Segmentation of specific speech signals from multi-dialog environment using SVM and wavelet

T.K. Truong a,*, Chien-Chang Lin a, Shi-Huang Chen b

a Department of Information Engineering, I-Shou University, Kaohsiung County 840, Taiwan, ROC
b Department of Computer Science and Information Engineering, Shu-Te University, Kaohsiung County 824, Taiwan, ROC

Available online 27 March 2007

Abstract

In this paper, a novel multi-speaker segmentation technique is presented. This technique makes use of wavelets and support vector machines (SVMs) to segment specific speakers’ speech signals from multi-dialog environments. The proposed method first applies wavelets to determine the acoustical features such as subband power and pitch information from a given multi-dialog speech data. Then the multi-speaker segmentation of the given multi-dialog speech data can be accomplished by the use of a bottom–up SVM over these acoustical features and additional parameters, such as frequency cepstral coefficients. A public audio database, Aurora-2, is used to evaluate the performances of the proposed method. Experimental results show that the accuracy of multi-speaker segmentation is 100% achieved in the combination of two speakers. And the segmental accuracy can achieve at least 94.12% and 85.93% for 4-speaker and 8-speaker conditions, respectively.

Keywords: Speech segmentation; Wavelets; Support vector machine (SVM)

1. Introduction

Both speech classification and segmentation provide useful information for speech retrieval as well as content understanding (Campbell et al., 2006). In general, a speaker classification problem involving several target speakers can be manipulated in two steps. First, characterize each input sound clip by a small set of parameters obtained from various feature extraction techniques. Next, apply classification algorithms to these parameters to accomplish the multi-classification task. The efficacy of a speaker classification system depends on the ability to capture proper acoustical features and to accurately classify each feature set to its own speaker (Lin et al., 2005). In this paper, these two tasks will be taking into account in the proposed speech segmentation method.

Recently, Li (2000) presented a method for content-based audio classification and retrieval. The features selected by Li are the combinations of mel-frequency cepstral features (MFCC) and other perceptual features including brightness, bandwidth, subband energies, and the shape of the frequency spectrum features. Li also proposed a new pattern classification method called the nearest feature line (NFL), and a comparison result with the nearest neighbor (NN) classification was included. His experiments show that NFL method performs better than nearest neighbor (NN), and nearest center (NC). However, in the same paper, in a series of 198 classifications of sounds selected from Muscle Fish database, 40 of them are not correct.

Many researches have proposed to improve the speech classification and segmentation by employing different features and methods. The most common approach for speech segmentation is to modify a hidden Markov model (HMM)
based speech recognition system (Ajmera et al., 2003; Wu et al., 2005). However, due to the limitation of speech recognizer, the HMM-based speech segmentation methods perform not that well in a multi-dialog environment.

Several statistical techniques, such as neural networks or support vector machines (SVMs), can also be used for the audio classification and segmentation. Among these techniques, the SVMs (Vapnik, 1998; Kecman, 2001), proposed by Vapnik, have been regarded as a new learning algorithm for various applications, such as audio classification (Clarkson and Moreno, 1999; Guo and Li, 2003; Melgani and Bruzzone, 2004) and pattern recognition (Schwenker, 2000; Fenglei and Bingxi, 2001; Ganapathiraju et al., 2004). By using SVMs instead of NFL method, Guo and Li (2003) improved significantly the previous work (Li, 2000) on classification performance. The achievement they made is to reduce the number of classification errors from 40 to 16 in the same number of 198 classifications.

Recently, wavelets and support vector machines have been successfully used in audio classification and categorization (Lin et al., 2005). Based on (Lin et al., 2005), this paper further applies wavelets and SVM to multi-speaker segmentation that can segment specific speech signals from multi-dialog environments. The proposed multi-speaker classification method consists of three steps. The first step is the voice activity detection (VAD) of the input speech signal using the AMR VAD OPT1 (ETSI EN 301 708 V7.1.1) algorithm. Next, derive both perceptual features and cepstral coefficients from speakers’ speeches by the use of Fourier transform and wavelets, such as normalized subband power, pitch information, frequency cepstral coefficients (FCC), etc. Finally, apply SVMs to the multi-speaker segmentation with pre-selected parameters. Experimental results show that the accuracy of multi-speaker segmentation is 100% achieved in all cases of the combination of two speakers. And the segmental accuracy reaches at least 94.12% and 85.93% for 4-speaker and 8-speaker cases, respectively.

The rest of this paper is organized as follows. In Section 2, the detailed descriptions of the feature extraction as well as wavelets are given. Then, the SVM for two-class classification problem is described in Section 3. Section 4 illustrates various experimental results under multi-dialog environments. Finally, conclusions are given in Section 5.

2. Feature extraction

Based on (Lin et al., 2005), a $(12 + 2L)$-dimensional feature vector, proposed in this paper for speech segmentation, is constructed from perceptual features and frequency cepstral coefficients. These features are extracted from each input sound and can be applied to facilitate further segmentation and classification. The detailed pre-processing and feature extraction processes are described in the following subsections.

2.1. Pre-processing

The original audio sounds in (Aurora-2) are down-sampled at 8000 Hz with 16-bit resolution. Each sound is divided into frames. The frame length is 256 samples (32 ms) with 192 samples (75%) overlap between adjacent frames. Each frame is processed via a pre-emphasizing filter that is defined as

$$s_n' = s_n - 0.96 \times s_{n-1}, \quad n = 1, \ldots, 255,$$

(1)

where $s_n$ is the $n$th sample of the frame $s$ and $s'_0 = s_0$. Then, the pre-emphasized frame is Hamming-windowed by

$$s^h = s' \ast h_i, \quad i = 0, \ldots, 255,$$

(2)

where $h_i = 0.54 - 0.46 \times \cos(2\pi i/255)$. The pre-processed frame will be detected as a non-silent frame for feature extraction if the total power is large, i.e.,

$$\sum_{i=0}^{255} (s^h_i)^2 > 400^2,$$

(3)

where 400 is an experience value (Li, 2000; Guo and Li, 2003; Lin et al., 2005).

As mentioned in (Holmberg et al., 2006), the short-term time constant lying between 80 to 300 ms has better recognition performance. In this paper, a frame block is therefore designed to contain 32 frames and then abandoning the silent frames such that a frame block has non-silent frames between 10 (104 ms) and 32 (280 ms) frames, dropping the frame block containing less than 10 frames.

2.2. Feature extraction from non-silent frames

In many previous works, wavelet transform (Mallat, 1998; Burrus et al., 1998) and the Fourier transform are two popular methods that map audio signal from time domain to feature domain. (Chen and Wang, 2002) and (Hsieh et al., 2002) proved that a three-level wavelet transform gives better performances for audio signals with sampling rate 8000 Hz. Hence, this paper combines the Fourier transform and the Daubechies length-8 orthogonal wavelet (Mallat, 1998; Burrus et al., 1998) to increase the capability to capture proper audio features.

There are totally $6 + L$ features, listed in Table 1, derived from FFT coefficients $f(u)$ and wavelet coefficients of each non-silent frame $s^h_i$. The detailed extraction process of each feature is given in the following:

<table>
<thead>
<tr>
<th>Table 1</th>
<th>List of extracted features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of feature</td>
<td>Feature</td>
</tr>
<tr>
<td><strong>Perceptual feature</strong></td>
<td>Subband power $P_j$</td>
</tr>
<tr>
<td></td>
<td>Pitch frequency $f_p$</td>
</tr>
<tr>
<td></td>
<td>Brightness $a_0$</td>
</tr>
<tr>
<td></td>
<td>Bandwidth $B$</td>
</tr>
<tr>
<td><strong>Cepstral coefficient</strong></td>
<td>Frequency cepstral coefficient (FCC) $c_n$</td>
</tr>
</tbody>
</table>
(1) Subband power $P_j$: Let $\omega$ be the half sampling frequency. Three sections of subband power calculated in wavelet domain are the subband intervals $[0, \omega/8]$, $[\omega/8, \omega/4]$, and $[\omega/4, \omega/2]$, corresponding to the approximation and detail coefficients $a_d(k)$, $d_0(k)$, and $d_1(k)$, respectively, of a given audio sound $a_3(k)$. The subband power is calculated by $P_j = \sum z^2_j(k)$, where $z^j(k)$ is the corresponding approximation or detail coefficients of subband $j$.

(2) Pitch frequency $f_p$. In this paper, the noise-robust pitch detection method using wavelet transform with aliasing compensation is used to extract the pitch frequency (Chen and Wang, 2002).

(3) Brightness $\omega_b$: The brightness is the frequency centroid and is computed as $\omega_b = \int_0^\infty u[F(u)]^2 du / \int_0^\infty [F(u)]^2 du$.

(4) Bandwidth $B$: It is the square root of the power-weighted average of the squared difference between the spectral components and the frequency centroid, i.e., $B = \sqrt{\int_0^\infty (u - \omega_b)^2 [F(u)]^2 du / \int_0^\infty [F(u)]^2 du}$.

(5) Frequency cepstral coefficient (FCC): The $L$-order coefficients are calculated as $c_n = \sqrt{2/256 \sum_{u=0}^{255} (\log_{10} F(u)) \cos(n(u - 0.5)\pi/256)}$, where $n = 1, 2, \ldots, L$.

For each frame block, after the $6 + L$ features extracted from each non-silent frame, the mean and the standard deviation of each of the $6 + L$ features are computed to yield a $(12 + 2L)$-dimensional feature vector.

2.3. Feature vector formation and normalization

Select randomly from the Aurora database two sets of speech files, one for training and one for testing. The detailed description is given in Section 4. For each speech in the set for training, several feature vectors can be obtained. Assume that $n_T$ is the number of the obtained feature vectors. The training set $T$ is then defined to be the $n_T \times (12 + 2L)$ array with row vectors being these feature vectors. Let $T(i,j)$ denote the $(i,j)$-position of $T$. Use this array $T$ to construct another $n_T \times (12 + 2L)$ array $T'$ whose $(i,j)$-position $T'(i,j)$ is defined to be $T'(i,j) = T(i,j) - \mu_j$, where $\mu_j = \sum T(i,j)/n_T$ is the average value of column $j$. Next, normalize $T'$ by computing $T''(i,j) = T'(i,j)/m_j$, where $m_j$ is the maximum of the absolute value of elements in column $j$. Thus, each feature will have similar weights after the normalization process. Finally, the values $\mu_j$ and $m_j$ computed above are also used in the testing-set part, i.e., $E(i,j) = E(i,j) - \mu_j$ and $E''(i,j) = E'(i,j)/m_j$. These normalized perceptual cepstral coefficients, $T''(i,j)$ and $E''(i,j)$, are used for training and testing in SVM, respectively.

3. Support vector machines (SVMs)

The support vector machines (SVMs) (Vapnik, 1998; Kecman, 2001) proposed by Vapnik have been recently evaluated as popular tools for learning from given data. The reason is that SVMs are more effective than the traditional pattern recognition approaches based on the combination of a feature selection procedure and a conventional classifier (Li, 2000; Zhang and Kuo, 2001; Lu et al., 2002). They are also much more effective than other conventional nonparametric classifiers (e.g., the RBF neural networks, nearest neighbor (NN), nearest center (NC), and the $k$-NN classifier (Schwenker, 2000)) in terms of classification accuracy and stability to parameter setting. In order to separate given points into two target classes, SVMs use a known kernel function with a parameter, upper-bound $C$, to define a hyperplane. An improved SVM called soft-margin SVM which can tolerate minor misclassifications (Vapnik, 1998; Clarkson and Moreno, 1999; Kecman, 2001) is considered to be more suitable for classification and therefore is used in this paper.

Suppose a training set $S = \{(x_i, y_i)\}_{i=1}^n \subseteq (X \times Y)^t$ is given, where $x_i \in X \subseteq R^r$ ($r$-dimensional real space) and $y_i \in Y = \{1, -1\}$, are the input vector and the target variable, respectively. Choose a kernel function $K(x_i, x_j) = \phi(x_i)(\phi(x_j))$, where $\phi$ maps the input space $X$ into another high dimensional feature space $F$. The given nonlinearly separable samples $S$ may be, therefore, linearly separated in $F$, as shown in Fig. 1.

Many hyperplanes, denoted by $(w, b)$, where $w \in R^r$, $b \in R$, consisting of all $x \in X$ satisfying $(\langle w, x \rangle + b) > 0$, can achieve the separation purpose. One can find a hyperplane that maximizes the margin (the minimal distance from the hyperplane to the given points). The problem can thus be formed as

$$
\begin{align*}
\text{minimize} & \quad ||w||^2 / 2 \\
\text{subject to} & \quad y_i(\langle w, x_i \rangle + b) \geq 1
\end{align*}
$$

(4)

The saddle point of Lagrange function is the solution of the optimization problem. Let $C$ be the upper-bound of the Lagrange multipliers $\alpha_i$ and then (4) can be formulated as

$$
L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle
$$

(5)

with constraints $\sum_{i=1}^n \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$. Suppose that $x^*_i$ maximize (5), then the optimal discriminant

![Fig. 1. A feature map can simplify the classification task.](image)
problems. The slack variables $\xi_i$ in plane. The soft-margin SVM is used to solve non-separable classification problems. Therefore, the ERBF kernel function and the bottom–up binary tree scheme has the best performance in terms of classification accuracy and time complexity. There are two common kernel functions for the nonlinear feature mapping. (1) exponential radial basis function (ERBF) $K(x, \tilde{x}) = \exp(-|x - \tilde{x}|^2 / 2\sigma^2)$, and (2) Gaussian kernel function $K(x, \tilde{x}) = \exp(-|x - \tilde{x}|^2 / 2\sigma^2)$, where parameter $\sigma^2$ is the variance of the Gaussian function. Many classification problems are always separable in feature space and able to obtain better accuracy by using ERBF kernel function than by Gaussian kernel function (Guo and Li, 2003; Lin et al., 2005). In addition to achieve multiclass classification by using a two-class SVM, as shown in Table 2, four commonly used schemes are (1) one-against-one, (2) one-against-all, (3) top–down binary tree, and (4) bottom–up binary tree. Many previous works (Guo and Li, 2003; Meldani and Bruzzone, 2004; Lin et al., 2005) show that the bottom–up binary tree scheme has the best performance in terms of classification accuracy and time complexity. Therefore, the ERBF kernel function and the bottom–up binary tree scheme are used in this paper.

### 4. Experimental results

In the following experiments, a public audio database named Aurora-2 is used to evaluate classification performance. This clean TIDigits database consists of sequences.

#### Table 2
Comparison among four schemes

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Total training complexity</th>
<th>Total testing times</th>
<th>Result combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-against-one</td>
<td>$[e(c - 1)/2] \cdot T(n,n)$</td>
<td>$d(c - 1)/2$</td>
<td>Vote/statistic</td>
</tr>
<tr>
<td>One-against-all</td>
<td>$cT(n, c - 1)n$</td>
<td>$e - 1$</td>
<td>Vote/statistic</td>
</tr>
<tr>
<td>Top–down binary tree requirement</td>
<td>$\sum_{i=1}^{\log_2 c - 1} T(cn/2^i, cn/2^i)$</td>
<td>$\log_2 c$</td>
<td>No</td>
</tr>
<tr>
<td>Bottom–up binary tree requirement</td>
<td>$[e(c - 1)/2] \cdot T(n,n)$</td>
<td>$e - 1$</td>
<td>No</td>
</tr>
</tbody>
</table>

Each of the $e$ classes contains $n$ members, where $e$ is a number of power of 2.

$T(U, V)$: The training complexity with $U$ vectors in the plus-class and $V$ vectors in the minus-class.
of at most seven arbitrary digits spoken by American
speakers. There are seven kinds of experiments correspond-
ing to 2 to 8 randomly chosen speakers. In the experiment
 corresponding to a certain number of speakers, all possible
combinations about the gender of speakers are considered.
For example, if the number of speakers is two, there are
three cases: namely, two female, one male and one female,
and two males. In each case, the training set is consisting of
all the speeches of the selected speakers in the “clean1”
folder of Aurora-2. On the other hand, the testing-set con-
spists of a single file created by composing of the same
selected speakers’ speeches in the folder “clean2”.

Every training or testing vector consists of a (12+2L-
dimensional feature vector that is extracted from each
frame block of sounds and then used in soft-margin SVMs
with bottom–up tree scheme to complete evaluations. The
accuracy of the ERBF kernel in SVMs is clearly better than
that of Gaussian kernel in most settings (Lin et al., 2005).
An interesting observation is that the ERBF kernel func-
tion will be more stable and accurate when both the
upper-bound C and variance $\sigma^2$ are at least 30 and 60,
respectively. This suggests us to choose ERBF as the kernel
function and $C = 30$, $\sigma^2 = 60$ to be the upper-bound and
variance, respectively, while FCC level $L$ varies from 1 to
99.

The steps to evaluate the classification accuracy used in
this paper are as follows:

Step 1: Segment a testing file into $k$ sentences by using
AMR VAD OPT1 algorithm (ETSI EN 301 708
V7.1.1).

Step 2: Make use of bottom–up SVMs to classify each of
the $k$ sentences.

(1) Partition the sentence into frame blocks.

(2) Recognize to which class indexed by the
speakers each frame block belongs.

(3) For each class, count the total number of
non-silent frames from those frame blocks
assigned to the class.

(4) Set the sentence to the class obtaining the
maximal number; if there are more than one
candidate then choose the first one.

Step 3: The accuracy of classification is defined to be the
number $(\text{Acc}/k) \times 100\%$, where $\text{Acc}$ is the number
of correctly classified sentences in Step 2.

There are 42 cases can be obtained from combinations
of 2 to 8 speakers. For each combination, $A_m$ is computed
as the maximum of accuracy and $L_m$ indicates which FCC
level $L$ the first $A_m$ happens. As shown in Table 3, when the
number of speakers is two or three, the accuracy is 100% in
every except one case, where the label “xM-yF” means that
the number of male and female speakers are $x$ and $y$,
respectively. As the number of speakers increases, the accu-
cracy decreases. Even the case of eight speakers, the accu-
cracy is still nearly 86%. The reason is that the speakers
who belong to the same gender have similar features.
One also observed that the closer the numbers of females
and males are, the better the accuracy is. This observation
will be pointed out in Fig. 3 which shows the accuracies
based on different numbers of speakers and different ratios
of male to female.

As shown in Fig. 4, the segmental accuracies among 6
cases, namely 0M-2F, 1M-1F, 2M-0F, 0M-8F, 4M-4F,
and 8M-0F, are illustrated in various FCC level $L$. It
reveals that the proposed method using ERBF kernel func-
tion performs the best with $35 \leq L \leq 99$ for the pre-
selected values $C = 30$ and $\sigma^2 = 60$. This proves once again

![Fig. 3. The accuracy in various ratios of male and female.](image-url)
that the accuracy of speech segmentation is more stable and accurate when the ERBF kernel function and the pre-selected upper-bound $C = 30$ and variance $\sigma^2 = 60$ are used in SVM.

5. Conclusions

The processing time is important in the pattern segmentation and classification. Starting with the method proposed by the authors (Lin et al., 2005), the proposed method in this paper reduces the size of feature set and this yields the reduction of processing time. In addition, by partitioning the speech into frame blocks, the proposed method is more suitable for real-time processing systems. Based on the experimental results shown in Fig. 4, the segmental experiments are more stable and accurate in the case of bottom-up SVM with upper-bound $C = 30$ when the ERBF kernel function has variance $\sigma^2 = 60$. These figures are therefore been chosen to be the default setting for the proposed method in the future research.

Acknowledgements

The authors are extremely grateful to Prof. Yaotsu Chang and Prof. Yukon Chang who critically read the entire paper and made numerous excellent comments.

References


ETSI EN 301 708 V7.1.1, Voice Activity Detector (VAD) for Adaptive Multi-Rate.


