Intrinsic Spectral Analysis Based on Temporal Context Features for Query-by-Example Spoken Term Detection

Peng Yang[1], *Cheung-Chi Leung*[2], *Lei Xie*[1], *Bin Ma*[2], *Haizhou Li*[2]

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Institute for Infocomm Research, A*STAR, Singapore
{pengyang,lxie}@nwpu-aslp.org, {ccleung,mabin,hli}@i2r.a-star.edu.sg

Abstract

We investigate the use of intrinsic spectral analysis (ISA) for query-by-example spoken term detection (QbE-STD). In the task, spoken queries and test utterances in an audio archive are converted to ISA features, and dynamic time warping is applied to match the feature sequence in each query with those in test utterances. Motivated by manifold learning, ISA has been proposed to recover from untranscribed utterances a set of nonlinear basis functions for the speech manifold, and shown with improved phonetic separability and inherent speaker independence. Due to the coarticulation phenomenon in speech, we propose to use temporal context information to obtain the ISA features. Gaussian posteriorgram, as an efficient acoustic representation usually used in QbE-STD, is considered a baseline feature. Experimental results on the TIMIT speech corpus show that the ISA features can provide a relative 13.5% improvement in mean average precision over the baseline features, when the temporal context information is used.

Index Terms: spoken term detection, intrinsic spectral analysis, Gaussian posteriorgram, dynamic time warping

1. Introduction

Spoken term detection (STD) refers to the task of finding the occurrences of a given query in an audio archive. Usually, the query is provided in text form. Under this condition, a sophisticated large vocabulary continuous speech recognition (LVCSR) system is needed to transcribe the utterances in the test archive into their textual representations, and then detection is done on the recognition lattices. This requires a large amount of annotated speech data to train the speech recognizer, and out of vocabulary (OOV) problem will occur if the query contains words that are not in the recognition vocabulary.

Query-by-example spoken term detection (QbE-STD) is another kind of scenario in which queries are spoken examples. In this situation, the detector can convert spoken queries and test utterances into sequences of acoustic features, and dynamic time warping (DTW) is applied to match two sequences of the features — one from a query and another from a test utterance. QbE-STD is suitable for low-resource languages and even unknown languages.

Posteriorgram features have been shown better performance than raw spectral features in QbE-STD [1, 2, 3] and other speech tasks [4, 5, 6]. In [1], phoneme posteriorgrams are generated using a well-trained phoneme recognizer. In [2], a Gaussian mixture model (GMM) trained using a set of unlabeled data is employed to generate Gaussian posteriorgrams. In [3], HMMs are trained in an unsupervised way to extract acoustic segment model (ASM) posteriorgrams. It is known that unsupervised posteriorgram has a performance gap compared with the supervised one for a QbE-STD task. However, training a phoneme recognizer usually requires at least hours of annotated speech data. For many languages, it is difficult to collect enough necessary resources. Thus we are interested to find an unsupervised way to extract a more useful feature representation to reduce the performance gap between using unsupervised and supervised methods.

Similar to Gaussian posteriorgrams, intrinsic spectral analysis (ISA) features provide a data representation which has been shown less sensitive to speaker variations. This motivates us to investigate the use of ISA for QbE-STD. ISA is formulated in [7] as an unsupervised manifold learning algorithm, which is derived from the unsupervised learning case of Manifold Regularization [8]. ISA provides a natural regularized out-of-example extension of Laplacian eigenmaps [9]. Laplacian eigenmaps, as a dimensionality reduction and data representation method, have been shown its success in story segmentation [10, 11] and image segmentation [12]. Spectral clustering, which has close connection to Laplacian eigenmaps, has been shown useful for unsupervised acoustic unit mining [13, 14]. Recently supervised ISA [15] has been applied in phone classification with performance improvement. The experiments on isolated word matching in [7] show ISA's superiority of data representation over traditional spectral features, such as MFCC and PLP. However, the comparison between ISA features and Gaussian posteriorgram features is not available in [7].

Due to the coarticulation phenomenon in speech, any sound in a continuous speech production process is influenced by its neighbor context. Some works [16, 17, 18] taking this phenomenon into consideration report performance improvement in their frameworks. In this paper, we propose to use temporal context information to have a more accurate graph Laplacian in order to obtain a better intrinsic projection maps in ISA, and eventually the improved ISA features can provide performance improvement for the QbE-STD task.

We are interested to compare the performance of ISA features and Gaussian posteriorgrams in a QbE-STD task, because both of them have been shown superior performance compared with spectral features in separate studies. To the best of our knowledge, this is the first attempt to do this comparison in QbE-STD or similar tasks using sequence matching on acoustic features. Note that intrinsic projection maps that generate ISA features and the Gaussian components that generate posteriorgram features can be considered as nonlinear transformations. However, there is significant difference between the two nonlinear transformations. While the intrinsic projection maps derive

ISA features from log mel spectrograms, the Gaussian components generate posteriorgrams from spectral features, such as MFCC, which is from discrete cosine transform of log mel spectrogram. Discrete cosine transform in MFCC is used to decorrelate each element of the features so that the resultant features can be modeled with Gaussian components with diagonal covariance matrices. Moreover, while intrinsic projection maps would preserve the neighborhood relations of the input data, each frame of unlabeled data is considered independently in the training of Gaussian components.

Our experimental results on the TIMIT corpus show that the ISA features can provide a relative improvement of 13.5% in MAP, 13.4% in P@N and 8.6% in P@10, over Gaussian posteriorgrams, when temporal context information is used in ISA. We also observe that the three parameters in ISA are not sensitive to the use of temporal context information.

2. Intrinsic spectral analysis

Intrinsic spectral analysis is an unsupervised learning algorithm which is designed to recover from unlabeled speech data a set of nonlinear basis functions for the speech manifold [7]. Given a set of unlabeled training data, $X = \{x_1, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d$ for all i, that forms a mesh of data points that lie on the manifold, we can construct a weighted, undirected adjacency graph $G(V, E)$ with one vertex per data point. As in [8], edges between vertices V_i and V_j are defined by either k-nearest neighbors or ϵ-neighborhoods, and the corresponding edge weights W_{ij} can be defined either using binary weights or heat kernel weights. A graph Laplacian $\mathbf{L}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with elements $D_{ii} = \sum_j W_{ij}$.

The nonlinear basis function f can be learned by solving the optimization problem

$$f^* = \arg\min_{f \in H_K} \|f\|_K^2 + \xi \mathbf{f}^T \mathbf{L} \mathbf{f}, \tag{1}$$

where H_K is the reproducing kernel Hilbert space (RHKS) for some positive semi-definite kernel function K, $\mathbf{f} = [f(x_1), \ldots, f(x_n)]^T$ is the vector of values of the basis function f for each point of the training data, and ξ is the regularization parameter controlling the relative importance of the two term: the first term is the extrinsic norm, representing the complexity of the solution, and the second term is the Laplacian eigenmaps objective function. By the RHKS representer theorem, the solution of Eq. (1) can be written as

$$f^*(v) = \sum_{i=1}^{n} a_i K(x_i, v), \tag{2}$$

where $\mathbf{a} = [a_1, \ldots, a_n]^T \in \mathbb{R}^n$ is the eigenvector of the following generalized eigenvalue problem

$$(\mathbf{I} + \xi \mathbf{L}\mathbf{K})\mathbf{a} = \lambda \mathbf{K}\mathbf{a}, \tag{3}$$

where $\mathbf{K}$ is the $n \times n$ Gram matrix defined on the unlabeled training data X by $K_{ij} = K(x_i, x_j)$, and $\mathbf{I}$ is the identity matrix. In this paper we chose the configuration as in [7]: k-nearest neighbors and binary weights are used to define the adjacency graph; radial basis function (RBF), $K(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$, is used as the kernel function K.

Before solving the general eigenvalue problem in Eq. (3), three parameters, including k, ξ, and σ, need to be chosen: k refers to the number of the nearest neighbors used to define the graph Laplacian $\mathbf{L}$, ξ refers to the weighting parameters in Eq. (3), and σ is the width of the RBF kernel mentioned above. A full spectrum of eigenvectors can be attained by solving Eq. (3). According to Eq. (2), each eigenvector corresponds to a nonlinear intrinsic basis function, and then out-of-sample data points can be transformed into a low-dimensional form using the first i ($i << d$) intrinsic basis functions (sorted by the eigenvalues in an ascendant order), which is believed to represent the underlying manifold structure. The corresponding non-linear projection maps are also known as intrinsic projection maps.

2.1. Making use of temporal context

We investigate to use temporal context information to better obtain the underlying manifold structure. This is motivated by the coarticulation phenomenon, in which any sound in a continuous speech production process is influenced by its neighor context. Since the utterances in the task are unannotated and there is no prior information about the distinctive sounds (e.g. phonemes) in the utterances, we make use of the temporal contextual information of each speech frame by concatenating consecutive frames of short-time spectral features (log mel spectrograms). More precisely, consider a short-time spectral feature $\mathbf{s}_i \in \mathbb{R}^d$ at time index i. The temporal context features $\mathbf{s}_i^c$ can be obtained as follows:

$$\mathbf{s}_i^c = [\mathbf{s}_{i-(c-1)/2}^T, \ldots, \mathbf{s}_i^T, \ldots, \mathbf{s}_{i+(c-1)/2}^T]^T, \tag{4}$$

where c denotes the number of contextual frames concatenated into a supervector. Note that when $c = 1$, $\mathbf{s}_i^c$ becomes the original features as in [7] without the context information. When temporal context is used, the number of dimensions of data points in X is increased by c times (i.e. $x_i \in \mathbb{R}^{cd}$ for all i). Note that the elements in the graph Laplacian $\mathbf{L}$ and the kernel matrix $\mathbf{K}$ are obtained by the observations of speech frames in the increased dimensions, but the matrices keep their original dimensions ($n \times n$). In practice, the first and last frame in an utterance is duplicated as necessary to let each frame have enough contextual frames to be concatenated.

3. The QbE-STD framework

Query-by-example spoken term detection (QbE-STD) involves two main modules: feature extraction and detection by dynamic time warping.

3.1. Feature extraction

The extraction of ISA features has been presented in Section 2. MFCC feature and Gaussian posteriorgram, which are usually used in the QbE-STD task, are considered as the baseline feature representations in this paper. The Gaussian posteriorgram is a feature representation generated from a GMM. Given a set of unlabeled training data, $X = \{x_1, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d$ for all i, a J-mixture GMM is trained, then for each data point x_i, its posterior probability distribution over all the Gaussian components, forms a J dimensional vector, $P_i = [p(m_1|x_i), \ldots, p(m_J|x_i)]^T$, which then can be used as the new representation of the original data point x_i.

3.2. Detection by dynamic time warping

Variants of dynamic time warping (DTW) have been used to align two sequences of acoustic features for various tasks, such as speech pattern discovery [1, 2, 3, 19, 20, 21], story segmentation [5] and speech summarization [22]. In this paper, to match

a spoken query with the subsequence of a test utterance, the algorithm in [20] is used. Given two sequences of acoustic features, one from a spoken query $u = [u_1, \ldots, u_r]$ and another from a test utterance $v = [v_1, \ldots, v_s]$, where r and s are the lengths of the sequences, the distance $d(i, j)$ between any two feature vectors u_i and v_j are computed. For the MFCC and ISA features, we use the cosine distance

$$d(i, j) = 1 - \frac{u_i^T v_j}{|u_i|\,|v_j|}. \quad (5)$$

For the Gaussian posteriorgram, we use the inner-product distance

$$d(i, j) = -\log(u_i^T v_j). \quad (6)$$

The DTW algorithm finds a path in the distance matrix $d(i, j)$ starting from a location $(1, b)$ to another location (r, e) (where $1 \leq b \leq e \leq s$), such that its average accumulated distance is minimum. The average accumulated distance of the best path ending at location (i, j) is defined as $cost(i, j) = a(i, j)/l(i, j)$, where $a(i, j)$ is the corresponding accumulated distance and $l(i, j)$ is the length of the best path. The value of $cost(i, j)$ can be obtained using a dynamic programming algorithm. For $i = 1, \ldots, r$, $a(i, 1) = (\sum_{k=1}^{i} d(i, 1))/l(i, 1)$, $l(i, 1) = i$. For $j = 1, \ldots, s$, $a(1, j) = d(1, j)$, $l(1, j) = 1$. For other locations (i, j), we choose a neighboring precedent point $(\hat{p}, \hat{q})$ from $(i - 1, j)$, $(i - 1, j - 1)$ and $(i, j - 1)$ such that the following distance is minimized:

$$\frac{a(\hat{p}, \hat{q}) + d(i, j)}{l(\hat{p}, \hat{q}) + 1}. \quad (7)$$

Then we obtain $a(i, j)$ and $l(i, j)$ as follows:

$$\begin{cases} a(i, j) = a(\hat{p}, \hat{q}) + d(i, j) \\ l(i, j) = l(\hat{p}, \hat{q}) + 1 \end{cases}. \quad (8)$$

Finally, for $e = 1, \ldots, s$, the minimum value of $cost(r, e)$ is set to be the dissimilarity score between the spoken query u and the matched subsequence in the test uttterance v. For the spoken query u, all the test utterances are ranked in an ascending order according to these disimilarity scores.

4. Experiments

The QbE-STD experiment is performed on the TIMIT corpus. From the 3,696 utterances of the training set, we extract 69 spoken queries which are at least 0.35s in duration and contain at least 6 English letters. The 944 utterances of the test set are used as the test archive. For each spoken query, an utterance is deemed a correct hit, if it contains the query term.

Three evaluation metrics are used for the performance measure: 1) Mean average precision (MAP), the mean of the average precision after each correct hit utterance is retrieved; 2) P@N, the average precision of the top N utterances, where N is the number of the correct hit utterances in the test set; 3) P@10, that is the average precision over the first 10 ranked utterances.

4.1. Details on feature extraction

A vector of 39 MFCC features, consisting of 12 cepstral coefficients, log energy, and their delta and acceleration cofficients, were computed every 10ms with a 25ms analysis window. Given a speech frame, Gaussian posteriograms were generated using a 50-component GMM [2], which was trained using the MFCC features in the training set of the TIMIT corpus.

Table 1: *Performance of ISA features. Average results in Gaussian posteriorgrams (GP) and ISA features. Values in parentheses indicate standard deviation of the performance. Note that MFCC and GP make use of temporal context information implicitly by delta and acceleration coefficients in MFCC.*

Feature	MAP	P@N	P@10
MFCC	0.243	0.241	0.213
GP	0.325(0.006)	0.314(0.008)	0.279(0.007)
ISA-1	0.296(0.003)	0.278(0.006)	0.250(0.008)
ISA-3	**0.369(0.005)**	**0.356(0.005)**	0.303(0.005)
ISA-5	0.367(0.005)	0.345(0.006)	**0.311(0.006)**
ISA-7	0.357(0.008)	0.334(0.012)	0.297(0.009)
ISA-9	0.340(0.008)	0.321(0.008)	0.287(0.009)

In ISA, log mel spectrograms, which were extracted with 40 mel channels and a 25ms analysis window at every 10ms as in [7], were used as the input for intrinsic projection maps. As described in Section 2.1, consecutive spectrograms were concatenated according to the parameter c, the number of features being concatenated. A set of 5,000 speech frames was randomly selected from the training set to form the mesh of data points, the graph Laplacian $\mathbf{L}$ and the kernel matrix $\mathbf{K}$. As in [7], the distance metric used to determine the k-nearest neighbors is the cosine distance defined in Eq. (5). The first 13 intrinsic components (skipping the first trivial one) were kept to project the original feature vector into 13 intrinsic components, and the delta and acceleration coefficients were appended to take into account the time derivatives of the basic static coefficients. The number of dimensions of the ISA features is 39, the same as that of the MFCC features.

Note that the MFCC and ISA features were post-processed with utterance-based cepstral mean and variance normalization. We found that this post-processing is important for the STD performance, especially when their delta and acceleration coefficients are included.

4.2. Comparison of different feature representations

Table 1 shows the performance of different feature representations on the QbE-STD task. The STD performance was varied by the 5,000 data points selected for the intrinsic projection maps in ISA. Similarly in Gaussian posteriorgrams, GMM, which is trained for generating posteriorgrams, depends on a set of the parameters initialized by k-means algorithm and varies the STD performance. So the performances of these features reported in the table are averaged results by running each individual system 20 times, and the values in the parentheses indicate the standard deviation of the performance. Note that the number c in the notation ISA-c indicates the number of log mel spectrograms being concatenated as a temporal context feature. In each set of ISA-c features, we chose $k = 10, \xi = 0.03$, and $\sigma = 0.8m$, where m is the average Euclidean distance between all graph vertices. We found this choice of the parameters in general provided good performance for different sets of ISA-c features.

As Table 1 shows, Gaussian posteriorgrams give better performance than MFCC features as expected. ISA features also provide better performance than MFCC features. When temporal context features are used ($c > 1$) , ISA features obviously outperform Gaussian posteriorgrams. Among different sets of ISA features, ISA-3 features provide the best performance in terms of MAP and P@N, and ISA-5 features provide the best performance in terms of P@10.

Table 2: *Temporal context features (log mel spectrograms; without intrinsic projection maps) provide better performance on QbE-STD.*

metric	c	MAP	P@N	P@10
Cosine	1	0.142	0.129	0.122
	3	0.170	0.164	0.138
Euclidean	1	0.165	0.180	0.158
	3	0.196	0.192	0.171

Moreover, we would point out that both MFCC features and Gaussian posteriorgrams make use of temporal context information implicitly by the delta and acceleration coefficients in MFCC features. We believe that temporal context features provide more accurate acoustic similarity for finding the k-nearest neighbors, and better representation for the graph Laplacian $\mathbf{L}$ and the kernel matrix $\mathbf{K}$, and thus better obtain the underlying speech manifold.

To verify whether the temporal contextual information obtained by concatenating consecutive features make the ISA better recover the speech manifold, we did an auxiliary experiment — a comparison of using log mel spectrogram with and without temporal context for the QbE-STD task. The corresponding results are in Table 2. We believe that if the log mel spectrograms with temporal context can do better in recovering the speech manifold, this temporal context spectrogram features can provide better STD performance than the original spectrogram features, though these log mel spectrogram features are not good candidates for the distance matrices for DTW detection.

As Table 2 shows, no matter whether the cosine or Euclidean distance is used, temporal context spectrogram features ($c = 3$) improve the STD performance. We believe that when temporal context spectrogram features are used in ISA, both the graph Laplacian $\mathbf{L}$, which is built based on the cosine distance defined in Eq. (5), and the RBF kernel $\mathbf{K}$, which is related to Euclidean distance, obtain the more accurate pairwise acoustic similarity measure between the sampled speech frames. This probably leads ISA to generate more reliable intrinsic basis functions. Note that although the log mel spectrograms reported in Table 2 provide obvious improvement when three consecutive frames are concatenated, this would not be the case for the systems reported in Table 1 when consecutive MFCC or ISA frames are concatenated for DTW detection. This is because appending delta and acceleration features has the similar effect as concatenating consecutive frames.

4.3. Effect of parameters in intrinsic spectral analysis

Lastly, we study the effect of the three parameters in ISA, including k, ξ and σ, on the QbE-STD task. In this last set of experiments, we varied k from 5 to 12, ξ from 0.003 to 30 (multiplying by 10 in each step), and σ from $0.2m$, $0.4m$, $0.6m$, $0.8m$ to $1.0m$. Since we observed that the performance changes were similar when different number of features are concatenated, we present the effect on performance using ISA-3 features in Figure 1. As mentioned in Section 4.2, since the STD performance was varied by the choice of a mesh of data points in ISA, each individual system was run five times and the results reported in the figures are averaged results. From the figures, we observe that the ISA features perform the best when ξ = 0.03 and σ = $0.8m$. The STD performance is relatively less sensitive to k, the number of the nearest neighbors used to define the adjacency graph and thus the graph Laplacian $\mathbf{L}$.

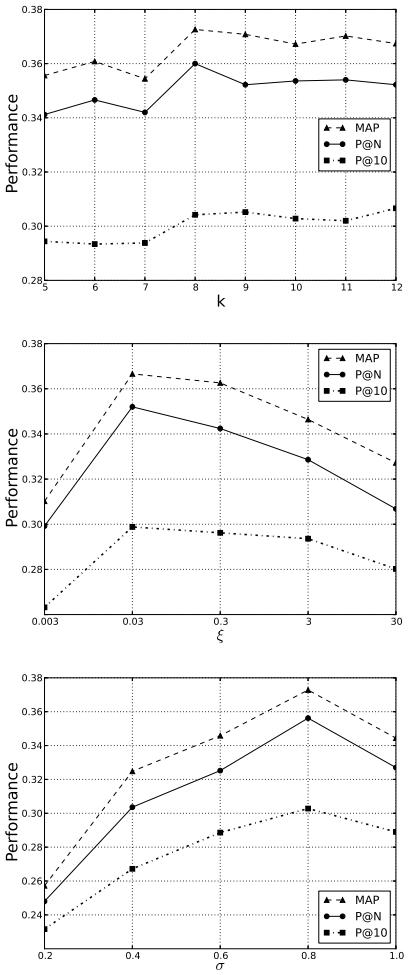

Figure 1: *The performance changes with* k $(\xi = 0.03, \sigma = 0.8)$, ξ $(k = 10, \sigma = 0.8)$ *and* σ $(k = 10, \xi = 0.03)$.

5. Conclusions

In this paper, we investigate to use temporal context information for intrinsic spectral analysis, and use it to generate feature sequences for dynamic time warping detection in query-by-example spoken term detection. Our experiments show that concatenating consecutive frames of log mel spectrograms in ISA provides obviously better performance than using Gaussian posteriorgrams on the spoken term detection task. We believe that the proposed method can be straightforwardly used in other zero-resource speech tasks [5, 22], which are based on searching for repeated acoustic patterns.

In the future, we would evaluate the features on speech in different domains, e.g., conversational telephone speech. We would investigate whether the learned intrinsic projection maps are portable across domains. Moreover, we would investigate the choice of parameters k, ξ and σ, when speech from different domains is involved. Ensemble manifold regularization [23] will be considered for more elegant choice of the parameters.

6. Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (61175018) and a grant from the Fok Ying Tung Education Foundation (131059).

7. References

[1] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009, pp. 421-426.

[2] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. ASRU*, 2009, pp. 398-403.

[3] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. ICASSP*, 2012, pp. 5157–5160.

[4] G. Aradilla, J. Vepa, H. Bourlard," Using posterior-based features in template matching for speech recognition," in *Proc. INTERSPEECH*, 2006, pp. 1186–1189.

[5] L. Zheng, C.-C. Leung, L. Xie, B. Ma and H. Li, "Acoustic texttiling for story segmentation of spoken documents" in *Proc. ICASSP*, 2012, pp. 5121–5124.

[6] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li. "Shifted-delta MLP features for spoken language recognition," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15–18, 2013.

[7] A. Jansen, and P. Niyogi, "Intrinsic spectral analysis," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1698–1710, 2013.

[8] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization:a geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data represenation," *Neural Comput.*, vol. 16, pp. 1373-1396, 2003.

[10] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, "Laplacian eigenmaps for automatic story segmentation of broadcast news," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 276–289, 2012.

[11] X. Lu, L. Xie, C.-C. Leung, B. Ma, and H. Li, "Broadcast news story segmentation using manifold learning on latent topic distributions," in *Proc. ACL*, 2013, pp. 190-195.

[12] J. Shi, and J. Malik, "Normalized cuts and image segmentation", *IEEE Trans. PAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[13] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Unsupervised mining of acoustic subword units with segment-level Gaussian posteriorgrams," in *Proc. INTERSPEECH*, 2013, pp. 2297–2301.

[14] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "A graph-based Gaussian component clustering approach to unsupervised acoustic modeling," in *Proc. INTERSPEECH*, 2014, to appear.

[15] R. Sahraeian, and D. Van Compernolle, "A study of supervised intrinsic spectral analysis for TIMIT phone classification," in *Proc. ASRU*, 2013, pp. 256–260.

[16] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Context dependent modeling of phones in continuous speech using decision trees," in *Proc. DARPA Speech and Natural Language Processing Workshop*, 1991, pp. 264–270.

[17] P. Schwarz, P. Matejka, and J. Cernock, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, 2006, pp. 325–328.

[18] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, "Exploiting contextual information for improved phoneme recognition," in *Proc. ICASSP*, 2008, pp. 4449–4452.

[19] P. Yang, L. Xie, Q. Luan, and W. Feng, "A tighter lower bound estimate for dynamic time warping," in *Proc. ICASSP*, 2013, pp. 8525-8529.

[20] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Proc. INTERSPEECH*, 2009, pp. 2843-2846.

[21] X. Anguera, R. Macrae, and N. Oliver, "Partial sequence matching using an unbounded dynamic time warping algorithm," in *Proc. ICASSP*, 2010, pp. 3582-3585.

[22] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. EMNLP*, 2010, pp. 460-470.

[23] B. Geng, D. Tao, C. Xu, Y. Yang, and X. S. Hua, "Ensemble manifold regularization," *IEEE Trans. PAMI*, vol. 34, no. 6, pp. 1227–1233, 2012.