‘/M/’, Fig. 8 shows the segmentation of a VCCV basic-unit, where the consonants ‘/y/ and /m/’ are glide and a nasal respectively. The classes of speech considered here are difficult to segment because they possess high co-articulation in combination with vowels. We found that the segmentation performance with MCC features is better than that with LPCC features. So, the results shown here have been obtained using MCC features. The figures show that the transition of the second formant frequency clearly matches with the duration between norm crossovers in all the spectrograms. In [7], it is shown that data-driven Principal component analysis (PCA) approach, though significantly easier to implement than Linear discriminant analysis (LDA), gives a comparable performance as LDA. So, we considered data-driven OPCa approach for basic-unit segmentation.

We applied our technique on 600 co-articulated units involving consonants such as ‘/y/’/ ‘/f/’/ ‘/l/’/ ‘/h/’ and ‘/N/’, in combination with vowels. When tested on the basic-units (distinct from the training set) from the same male speaker (Tamil mother tongue), we obtain correct segmentation of the V-C boundary in more than 85% of the cases. Further, when the same is applied on units from a speaker of opposite sex, speaking a different language (Kannada), resulted in correct segmentation in 80% of the co-articulated units tested.

6. CONCLUSION

We have proposed a segmentation algorithm that effectively handles co-articulated basic-units. The filter bank interpretation of the feature transformation throws light on the relative significance of different frequency bands for vowels and consonants. We found that mid-frequency region in the speech spectrum has more vowel information; low and high frequency regions have higher consonant information. The vowel norm-contour clearly shows a dip in the consonant region and is not affected by speaker variations in a test basic-unit. Consonant norm-contour is, to some extent, affected by speaker variations in a test basic-unit. This can be explained by the presence of speaker information in low and high frequency regions of speech spectrum as shown in [5].

7. REFERENCES


Fig. 6. (a) Speech signal ‘/aimi/’. (b) Its segmentation into vowel ‘/ai/’ and ‘/m/’ and consonant ‘/n/’ regions, using vowel and consonant subspaces. (c) Spectrogram of ‘/aimi/’.

Fig. 7. (a) Speech signal ‘/aui/’. (b) Its segmentation into vowel ‘/au/’ and ‘/u/’ and consonant ‘/l/’ regions, using vowel and consonant subspaces. (c) Spectrogram of ‘/aui/’.

Fig. 8. (a) Speech signal ‘/aimol/’. (b) Its segmentation into vowel ‘/ai/’ and consonant ‘/m/’ and ‘/l/’ regions, using our algorithm. (c) Spectrogram of ‘/aimol/’.