interspeech 2008 incorporating SST 08

Robust Speaker Change Detection Using Kernel-Gaussian Model

Jie Gao, Xiang Zhang, Qingwei Zhao,Yonghong Yan

Institute of Acoustics, Chinese Academy of Sciences, Beijing, P.R.China

{jgao, xzhang, qzhao, yonghong.yan}@hccl.ioa.ac.cn

Abstract

This paper introduces and evaluates a novel approach for unsupervised speaker change detection. In many unsupervised speaker change detection algorithms, each audio segment is typically modeled with a multivariate single Gaussian density, where it is assumed that the distribution of the speech features of the segment is Gaussian. However, this assumption is too strong in many cases. Therefore, this paper presents an alternative to the single Gaussian model: Gaussian model in reproducing kernel Hilbert space (RKHS) or Kernel-Gaussian model (KGM). KGM first projects speech features into RKHS via a nonlinear mapping. Then it models the features in RKHS with a Gaussian density. The mapping procedure enables KGM to capture nonlinear structure of speech features. An implementation of KGM is proposed and evaluated. Experiments on different datasets show that better results are achieved by KGM compared to the single Gaussian model.

Index Terms: Speaker change detection, kernel method, Gaussian distribution, generalized likelihood ratio

1. Introduction

Speaker change detection (SCD) aims at determining the time indices of speaker change points in an input speech stream and dividing the stream into speaker homogenous regions. SCD is an important preprocessing step of many applications including audio data mining, speech transcription, speaker recognition and speaker tracking.

The SCD problem has been well examined by researchers in recent years. Previous approaches to automatic speech segmentation can be classified into two categories: supervised and unsupervised. Supervised approaches include decoder-guided and model-based segmentation. In the decoder-guided approaches the input speech stream is first decoded using a recognition system, then change points are obtained by the detected silence locations [1]. In model-based approaches, initial models are created for a closed set of acoustic classes using training data. Then the input speech stream is classified by maximum likelihood selection using these models [2]. The boundaries between classes become potential change points.

Unsupervised SCD avoids requirements of the prior knowledge and the training stage in supervised approaches. It typically involves sliding an analysis window through the audio stream, building parametric models of two neighboring windows and measuring dissimilarity between the two models. Potential change points are found around local optima (maxima or minima) of the resulting distance graph [3, 4, 5, 6, 7]. Due to the simplicity of implementation, investigations are performed on many aspects of unsupervised SCD. Effects of different window sliding schemes and window sizes are investigated in [3, 5]. A variety of features, feature combination [8] and feature transformation [4] is exploited, aiming at finding more effective representations of speech segments. Different metric functions have also been investigated, including the BIC related [3, 5], GLR [9], Kullback-Leibler distance [5], and their combination [5].

Work also has been done for the modeling aspect in unsupervised SCD. Many systems use the single Gaussian model. This model implicitly assumes the linear distribution of speech features for each segment. However, previous study in the field of speaker recognition reports the distribution of speech features may be of arbitrary shape density [10]. Therefore, Gaussian Mixture Models (GMMs) are deployed in some unsupervised SCD methods [6, 7]. However, finding alternative models to effectively represent speech features for SCD is still an open issue.

The main focus of this paper is to introduce and evaluate an alternative to single Gaussian model for unsupervised SCD algorithms, namely a Gaussian model in reproducing kernel Hilbert space (RKHS) or Kernel-Gaussian model (KGM). KGM tries to model the nonlinearity speech features through a kernel method. KGM firstly represents the speech features as samples in a high dimensional feature space (RKHS) via nonlinear mapping. Then a Gaussian model is fitted for the samples in RKHS. Since a nonlinear function is used in the mapping, the derived distribution can account for the complex structure of the speech features. An implementation of KGM is also proposed in this work, which is equivalent to performing probabilistic Kernel-PCA (PKPCA) on features. Therefore, this implementation potentially makes speech features more discriminative for speaker change detection.

The remainder of this paper is organized as follows. In the next section, the motivation of using KGM as an alternative to a single Gaussian model is presented. In Section 3, the concept of KGM and the proposed implementation is detailed. The SCD algorithm used in the evaluation experiments is also described. Section 4 presents the experiments and the results. Finally, Section 5 concludes this paper and discusses future work.

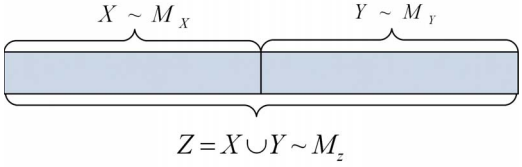
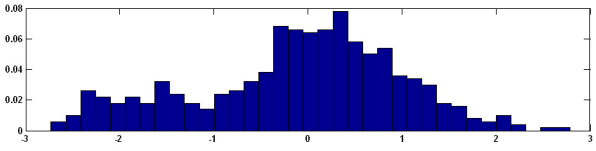
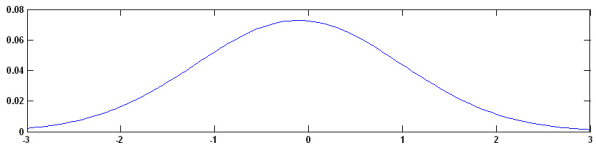


Figure 1: Two neighboring speech segments X, Y and their union Z with corresponding fitted models M_x, M_y and M_z

Accepted after peer review of full paper



(a) Histogram of a single Mel frequency cepstral coefficient from a 5-second segment by a male speaker



(b) Gaussian model by maximum likelihood estimation

Figure 2: Nonlinear distribution of features of speech segments

2. Motivation of Kernel-Gaussian Model

As shown in Figure 1, a procedure used by many unsupervised SCD algorithms is to fit two neighboring segments X, Y and their union Z with certain parametric models M_x , M_y and M_z. Then a metric function $d(M_x, M_y, M_z)$ based on these models is evaluated. A Gaussian model is commonly adopted because of its simplicity. However, this makes the assumption that the distribution of speech features is Gaussian, which may be not the case. An example is shown in Figure 2. Figure 2(a) gives the histogram of a single Mel frequency cepstral coefficient from a 5-second segment by a male speaker. Nonlinear structure of speech features is observed. Figure 2(b) gives the Gaussian model of the segment estimated by maximum likelihood estimation (MLE) and shows that the Gaussian model fails to model this distribution effectively. Hence a model with better modeling capability is desired for speaker change detection.

In addition, speech features are easily corrupted by channel distortion, additive noise or other factors, which reduce their discriminability for SCD. Effective feature transformations that produce more robust feature representations will be helpful to SCD.

The Gaussian model in reproducing kernel Hilbert space (KGM) manipulates data through a kernel method, which enables it to model the nonlinear structure or high-order statistical information of data. Hence, KGM is a suitable alternative to a single Gaussian model. Besides, a proper implementation of KGM simultaneously performs an effective feature transformation and provides a more robust feature representation.

3. Models in unsupervised speaker change detection

In this section, we describe the algorithm framework of the unsupervised SCD used in this paper. We also detail the concept of KGM and the proposed implementation.

Although many distance measures can incorporate the proposed model readily, the GLR distance is chosen here [9]. The GLR distance for X and Y in Figure 1 is defined as:

$$R = \frac{L(Z|M_z)}{L(X|M_x)L(Y|M_y)} \quad (1)$$

where $L(X|M_x)$ denotes the likelihood of the vector X given the model M_x and N_x is the number of samples in X. In the case of the single Gaussian model $M \sim N(\mu, \Sigma)$, a closed-form expression of GLR is derived as:

$$R_{Gaussian} = \frac{N_z}{2} log|\Sigma_z| - \frac{N_x}{2} log|\Sigma_x| - \frac{N_y}{2} log|\Sigma_y| \quad (2)$$

The segmentation algorithm used is similar to the first stage of the DISTBIC procedure in [5]. Given N feature vectors $X = \{x_1, x_2..x_N\}$ of an audio segment, modeling algorithms of the single Gaussian model and KGM are detailed below.

3.1. Single Gaussian Model

X is typically modeled by a single multivariate Gaussian distribution $N(\mu, \Sigma)$. The mean μ and covariance Σ are estimated by MLE, given as:

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i = XS \qquad S = \frac{1}{N} \times \mathbf{1} \quad (3)$$

$$\hat{\Sigma}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})(x_i - \hat{\mu}_{MLE})^T \quad (4)$$

$$= XJJ^TX^T$$

$$J = \frac{1}{N^{1/2}}(I_N - S \cdot \mathbf{1}^T) \quad (5)$$

where S is the weighting vector for sample mean, J is a $N \times N$ matrix called *centering matrix* and $\mathbf{1}$ is a $N \times 1$ vector of 1's. The GLR distance can be evaluated by (2) with the fitted models.

3.2. Kernel-Gaussian Model

The concept of KGM is illustrated in Figure 3. Given a sequence of feature vectors $X = \{x_1, x_2..x_N\}$ extracted from a speech segment, our proposed approach works in three steps. First, it maps X in the original data space Ω (named input space) to training samples $\Phi(X) = \{\phi_1, \phi_2..\phi_N\}$ in a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ using a nonlinear mapping function Φ. Second, a Gaussian distribution (KGM) is fitted for mapped data samples $\Phi(X)$ in RKHS. Finally, a metric function based on KGM is evaluated. The essence of KGM is to use the single Gaussian model within the nonlinear space RKHS, which enables it to capture complex structure of data in the input space. The key implementation issue in building KGM is the proper approximation of its sufficient statistics in RKHS: mean and covariance matrix.

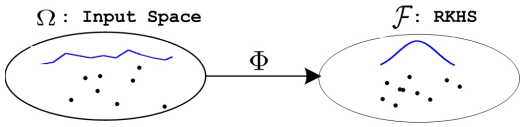


Figure 3: Concept of Kernel-Gaussian Model

KGM handles speech data through a kernel method. It does not manipulate samples in RKHS ($\Phi(X)$) directly. Instead, it represents them in the so-called *Gram Matrix:* $\mathbf{K} = \Phi\Phi^T$. The element $k_{i,j}$ of the Gram matrix is the dot product of the i^{th} sample ϕ_i and the j^{th} sample ϕ_j. This can be easily computed using the kernel trick : $k_{i,j} = K(x_i, x_j) = \phi_i \cdot \phi_j$ [12], where $K(\cdot,\cdot)$ is a pre-specified kernel function. Further, the centered version of Gram Matrix is denoted as *Centered Gram Matrix:* $\mathbf{K}^* = J\Phi\Phi^T J^T$, where J is the centering matrix given by (5).

With these variables, we propose to approximate the mean and the covariance matrix of KGM following the method in [11].

The mean in RKHS is estimated by MLE, given by

$$\hat{\mu}_{RKHS} = \frac{1}{N}\sum_{i=1}^{N}\phi_i \tag{6}$$

However, the covariance matrix estimated from MLE fails to be a good approximation due to its rank-deficiency. Instead, the approximation of the covariance matrix is in the regularized form:

$$\hat{\Sigma}_{RKHS} = \Phi JQQ^T J^T \Phi^T + \rho I_f \tag{7}$$

where ρ is a constant, I_f is the identity matrix in RKHS and Q is an ancillary matrix. Q is obtained by solving the following eigenvalue equation:

$$\Lambda\nu = \mathbf{K}^*\nu \tag{8}$$

Then the solution of Q matrix is given as:

$$Q_{N\times r} = V_r(I_r - \rho{\Lambda_r}^{-1})^{1/2}R \tag{9}$$

where R can be any orthogonal matrix, which is assumed to be an identity matrix in this work. $\Lambda_r = Diag[\lambda_1, \cdots, \lambda_r]$ is a diagonal matrix whose diagonal elements are the r largest eigenvalues $\{\lambda_1, \cdots, \lambda_r\}$ of $\mathbf{K}^*$ and $V_r = [v_1, \cdots, v_r]$ is the matrix of corresponding eigenvectors.

The determinant and inversion of the covariance matrix of are also frequently used. In KGM, they are derived as :

$$\textbf{Determinant}: |\hat{\Sigma}_{RKHS}| = \rho^{f-r}|\Lambda_r| \tag{10}$$

$$\textbf{Inversion}: \hat{\Sigma}^{-1}_{RKHS} = \rho^{-1}(I_f - \Phi B\Phi^T) \tag{11}$$

where

$$B = JV_r(\Lambda_r^{-1} - \rho\Lambda_r^{-2})V_r^T J^T \tag{12}$$

Evaluation of commonly used probability distances [3, 5, 9] becomes trivial with these statistics, without explicitly evaluating Φ [11]. Therefore, KGM can be easily incorporated into many unsupervised SCD algorithms. For example, GLR in equation (2) becomes:

$$R_{KGM} = \frac{N_z}{2}log|\Lambda_r^z| - \frac{N_x}{2}log|\Lambda_r^x| - \frac{N_y}{2}log|\Lambda_r^y| \tag{13}$$

This implementation of KGM is equivalent to performing PKPCA on the feature vectors, that is, only r principal components of data in RKHS remained. Since the remaining principal components are of the largest variance, PKPCA works similar to PCA and Kernel-PCA [4, 8, 13]. It is believed that factors corrupting the speaker information of the features can be partly removed from the remained principal components. A more robust feature presentation may be obtained.

4. Experiments and result

4.1. Data

In order to evaluate the performance of KGM, different types of corpora are used, including both artificial data and real audio:

- **SIMULATION**: Data artificially created by concatenating 103 randomly chosen sentences from the TIMIT speech database. (Clean speech, about 5 minutes and 102 target speaker changes)
- **BN**: Subset of Hub4 1997 Mandarin Broadcast News speech. (Clean and prepared speech, 60 minutes, 117 target speaker changes)

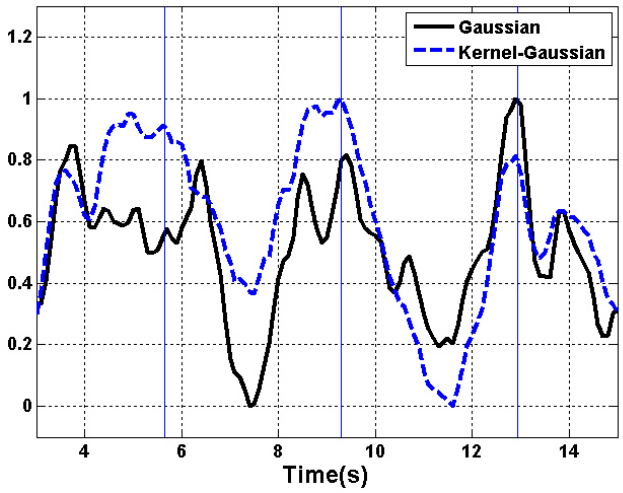


Figure 4: Excerpt of GLR distance graphs on the TIMIT dataset, with vertical bars bein g the real speaker change points.

- **CTS**: Telephone conversations extracted from self-collected data corpora. (Summed-channel condition, 40 minutes, 876 target speaker changes)

In our experiments all the data are sampled at 8 kHz and encoded by 16 bits, then windowed to 20ms frames with a 10ms overlap. 13- dimension MFCC features are calculated for each frame.

4.2. Evaluation Metrics

Performance of speaker change detection algorithms is evaluated on two types of errors. A *Type-I* error occurs when the process does not detect an existing speaker change point. A *Type-II* error occurs when a speaker change is detected although it does not exist. Type-I and II errors are also evaluated as precision (PRC) and recall (RCL) respectively, which are defined as:

$$PRC = \frac{\text{number of correctly found changes}}{\text{total number of changes found}} \tag{14}$$

$$RCL = \frac{\text{number of correctly found changes}}{\text{total number of correct changes}} \tag{15}$$

In order to compare the performance of different systems, an F measure combining PCR and RCL is defined as:

$$F = \frac{2 \times PRC \times RCL}{PCR + RCL} \tag{16}$$

The F measure varies from 0 to 1, the higher the F measure is the better performance it indicates.

4.3. Results

In our experiments, for the GLR evaluation, the window length (X or Y in Figure 1) is set to 1.5 seconds for **SIMULATION** and **CTS** datasets and 2 seconds for the **BN** dataset. Window shifts for all datasets are set to 0.1 seconds. The RBF kernel $K(x, y) = exp(-\gamma \cdot \parallel x - y \parallel)$ is adopted for KGM, where the kernel width γ is set empirically to 10. The dimension r in equation (9) is also determined empirically as 50.

Figure 4 shows an excerpt of GLR distance graphs on TIMIT data generated by the Gaussian model and KGM. Potential change points lie around local maxima. Both graphs are

normalized to be in the range 0 to 1 for the convenience of comparison. The figure shows that the Gaussian model may fail to detect the first change point, while KGM can correctly recall this one. This can be explained by the better modeling capability of KGM. Another point from Figure 4 is that the graph for the Gaussian model has more local peaks, which become potential false alarms. Meanwhile, the graph of KGM is smoother. This can be attributed to the PKPCA performed by our implementation of KGM. The quasi-PCA behavior may help to reduce factors affecting the discriminability of features, and a smoother version of distance graph is obtained.

All evaluation metric scores on the three datasets are shown in Table 1. Although the characteristics of the three datasets vary greatly, it is observed that KGM leads to improvements of 2% to 3% in F measure over the Gaussian model on all of them. These improvements are not very significant, however they are consistent. Hence, KGM is still a robust and effective alternative to the single Gaussian model.

Model:	Precision	Recall	F-Measure
SIMULATION			
Gaussian Model	0.646	0.803	0.716
KGM	0.672	0.823	0.740
BN			
Gaussian Model	0.558	0.786	0.652
KGM	0.587	0.803	0.683
CTS			
Gaussian Model	0.586	0.671	0.626
KGM	0.573	0.751	0.650

Table 1: *Evaluation metric scores on three datasets*

However, *there is no free lunch*. One of the issues to be addressed in adopting KGM is its computational complexity. The data mapping and parameter approximation procedure can be both expensive. Given N samples in the input space, N^2 kernel functions $K(\cdot,\cdot)$ have to be evaluated. An empirical evaluation is conducted on the processing speed of the proposed implementation versus the analysis window length in GLR. The BN dataset, 1 hour long, is used for this purpose. Results are shown in Table 2 in terms of processing time and real time norm ($\times RT$). It is observed that the efficiency of the proposed implementation is somewhat frustrating. The processing speed decreases dramatically with the window length and falls below real time. This means unsupervised SCD based on KGM can't be directly applied to real-time applications before some improvement to reduce its computational cost.

Window Length (Seconds)	1	1.5	2
Processing Time (hrs)	1.32	4.55	10.22
Processing Speed ($\times RT$)	0.76	0.22	0.098

Table 2: *Computational consumption versus analysis windows length of GLR on the **BN** dataset of one hour long*

5. Conclusion and future work

In this paper, a study on a novel Kernel-Gaussian model (KGM) for the unsupervised speaker change detection is conducted. The KGM is attractive because it can model nonlinear distribution of speech data. An implementation of KGM is proposed and evaluated. Experiments on different kinds of speech corpora shows that it leads to consistent performance improvements compared to single Gaussian model in terms of F measure. An investigation is also performed on the computational efficiency of the proposed implementation.

This work is still a preliminary study of KGM. As our experiments show, further work needs to be done to reduce computation cost of KGM to make it usable in real speaker change detection applications.

6. Acknowledgments

This work is partially supported by National Natural Science Foundation of China (10574140, 60535030), MOST (973program,2004CB318106), The National High Technology Research and Development Program of China (863 program, 2006AA010102,2006AA01Z195).

7. References

[1] P.Woodland, M.Gales, D.Pye, and S.Young, "The development of the 1996 htk broadcast news transcription system," in *Speech Recognition Workshop*, 1997, pp. 90–93.

[2] T.Kemp, M.Schmidt, M.Westphal, and A.Waibel, "Strategies for automatic segmentation of audio data," in *Proc.ICASSP'00*. IEEE, 2000, vol. 2, pp. 1423–1426.

[3] M.Cettolo, M.Vescovi, and R.Rizzi, "Evaluation of BIC-based algorithms for audio segmentation," *Computer Speech and Language*, vol. 19, no. 2, pp. 147–170, April 2005.

[4] J.Hung, H.Wang, and L.Lee, "Automatic metric-based speech segmentaion for broadcast news via principal component analysis," in *Proc.ICSLP'00*, 2000, vol. 6, pp. 121–124.

[5] P.Delacourt and C.J.Wellekens, "DISTBIC:a speaker-based segmentation for audio data indexing," *Speech Communncation*, vol. 32, pp. 111–126, January 2000.

[6] J.Ajmera, I.McCowan, and H.Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, August 2004.

[7] A.S.Malegaonkar, A.M.Ariyaeeinia, and P.Sivakumaran, "Efficient speaker change detection using adapted gaussian mixture models," *IEEE Tran.Audio Speech and Language Processing*, vol. 15, no. 6, pp. 1859–1869, August 2007.

[8] R.Q.Huang and J.H.L.Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *IEEE Tran. Audio Speech and Language Processing*, vol. 14, no. 3, pp. 907–919, May 2006.

[9] H.Gish, M.Siu, and R.Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc.ICASSP'91*. IEEE, 1991, vol. 2, pp. 873–876.

[10] D.Reynolds and R.C.Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Tran.Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

[11] S.Zhou and R.Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 917–929, June 2006.

[12] B. Scholkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, 2002.

[13] A.Lima, H.Zen, Y.Nankaku, C.Miyajima, K.Tokuda, and T.Kitamura, "On the use of kernel PCA for feature extraction in speech recognition," in *Proc.Eurospeech'03*. IEEE, 2003, pp. 2625–2628.