Unsupervised Bottleneck Features for Low-Resource Query-by-Example Spoken Term Detection

Hongjie Chen[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

{hjchen,lxie}@nwpu-aslp.org, {ccleung,mabin,hli}@i2r.a-star.edu.sg

Abstract

We propose a framework which ports Dirichlet Gaussian mixture model (DPGMM) based labels to deep neural network (DNN). The DNN is used to extract a low-dimensional unsupervised speech representation, named as unsupervised bottleneck feature (uBNF), which captures considerable information for sound cluster discrimination. We investigate the performance of uBNF in query-by-example spoken term detection (QbE-STD) on the TIMIT English speech corpus. Our uBNF performs comparably with the cross-lingual bottleneck features (M-BNF) extracted from a DNN trained using 171 hours of transcribed telephone speech in another language (Mandarin Chinese). With the score fusion of uBNF and M-BNF, we gain about 10% relative improvement in term of mean average precision (MAP) comparing with M-BNF. We also studied the performance of the framework with different input features and different lengths of temporal context.

Index Terms: unsupervised feature learning, low-resource speech processing, Dirichlet process Gaussian mixture model, spoken term detection, bottleneck feature

1. Introduction

Query-by-example spoken term detection (QbE-STD) [1, 2] search queries in an audio archive by acoustic pattern matching between spoken examples and test utterances in the archive. Since this task does not necessarily require the linguistic expertise and transcription of the target data, it has gained attention from researchers in recent years [1, 2, 3, 4].

Various types of unsupervised acoustic features, which are based on spectral features [1, 2], frame clustering [1, 2, 5], segment clustering with GMM-HMM modeling [6], speech manifold [7, 8, 4] and deep neural networks (DNN) [9, 10], has been studied in QbE-STD or related tasks. These unsupervised frameworks attempt to discover the phonetic or phonetic-like units, and model them purely from speech data without linguistic expertise or transcription. On the other hand, some works [11, 12, 13, 14, 15] studied the acoustic features whose models were trained supervisedly using the linguistic expertise and transcription of high-resource non-target languages. Moreover, fusion of unsupervised and/or supervised features has also been studied in [16, 3, 17, 14, 15], which suggests that unsupervised learning and supervised learning can gather complementary knowledges from the target speech data and benefits the down-stream tasks.

In this paper, inspired by [12] which studied deep Boltzmann machines (DBMs) supervised by Gaussian mixture model (GMM) based labeling to extract better posteriorgrams, we propose to extract a low-dimensional data presentation based on deep neural networks (DNNs) in an unsupervised way. Firstly, a Dirichlet process Gaussian mixture model (DPGMM) is trained on speech frames with no transcription and each speech frame is transcribed with the DPGMM's component assignment. Then a bottleneck-shape DNN (BN-DNN) is trained with the transcribed speech frames. By forward passing feature vectors of the queries and test utterances through the trained DNNs, the output of an inner bottleneck layer yield a low-dimensional representation, referred to as bottleneck features (BNFs). Since DPGMM is an unsupervised clustering algorithm, we refer to our proposed DNNs and the corresponding BNFs as unsupervised DNNs (uDNNs) and unsupervised BNFs (uBNFs) respectively.

Our proposed features are evaluated on the TIMIT speech corpus. We employ subsequence-DTW (SDTW) [18] for acoustic pattern matching in QbE-STD. Recently cross-lingual BNFs are widely used in QbE-STD [15, 19]. Cross-lingual BNFs (M-BNF) extracted from a DNN supervisedly trained using the HKUST Mandarin Chinese telephone speech corpus (LDC2005S15) are considered as baseline features in our experiments. We conduct comparison between uBNFs, uDNN-based posteriorgrams (uDNN-PG), DPGMM-based posteriorgrams (PG) and the baseline features. To investigate whether our uBNF and M-BNF can provide complementary information for QbE-STD, we perform the score fusion of these two sets of features. Moreover, DNNs do not require uncorrelated input features. The work in ASR [20] showed that the DNN trained using filter-bank features (FBank) can outperform that trained using MFCC. We would study whether similar observation is also made in our unsupervised training scenario. Moreover, we study the effect of temporal context length of input features on uBNFs for QbE-STD.

Note that our framework differs from [12] in the following aspects. 1) We use DPGMM instead of GMM. Note that [1, 2, 12] utilized small numbers (e.g. 50 or 61) of Gaussian components while [21, 14, 15] used hundreds of Gaussain components. We argue that a small number of Gaussian components is usually not sufficient to represent the speech data and the model complexity should be tuned on a development dataset. Meanwhile, our previous study [5] illustrated that DPGMM can learn its model complexity (i.e. the number of Gaussian components) automatically according to the observed speech data. Moreover, although a set of hyper-parameters is involved for configuring DPGMM, as shown in [5], DPGMM is not sensitive to the choice of hyper-parameters and parallel inferring DPGMM is scalable to a large amount of speech frames. In summary, porting DNNs to DPGMM labeling makes our uDNNs more suitable for QbE-STD in scenarios where development dataset or linguistic knowledge is inaccessible, and it makes the uDNNs feasible to large-scale speech dataset. 2) We extract features from the inner bottleneck layer instead of the last softmax layer. The bottleneck features provide a more compact feature representation than the posterior features provided in the last softmax layer, while retaining considerable information for sound cluster classification. The low-dimensional feature representation reduces tremendous time cost and storage load for dynamic time warping (DTW) in QbE-STD. 3) We show that using FBank together with F0 features in training the uDNNs brings performance gain while MFCC were used in [12].

2. Unsupervised Bottleneck Features

The proposed unsupervised bottleneck features (uBNFs) are extracted from bottleneck-shaped DNNs trained with labels from an unsupervised clustering method. Here we refer to this type of DNNs as uDNNs. Our proposed framework for training the uNNs consists of two modules, unsupervised cluster labeling using Dirichlet process mixture model (DPGMM) and DNN training using the unsupervised labels.

2.1. Unsupervised Labeling Using DPGMM

Studies in [2, 12, 21, 14, 15, 5] suggest that the complexity of GMM needs fine-tuning on a development set to represent speech sufficiently. Thus instead of GMM, we employ a Bayesian nonparametric model, Dirichlet process Gaussian mixture model (DPGMM) also referred to as infinite Gaussian mixture model (IGMM), to represent speech since DPGMM is able to learn a suitable number of components. DPGMM can be depicted as a graphical model illustrated in Figure. 1.

Meanwhile, as studied in [5], parallel inference of DPGMM based on split/merge sampling [22] can be scalable to a large amount of speech frames and DPGMM is insensitive to the hyper-parameters, α and $\boldsymbol{\theta}_0 = \{\mathbf{m}_0, \mathbf{S}_0, \kappa_0, \nu_0\}$. This leads DPGMM to be a suitable and feasible choice to represent speech data, especially for low-resource language scenarios. To be concise, we recall the physical meaning of these hyper-parameters without details of DPGMM modeling. Specifically, α specifies the prior distribution of the mixing weights. $\mathbf{m}_0$ is the prior mean for the mean of each component. $\mathbf{S}_0$ is proportional to the prior mean for the covariance matrix for each component. κ_0 and ν_0 measure the belief-strength in $\mathbf{m}_0$ and $\mathbf{S}_0$, respectively.

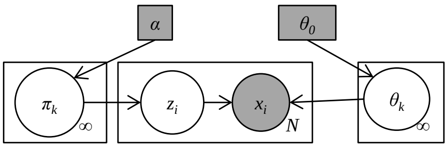

Figure 1: Graphical representation of Dirichlet process Gaussian mixture model (DPGMM).

Note that some other Bayesian nonparametric models [23, 24] were developed to model speech data. Considering the inference efficiency of Bayesian nonparametric models, this paper merely investigates DNNs with DPGMM labeling.

Given feature vectors of speech frames $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, the inference of DPGMM results in K components together with their mixing weights, $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$, mean vectors, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^K$ and covariance matrix, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k\}_{k=1}^K$. The posterior probability of the k-th component conditioned on i-th observed speech frame, $\mathbf{x}_i$, can be computed as follows:

$$p_{i,k} = p(c_k|\mathbf{x}_i) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \Sigma_j)}. \quad (1)$$

Then $\mathbf{p}_i = (p_{i,1}, ..., p_{i,K})(i = 1, ..., N)$ forms a posteriorgram. The label of $\mathbf{x}_i$, denoted as l_i is computed as follows:

$$l_i(\mathbf{x}_i) = \arg \max_{1 \le k \le K} p_{i,k}. \quad (2)$$

2.2. DNN Training Using Unsupervised Labels

For low-resource settings, DNN training is infeasible due to the lack of transcribed data. However, [12] argues that a weaker classifier is able to initialize a stronger classifier and it presented a way to train Deep Boltzmann Machines (DBM) in which the fully unsupervised learnt labels using GMM are used as transcription. Based on this assumption and their positive results, we train DNNs by incorporating DPGMM labeling presented in Sec. 2.1. Since DPGMM is a unsupervised clustering method, we refer to our DNNs as uDNNs. Noted that the number of targets in DBM in [12] is set to 61, the size of English phoneme set. Since DPGMM is free of manual selection of number of components which is more suitable to low-resource scenario. Meanwhile, to save time during downstream searching steps and space for storage, we form the DNNs with bottleneck layer, as illustrated in Figure. 2, to extract low-dimensional speech representation, referred to as uBNF.

Specifically, the first layer of our uDNNs expands the raw feature by concatenating the input vector with its left and right N (+/-N) feature vectors to utilize temporal context information. The internal bottleneck layer consists of linear transformation units and the last layer consists of linear transformation units and Softmax operation. All other hidden layers consist of linear transformation units together with Sigmoid operation. The configuration of our DNNs is illustrated in Table. 1.

For training, firstly we employ Restricted Boltzmann Machines (RBMs) to conduct pre-training and then the uDNNs are trained by minimizing cross-entropy with maximum 20 iterations. Both of pre-training and training procedures are run with Kaldi [25].

3. Query-by-example Spoken Term Detection

Query-by-example spoken term detection (QbE-STD) consists of two successive modules, including feature extraction and detection based on pattern matching. The framework of our QbE-STD system is illustrated in Figure. 2.

3.1. Feature Extraction

In feature extraction, various acoustic features, such as Mel-frequency cepstral coefficients (MFCC), posteriorgrams and bottleneck features (BNFs), are commonly used in QbE-STD. In this paper, the uBNFs and uDNN-PGs are obtained by forward-passing raw speech features through the proposed uDNNs by taking the outputs of the inner bottleneck layer and the last layer of the uDNN presented in Sec. 2.2, respectively. DPGMM-based posteriorgrams can be computed as described in Sec. 2.1.

3.2. Detection Based on Pattern Matching

Acoustic pattern matching can be implemented in various variants of dynamic time warping (DTW) which is used to align

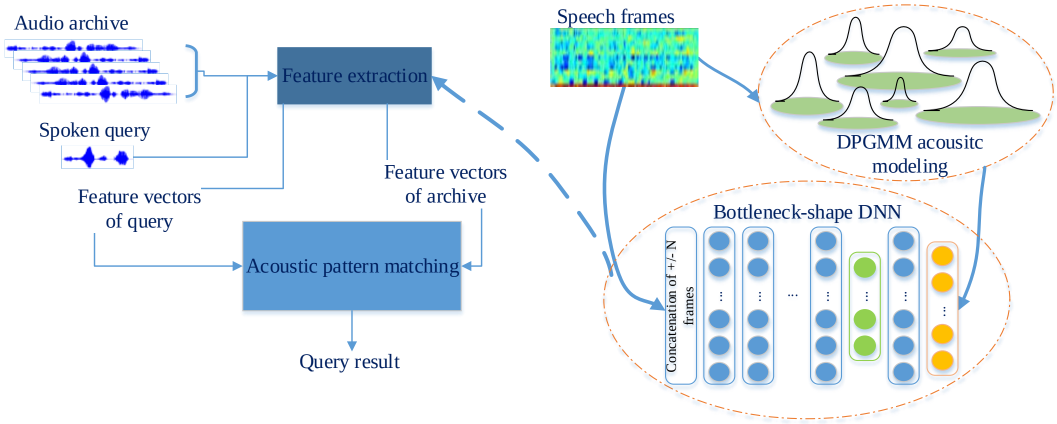

Figure 2: Framework of QbE-STD based on unsupervised bottleneck features (uBNFs).

two sequences of acoustic feature vectors in various tasks, e.g. speech pattern discovery [1, 2], speech summarization [26], story segmentation [27, 28], etc. In this paper, we employ subsequence-DTW (SDTW) described in [18] to conduct acoustic pattern matching between a query and a test utterance in the retrieval archive.

Given two sequences of acoustic feature vectors from a spoken query and an test utterance, $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_M)$ and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N)$ where the M and N are the lengths of the query and the test utterance, for MFCC and BNFs, the distance $d_{i,j}$ between arbitrary two vectors, $\mathbf{q}_i$ and $\mathbf{u}_j$ is computed as

$$d_{i,j} = 1 - \frac{\mathbf{q}_i^T \mathbf{u}_j}{|\mathbf{q}_i||\mathbf{u}_j|}. \tag{3}$$

For posteriorgrams, we use negative logarithm of inner-product

$$d_{i,j} = -\log(\mathbf{q}_i^T \mathbf{u}_j). \tag{4}$$

Using the SDTW based on dynamic programming (DP) on the distance matrix composed by aforementioned distances, we can find an optimal path with minimum distance cost which can be regarded as the dissimilarity between the query, $\mathbf{Q}$, and the test utterance, $\mathbf{U}$. For the spoken query $\mathbf{Q}$, all the test utterances $\mathcal{U} = \{\mathbf{U}_k\}_{k=1}^{K_U}$ are ranked in an ascending order according to the dissimilarities.

4. Experiments

4.1. Data and Experimental Setup

Our QbE-STD experiments are conducted on the TIMIT speech corpus which consists of a training set of 4620 utterances and a test set of 944 utterances. The training set is used to train the DPGMM and then divided into two subsets, training subset and cross-validation subset with the size ratio of 9:1, for training DNNs. We extracted 69 queries, totally 346 spoken examples, which consist of at least 6 English letters and are at least 0.35s. The test set is used as the retrieval archive. For each query, a correct hit is counted if an utterance in the retrieval contains the query.

We conducted the comparison study of MFCC, BNFs and posteriorgrams. Firstly, 39-dimensional MFCC (13-dimensional MFCC+Δ+$\Delta\Delta$) are extracted and then post-processed by cepstral mean and variance normalization (CMVN), which is our weak baseline feature. Secondly, 40-dimensional BNF extracted from a Mandarin Chinese tokenizer (M-BNF) which is a stacked bottleneck-shaped DNN [19] trained on 171 hours of speech from the HKUST Mandarin telephone corpus (LDC2005S15). In the stacked hierarchical network, the first-stage network takes filter-bank and pitch features as input. The first-stage and the second-stage networks have the topology of 1500-1500-80-412 and 1500-1500-40-412 respectively, where the 412 is the number of senones. We regard M-BNF as a strong baseline feature since it is extracted from a tokenizer trained using manual transcripts. Thirdly, we test uBNFs denoting 40-dimensional BNFs extracted from our proposed uDNNs. During unsupervised labeling, we trained DPGMM on MFCC of the training set setting parameters following [5]. Specifically, these parameters including α and $\boldsymbol{\theta_0} = \{\mathbf{m}_0, \mathbf{S}_0, \kappa_0, \nu_0\}$. $\mathbf{m}_0$ and $\mathbf{S}_0$ are set as the global mean and covariance of the post-processed MFCC, respectively. ν_0 is set to 41. α and κ_0 are set to 1. After training, we obtained a 306-component DPGMM. Then we trained the uDNNs with the configurations listed in Table. 1. Finally, posteriorgrams from the DPGMM (PG) and DPGMM-supervised DNN (uDNN-PGs) are tested. Additionally, we conducted a score fusion of the language-mismatched BNF and the best uBNF in QbE-STD, where the fusion weight of each type of BNFs is set to 0.5.

Table 1: Configurations of DNNs.

Input Features	Ndim	Len. Con.	Size
MFCC	39	+/-1,+/-3,+/-5	1024x4,40,1024,306
FBank	36	+/-1,+/-3,+/-5	1024x4,40,1024,306
FBank+F0	39	+/-1,+/-3,+/-5	1024x4,40,1024,306

Moreover, comparison of different input features for extraction of uBNFs are conducted in term of the QbE-STD performance. MFCC, FBank and FBank with pitch (FBank+F0) are studied in the comparison to see whether our uDNN requires input features to be correlated. In the end, we studied the effect of temporal context length of uBNF with different input features.

All spectral features, including MFCC, FBank, and FBank+F0, are extracted using Kaldi and have the same window length (25 ms) and shift length (10 ms). And, the QbE-STD is evaluated in term of three metrics: 1) **MAP**: the mean average precision for correct hits in the retrieval; 2) **P@N**: the average precision of the top N utterances where N is the number of the correct hit utterances in the retrieval archive; 3) **P@10** : the average precision over the first 10 ranked utterances.

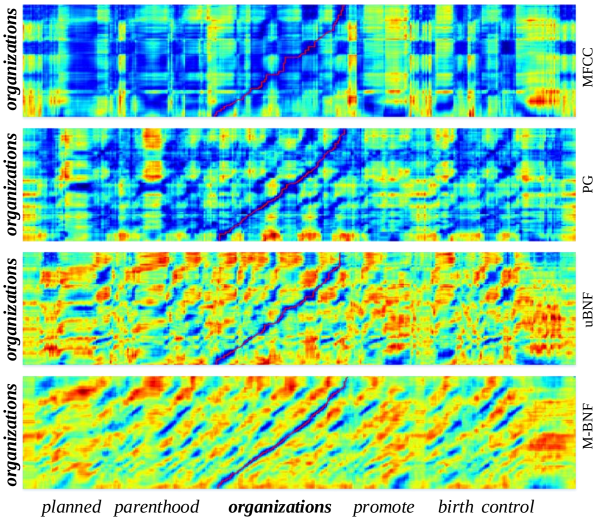

Figure 3: Distance matrix of feature representations for a query, *organizations* and a test utterance *planned parenthood organizations promote birth control* with optimal path found by S-DTW.

Table 2: Comparison of features with score fusion.

Feature/Method	NDim	MAP	P@N	P@10
MFCC	39	0.285	0.289	0.247
ISA [4]	39	0.369	0.356	0.303
PG	306	0.405	0.390	0.340
uDNN-PG	306	0.412	0.396	0.349
M-BNF	40	0.494	0.459	0.413
uBNF	40	**0.494**	**0.47**	**0.412**
Score fusion (M-BNF & uBNF)	-	**0.543**	**0.508**	**0.453**

4.2. Results and Discussion

Table. 2 gives the comparison of different feature representations with score fusion. Here we also report the best result of intrinsic spectral analysis (ISA) based feature from [4] since ISA is a low-dimensional unsupervised representation although there is a slight difference in the query and retrieval archive.

As illustrated in Table. 2, uBNF clearly outperform MFCC, PG and uDNN-PG in term of the three evaluation metrics. The uDNN is shown capable of predicting the Gaussian component labels given by DPGMM, and even improves PG by its deep structure. This is similar to the observation in [12]. Compared with uDNN-PG, the gain made by uBNF may be attributed to the difference of distance measure used in these two different types of features. It is worth noting that uBNF also provide a more compact representation than uDNN-PG.

Moreover, uBNF perform comparably with M-BNF. Note that the M-BNF extractor is trained using a large amount of transcribed out-of-domain data. It has mismatches in languages, channels and speaker styles with the target data. These mismatches are not uncommon when working on a low-resource language. M-BNF is used to show a baseline performance which is achievable when transcribed data in the target language is not available. We believe that our proposed framework is effective to capture considerable frame-based information for sound cluster discrimination.

Figure. 3 shows the distance matrix of a pair of query and test utterances, with the spoken word(s) of the query (test) utterance listed on the vertical (horizontal) axis. Colours in the distance matrix depict the distance between speech frames, with red (blue) indicating large (small) distances. A more salient blue diagonal band in the hit region, and larger areas of red and yellow in other regions are revealed in uBNF. The distance matrix of uBNF and M-BNF show the similar observation. These observations are consistent with the results shown in Table. 2.

Table 3: Effect of raw input feature of uBNFs.

Feature	MAP	P@N	P@10
MFCC	0.459	0.432	0.391
FBank	0.489	0.459	0.411
FBank+F0	**0.494**	**0.47**	**0.412**

Table 4: Effect of temporal context length.

Feature(Raw Feature)	Len. Con.	MAP	P@N	P@10
uBNF (MFCC)	+/-1	**0.459**	**0.432**	**0.391**
uBNF (MFCC)	+/-3	0.449	0.427	0.375
uBNF (MFCC)	+/-5	0.444	0.419	0.381
uBNF (FBank)	+/-1	0.4	0.383	0.338
uBNF (FBank)	+/-3	0.468	0.440	0.395
uBNF (FBank)	+/-5	**0.489**	**0.459**	**0.411**
uBNF (FBank+F0)	+/-1	0.436	0.416	0.366
uBNF (FBank+F0)	+/-3	0.488	0.457	0.405
uBNF (FBank+F0)	+/-5	**0.494**	**0.47**	**0.412**

Noteworthily, by score fusion of M-BNF with our uBNF, the performance gains a boost in QbE-STD with about 10% relative improvement in term of MAP comparing with using M-BNF, respectively. This tell us that supervised learnt feature and unsupervised learnt feature can capture complementary phonetic information and fusion of them is a simple but powerful strategy to improve the QbE-STD performance.

Table. 3 illustrates the effect of different input features in uBNFs extraction for QbE-STD. As shown in Table. 3, FBank performs better than MFCC. This is consistent with the observation in conventional ASR [20], that DNN is a more flexible model which does not require the input features to be uncorrelated. Meanwhile, FBank+F0 perform the best, showing that pitch does not hurt the performance in the non-tonal target language which is consistent with the observation in [29].

Table. 4 shows the effect of different temporal context lengths in uBNFs extraction on QbE-STD. When MFCC are used as input features, concatenation with +/- 1 frame context gives the best performance. However, FBank and FBank+F0 perform the best when concatenation with +/- 5 frame context is used. Moreover, uBNFs based on FBank or FBank+F0 generally perform better than MFCC when larger than +/- 3 frame context is used.

5. Conclusions

In this paper, we proposed a framework which ports DPGMM-based labels to DNNs. Since the DPGMM can cluster speech frames effectively in an unsupervised fashion and does not require to set its model complexity manually, our framework is suitable to low-resource QbE-STD and scenarios where no development dataset is accessible. Within this framework, we can extract a low-dimensional speech representation which can capture considerable information for sound cluster discrimination comparably with cross-lingual BNF. By the score fusion of the proposed uBNF and M-BNF, we can obtain about 10% relative improvement comparing with M-BNF in QbE-STD in term of MAP. Moreover, this framework inherits the feature of DNNs that it is insensitive to whether the input feature is uncorrelated. FBank+F0 is suitable for uBNF extractions to conduct QbE-STD when longer temporal context is used.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61571363).

7. References

[1] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 186–197, 2008.

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

[3] H. Wang, T. Lee, C. C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with dtw matrix combination for low-resource spoken term detection," in *Proc. ICASSP*. IEEE, 2013, pp. 8545–8549.

[4] P. Yang, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection," in *Proc. INTERSPEECH*. ISCA, 2014, pp. 1722–1726.

[5] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel Inference of Dirichlet Process Gaussian Mixture Models for Unsupervised Acoustic Modeling: A Feasibility Study," in *Proc. INTERSPEECH*. ISCA, 2015, pp. 3189–3193.

[6] 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*. IEEE, 2012, pp. 5157–5160.

[7] A. Jansen, S. Thomas, and H. Hermansky, "Intrinsic spectral analysis for zero and high resource speech recognition." in *Proc. INTERSPEECH*. ISCA, 2012, pp. 879–882.

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

[9] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An autoencoder based approach to unsupervised learning of subword units," in *Proc. ICASSP*. IEEE, 2014, pp. 7634–7638.

[10] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak topdown constraints," in *Proc. ICASSP*. IEEE, 2015, pp. 5818–5822.

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

[12] Y. Zhang, R. Salakhutdinov, H. A. Chang, and J. Glass, "Resource configurable spoken query detection using deep boltzmann machines," in *Proc. ICASSP*. IEEE, 2012, pp. 5161–5164.

[13] G. S. T. Schatz and E. Dupoux, "Phonetics embedding learning with side information," in *Proc. SLT*. IEEE, 2014, pp. 106–111.

[14] P. Yang, H. Xu, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow, B. Ma, C. E. Siong, and H. Li, "The NNI query-by-example system for mediaeval 2014," in *Working Notes Proc. MediaEval 2014*, 2014.

[15] J. Hou, V. T. Pham, C.-C. Leung, L. Wang, H. Xu, H. Lv, L. Xie, Z. Fu, C. Ni, X. Xiao, H. Chen, S. Zhang, S. Sun, Y. Yuan, P. Li, T. L. Nwe, S. Sivadas, B. Ma, E. S. Chng, and H. Li, "The NNI Query-by-Example System for MediaEval 2015," in *Working Notes Proc. Mediaeval 2015*, 2015.

[16] A. Muscariello, G. Gravier, and F. Bimbot, "Zero-resource audio-only spoken term detection based on a combination of template matching techniques," in *Proc. INTERSPEECH*. ISCA, 2011, pp. 921–924.

[17] B. Ludusan, A. Caranica, H. Cucu, A. Buzo, C. Burileanu, and E. Dupoux, "Exploring multi-language resources for unsupervised spoken term discovery," in *Proc. SpeD*. IEEE, 2015, pp. 1–6.

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

[19] K. Veselỳ, M. Karafiát, and F. Grézl, "Convolutive bottleneck network features for LVCSR," in *Proc. ASRU*. IEEE, 2011, pp. 42–47.

[20] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, November 2012.

[21] 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*. ISCA, 2013, pp. 2297–2301.

[22] J. Chang and J. W. Fisher III, "Parallel sampling of DP mixture models using sub-clusters splits," in *Proc. NIPS*, 2013, pp. 620–628.

[23] C.-Y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proc. ACL*. ACL, 2012, pp. 40–49.

[24] A. H. Harati Nejad Torbati, "Nonparametric bayesian approaches for acoustic modeling," Ph.D. dissertation, TEMPLE UNIVERSITY, 2015.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, no. EPFL-CONF-192584. IEEE, 2011.

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

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

[28] H. Chen, L. Xie, W. Feng, L. Zheng, and Y. Zhang, "Topic segmentation on spoken documents using self-validated acoustic cuts," *Soft Computing*, vol. 19, pp. 47–59, 2015.

[29] F. Metze, Z. A. W. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," in *Proc. ASRU*. IEEE, 2013, pp. 261–266.