Automatic Metric-based Speech Segmentation for Broadcast News via Principal Component Analysis

Jeih-weih Hung[1,2], *Hsin-min Wang*[1] *and Lin-shan Lee*[1,2]
[1]Institute of Information Science, Academia Sinica
[2]Dept of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
phone : 886-2-2788-3799 ext 1507, fax : 886-2-2782-4814
email : jwhung@iis.sinica.edu.tw

ABSTRACT

In this paper, we proposed an algorithm used to improve the performance of the metric-based segmentation techniques, by which the segmentation points are found at maxima of a distance measured between two contiguous windows shifted along the stream of speech features. In our proposed method, the PCA processes are first performed on the speech features to obtain more robust features, and then the above metric-based segmentation was applied on the PCA-derived features to decide the segmentation points. Experiment results show that our proposed method can efficiently improve the detection rates of the segmentation points up to 7% while the false alarm rates remain unchanged.

I. Introduction

Accurate segmentation of audio signal streams is a key process to improve the performance for recognition, transcription and retrieval of audio signals. Various schemes have been proposed to perform the speech segmentation automatically. According to [1], these approaches were roughly divided into three classes. : decoder-based splitting, model-based splitting and metric-based splitting. The metric-based splitting method has been found very useful [1,2] and very flexible, since no or little information about the speech signal is needed a priori to decide the segmentation points. In this kind of method, an acoustic distance measure is defined to evaluate the similarity between two contiguous windows shifted along the speech signal. The locations of distance peaks in the audio signal are detected as the candidates for the segmentation points. The final segmentation points are then chosen by some heuristic thresholds and criterions.

As we know, the performance of the metric-based segmentation approach depends not only on the distance measures used, but also on the feature representation of the audio signals. More discriminative or robust features are helpful, especially when the speech signal is corrupted by channel distortion or additive noise. The Principal Component Analysis (PCA) [3] has been widely used in various problems to obtain more effective representation of features. In this approach, the eigenvectors of the covariance matrix for the original features corresponding to the largest k eigenvalues are taken as the basis of the eigenspace. The original features are then mapped onto this eigenspace to obtain the PCA-derived features. In this paper, an improved metric-based segmentation approach using principal component analysis is presented, and significant improvements were obtained in segmenting the broadcast news.

The remainder of the paper is organized into 5 sections. In section 2, the Principal Component Analysis (PCA) is briefly reviewed. Then section 3 introduces the metric-based segmentation and several popular used distance measures. In section 4, we describe our proposed segmentation algorithm. Section 5 then presents some preliminary experimental results using the proposed approach. Finally, a short conclusion is given in section 6.

II. Principal Component Analysis (PCA)

It is well known that Principal Component Analysis (PCA) is widely applied for the data analysis and dimensionality reduction. Briefly speaking, for a zero-mean random vector of dimension N, PCA tries to find k ($k \le N$) orthonormal vectors so that when the inner product (a random variable) of the random vector and the individual orthonormal vector will have largest variance. The k ($k \le N$) orthonormal vectors are in fact the eigenvectors of the covariance matrix for the random vector corresponding to the largest k eigenvalues. Let's show it briefly as follows.

Let $\boldsymbol{x} \in R^N$ and $E(\boldsymbol{x})=\boldsymbol{0}$, at first we wish to find $\boldsymbol{e}_1 \in R^N$ and $|\boldsymbol{e}_1|^2 = 1$, such that $\mathrm{Var}(\boldsymbol{e}_1^T \boldsymbol{x}) = \boldsymbol{e}_1^T E(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{e}_1$ is maximum. Using the method of Lagrange multipliers, we wish to maximize

$J(\boldsymbol{e}_1)=\boldsymbol{e}_1^T E(\boldsymbol{x}\boldsymbol{x}^T)\,\boldsymbol{e}_1-\lambda_1(|\boldsymbol{e}_1|^2-1)$, (1)

where λ_1 is a Lagrange multiplier. After differentiating $J(\boldsymbol{e}_1)$ w.r.t $\boldsymbol{e}_1$ and set it to zero, we obtain $E(\boldsymbol{x}\boldsymbol{x}^T)\,\boldsymbol{e}_1=\lambda_1\,\boldsymbol{e}_1$, so λ_1 is the eigenvalue of $E(\boldsymbol{x}\boldsymbol{x}^T)$ and $\boldsymbol{e}_1$ is the corresponding eigenvector. Furthermore, since $\mathrm{Var}(\boldsymbol{e}_1^T\boldsymbol{x})=\lambda_1$ is to be maximized, λ_1 is chosen as the largest eigenvalue of $E(\boldsymbol{x}\boldsymbol{x}^T)$ and $\boldsymbol{e}_1$ is the corresponding eigenvector.

Secondly, we wish to find $\boldsymbol{e}_2\in R^N$ and $|\boldsymbol{e}_2|^2=1$, such that $\mathrm{Var}(\boldsymbol{e}_1^T\boldsymbol{x})=\boldsymbol{e}_1^T E(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{e}_1$ is maximum and $\boldsymbol{e}_2^T\boldsymbol{e}_1=0$. Again using the method of Lagrange multipliers, we wish to maximize

$J(\boldsymbol{e}_2)=\boldsymbol{e}_2^T E(\boldsymbol{x}\boldsymbol{x}^T)\,\boldsymbol{e}_2-\lambda_2(|\boldsymbol{e}_2|^2-1)-p(\boldsymbol{e}_2^T\boldsymbol{e}_1-0)$, (2)

where λ_2 and p are Lagrange multipliers. After differentiating $J(\boldsymbol{e}_2)$ w.r.t $\boldsymbol{e}_2$ and set it to zero, again we obtain $E(\boldsymbol{x}\boldsymbol{x}^T)\,\boldsymbol{e}_2=\lambda_2\,\boldsymbol{e}_2$, so λ_2 is the eigenvalue of $E(\boldsymbol{x}\boldsymbol{x}^T)$ and $\boldsymbol{e}_2$ is the corresponding eigenvector. Furthermore, since $\mathrm{Var}(\boldsymbol{e}_2^T\boldsymbol{x})=\lambda_2$ is to be maximized and $\boldsymbol{e}_2^T\boldsymbol{e}_1=0$, λ_2 is chosen as the second largest eigenvalue of $E(\boldsymbol{x}\boldsymbol{x}^T)$ and $\boldsymbol{e}_2$ is the corresponding eigenvector.

Following the above steps, k orthonormal vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\ldots, \boldsymbol{e}_k$ can be obtained by choosing the k eigenvectors corresponding to the largest k eigenvalues of $E(\boldsymbol{x}\boldsymbol{x}^T)$, which is the covariance matrix of $\boldsymbol{x}$ since $E(\boldsymbol{x})=\boldsymbol{0}$. The k orthonormal vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\ldots, \boldsymbol{e}_k$ form a basis of a subspace of R^N, and when $\boldsymbol{x}$ is projected to this subspace, it can be proven that the resulted random vector $\boldsymbol{y}$ is "closest" to $\boldsymbol{x}$ (the mean square error of $\boldsymbol{y}-\boldsymbol{x}$ is minimum) over the projection of $\boldsymbol{x}$ upon any other subspace of R^N spanned by k orthnormal vectors. The PCA-derived random vector $\boldsymbol{z}\in R^k$ from $\boldsymbol{x}\,R^N$ is referred to as the vectors of k projection coefficients of x upon the k eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\ldots, \boldsymbol{e}_k$. That is, $\boldsymbol{z}=[\,\boldsymbol{e}_1\,\boldsymbol{e}_2\,\ldots\ldots\,\boldsymbol{e}_k]^T\boldsymbol{x}$. The PCA-derived vector is often called the most "expressive" features since its components are of largest variances, which amounts to extracting the most parts of randomness of the original random vector.

III. Metric-based Segmentation

For the metric-based segmentation approaches, the speech signal is first encoded in terms of acoustic feature vectors. Then a dissimilarity (or called distance) value is measured between two consecutive parts (called window) of the acoustic features. Since it's complicated to directly measure the dissimilarity between two collections of vectors, both windows of features are often individually first modeled parametrically such as single or multiple Gaussian distributions, and then there are many distance measures between two parametric statistical models can be applied here. We list several distance measures [4] d^{12} between two multivariate Gaussian distributions, $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$, as follows, which we will used in our later experiments.

Kullback-Leibler distance:

$$d_{KL}^{12}=\tfrac{1}{2}(\mu_1-\mu_2)^T\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)(\mu_1-\mu_2)+\tfrac{1}{2}tr\left(\Sigma_1^{-1}\Sigma_2+\Sigma_2^{-1}\Sigma_1-2\mathbf{I}\right) \quad (3)$$

Mahalanobis distance:

$$d_{MAH}^{12}=\tfrac{1}{2}(\mu_1-\mu_2)^T(\Sigma_1\Sigma_2)^{-1}(\mu_1-\mu_2) \quad (4)$$

Bhattacharyya distance:

$$d_{BHA}^{12}=\tfrac{1}{4}(\mu_1-\mu_2)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)+\tfrac{1}{2}\log\frac{|\Sigma_1+\Sigma_2|}{2\sqrt{|\Sigma_1\Sigma_2|}} \quad (5)$$

Obviously a high distance value indicates a possible acoustic change, whereas a low value show that two portions of signal correspond to the same acoustic environment. Such distance measure is continually computed between two contiguous windows shifted along the speech features stream to form a distance curve. This distance curve is often low-pass filtered and then the local peaks of the filtered distance curve are then detected as the candidates for the segmentation points. Based on these candidates, different criterions and heuristics are applied to decide the final segmentation points. The overall procedure of metric-based segmentation is depicted in the following Figure 1.

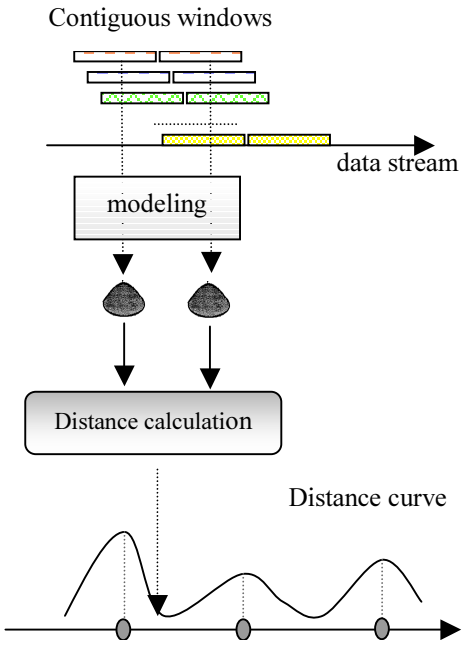

Figure 1. The procedures of metric-based segmentation

IV. Proposed Algorithm

From the previous statements, the performance of the metric-based segmentation is mainly influenced by several factors, and they are the feature representation of speech signal, the used distance measure, and the criterion to decide the final splitting points. Other factors, such as the size of window chosen, the filter applied for the distance curve and so on, also more or less affect its performance. In this paper we will pay attention to the used feature representation of speech signal to be segmented. It's well known that the current widely used speech features, such as MFCC or LPC, are easily corrupted by additive noise or channel distortion, and thus more discriminative or expressive presentations of speech to be segmented become necessary. In this paper, we suggested that the PCA processes are performed on the original speech feature vectors, and then the PCA-derived feature vectors are used for metric-based segmentation.

As stated in section III, the metric-based segmentation continuously computes the distance of two contiguous windows shifted along the speech features. In this paper we proposed two approaches for extracting the PCA-derived features for the metric-based segmentation..

In the first approach, two eigenspaces for the two contiguous windows of original speech features of dimension *N* are constructed separately. Then the original speech features of each window are mapped respectively onto the corresponding eigenspaces to form the PCA-derived features. Stated mathematically, consider the two windows of feature vectors $\mathbf{X}^L = \{\boldsymbol{x}_i^L\}_{i=1,\ldots,n}$, and $\mathbf{X}^R = \{\boldsymbol{x}_i^R\}_{i=1,\ldots,n}$. Assume $\mathbf{X}^L$ and $\mathbf{X}^R$ are the samples of two zero mean random vectors $\boldsymbol{x}^L$ and $\boldsymbol{x}^R$ respectively, then the covariance matrices of $\boldsymbol{x}^L$ and $\boldsymbol{x}^R$ can be approximated as $\Sigma_{x^L} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i^L)^2$ and $\Sigma_{x^R} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i^R)^2$ respectively. Let $\mathbf{V}^L = [\boldsymbol{e}_1^L \quad \boldsymbol{e}_2^L \quad \cdots \quad \boldsymbol{e}_k^L]$, whose columns are the eigenvectors corresponding to the largest *k* eigenvalues of Σ_{x^L}, and let $\mathbf{V}^R = [\boldsymbol{e}_1^R \quad \boldsymbol{e}_2^R \quad \cdots \quad \boldsymbol{e}_k^R]$, whose columns are the eigenvectors of the largest *k* eigenvalues of Σ_{x^R}. Then the PCA-derived features of the two windows can be presented as $\mathbf{Y}^L = \{\boldsymbol{y}_i^L = \mathbf{V}^L(\mathbf{V}^L)^T \boldsymbol{x}_i^L\}_{i=1,\ldots,n}$ and $\mathbf{Y}^R = \{\boldsymbol{y}_i^R = \mathbf{V}^R(\mathbf{V}^R)^T \boldsymbol{x}_i^R\}_{i=1,\ldots,n}$, respectively. After the PCA-derived features of the two windows are obtained, they are modeled and the distance between the two models is computed as the normal processes of the metric-based segmentation.

In the second approach, a single common eigenspace is derived using all the speech features of the two contiguous windows, and the speech features of the two contiguous windows are then mapped onto the common eigenspace to form the PCA-derived features. Stated mathematically, consider the two windows of feature vectors $\mathbf{X}^L = \{\boldsymbol{x}_i^L\}_{i=1,\ldots,n}$, and $\mathbf{X}^R = \{\boldsymbol{x}_i^R\}_{i=1,\ldots,n}$. Assume all $\mathbf{X}^L$ and $\mathbf{X}^R$ are the samples of a zero mean random vector $\boldsymbol{x}$, then the covariance matrix of $\boldsymbol{x}$ can be approximated as $\Sigma_x = \frac{1}{2n-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i^L)^2 + \sum_{i=1}^{n}(\boldsymbol{x}_i^R)^2\right)$. Let $\mathbf{V} = [\boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \cdots \quad \boldsymbol{e}_k]$, whose columns are the eigenvectors corresponding to the largest *k* eigenvalues of Σ_x. Then the PCA-derived features of the two windows can be presented as $\mathbf{Y}^L = \{\boldsymbol{y}_i^L = \mathbf{V}(\mathbf{V})^T \boldsymbol{x}_i^L\}_{i=1,\ldots,n}$ and $\mathbf{Y}^R = \{\boldsymbol{y}_i^R = \mathbf{V}(\mathbf{V})^T \boldsymbol{x}_i^R\}_{i=1,\ldots,n}$. After the PCA-derived features of the two windows are obtained, they are modeled and the distance between the two models is computed as the normal processes of the metric-based segmentation.

V. Experimental Results

Our proposed approaches were tested with the broadcast news reports provided by CTS (Chinese Television System in Taiwan). The experimental data consist of 29 sections of news reports and is about 46 minutes long. The speech signals were recorded with a sampling rate of 8K Hz. These data were first hand-segmented according to the speaker changes, environmental changes and silence periods, and 1023 segmentation points were decided. A 32ms Hamming window shifted with 16ms steps and a pre-emphasis factor of 0.95 were used to evaluate 15 mel-frequency cepstral coefficients (MFCCs) as the original speech features. The window size was chosen as 2 seconds and the window shift was 100 ms. Various distance measures listed in Section III between the two contiguous windows were evaluated while Gaussian densities were assumed for each window. The local maxima within an interval of 1.5 seconds were chosen as the candidates of segmentation points. Two approaches stated in Section IV are used to obtain the PCA-derived features based on the original MFCC features for segmentation. Furthermore, in our experiments we use different number of eigenvectors for the eigenspace to see its influences on the performance. In the following, Tables 1-3 show the detection rates of various approaches where different distance measures are applied.

features	# of eigenvec-tors	Approach 1 : two eigenspaces for two windows	Approach 2 : one eigenspace for two windows
PCA-derived features from MFCC	1.	94.04%	90.62%
	3	91.40%	87.98%
	5	89.35%	89.44%
	7	89.54%	91.20%
	9	89.15%	90.42%
	11	89.25%	89.64%
	13	88.86%	89.64%
MFCC		88.47%	

Table 1. The detection rates of two approaches when Mahalanobis distance is used

features	# of eigenvec-tors	Approach 1 : two eigenspaces for two windows	Approach 2 : one eigenspace for two windows
PCA-derived features from MFCC	1.	94.62%	90.81%
	3	92.96%	88.86%
	5	90.52%	88.17%
	7	90.52%	87.49%
	9	89.74%	88.66%
	11	89.35%	88.66%
	13	87.98%	88.47%
MFCC		87.68%	

Table 2. The detection rates of two approaches when Kullback-Leibler distance is used

features	# of eigenvec-tors	Approach 1 : two eigenspaces for two windows	Approach 2 : one eigenspace for two windows
PCA-derived features from MFCC	1.	91.20%	90.52%
	3	91.98%	88.56%
	5	90.62%	88.47%
	7	89.83%	87.00%
	9	90.13%	88.66%
	11	89.15%	88.56%
	13	88.17%	88.37%
MFCC		87.88%	

Table 3. The detection rates of two approaches when Bhattacharyya distance is used

First of all, from Tables 1-3 we see that when MFCCs were used as the features for segmentation, three different distance measures give comparable results. However, it is obvious that both approaches proposed in this paper give significant improvements in detection rates, especially when Mahalanobis distance or Kullback-Leibler distance is used. Secondly, we also see that in most cases the first approach, where two eigenspaces for the two contiguous windows are used, outperforms the second approach, where only a common eigenspace is used for the two contiguous windows. This can be explained briefly as follows. While two contiguous windows of speech features belong to two different acoustic environments, using only one eigenspace for both windows of features to project on will decrease the difference (distance) between them. However, while two contiguous windows of speech features belong to the same acoustic environment, using one common eigenspace is more suitable than using two difference eigenspaces, since the latter seems to increase the false alarm rate, which doesn't happen in our experiments though. Finally, we find that in most cases both approaches using fewer eigenvectors for constructing the eigenspace give rise to better performance, which is expected since computation complexity of the eigenspace construction and the mapping of the features can be efficiently reduced.

VI. Conclusion

In this paper, we proposed two approaches to apply the PCA on the speech features in order to obtain more robust features for metric-based segmentation. Improvements in detection rates of segmentation points show the effectiveness of our proposed approaches. Such improvements are believed to increase the performance of following speech signal processing and speech recognition.

References

[1] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", Proc. of DARPA Broadcast News Transcription and Understanding Workshop, 1998

[2] Laurent Couvreur and Jean-Marc Boite, "Speaker Tracking in Broadcast Audio Material in the Framework of the THISL Project", Proc. of European Speech Communication Association (ESCA) European Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio, 1999

[3] Fukunaga, Keinosuke, "Introduction to Statistical Pattern Recognition", 2nd ed. Boston : Academic Press, c1990.

[4] M. Basseville, "Distance Measures for Signal Processing and Pattern Recognition", Signal Processing, Vol. 18, 1989