UNSUPERVISED DISCOVERY OF LINGUISTIC STRUCTURE INCLUDING TWO-LEVEL ACOUSTIC PATTERNS USING THREE CASCADED STAGES OF ITERATIVE OPTIMIZATION

Cheng-Tao Chung[#1], *Chun-an Chan*[*2], *and Lin-shan Lee*[#*3]

Graduate Institute of Electrical Engineering, National Taiwan University[#]
Graduate Institute of Communication Engineering, National Taiwan University[*]
r01921031@ntu.edu.tw[1], chunanchan@gmail.com[2], lslee@gate.sinica.edu.tw[3]

ABSTRACT

Techniques for unsupervised discovery of acoustic patterns are getting increasingly attractive, because huge quantities of speech data are becoming available but manual annotations remain hard to acquire. In this paper, we propose an approach for unsupervised discovery of linguistic structure for the target spoken language given raw speech data. This linguistic structure includes two-level (subword-like and word-like) acoustic patterns, the lexicon of word-like patterns in terms of subword-like patterns and the N-gram language model based on word-like patterns. All patterns, models, and parameters can be automatically learned from the unlabelled speech corpus. This is achieved by an initialization step followed by three cascaded stages for acoustic, linguistic, and lexical iterative optimization. The lexicon of word-like patterns defines allowed consecutive sequence of HMMs for subword-like patterns. In each iteration, model training and decoding produces updated labels from which the lexicon and HMMs can be further updated. In this way, model parameters and decoded labels are respectively optimized in each iteration, and the knowledge about the linguistic structure is learned gradually layer after layer. The proposed approach was tested in preliminary experiments on a corpus of Mandarin broadcast news, including a task of spoken term detection with performance compared to a parallel test using models trained in a supervised way. Results show that the proposed system not only yields reasonable performance on its own, but is also complimentary to existing large vocabulary ASR systems.

Index Terms— unsupervised learning, hidden Markov models, spoken term detection, zero resource speech recognition, iterative optimization

1. INTRODUCTION

Supervised training of HMMs for automatic speech recognition relies on not only collecting huge quantities of acoustic data, but also obtaining the corresponding precise labels. Such supervised training method yields adequate performance in most circumstances but with high cost, and in many situations such annotated data sets are simply not available. This is why substantial effort [1]-[18] has been made for unsupervised discovery of acoustic patterns from huge quantities of acoustic data which may be easily obtained nowadays, without manual labels and corresponding knowledge. Most of such effort discovered only one level of phoneme-like acoustic patterns. However, it is well known that speech signals have multi-level structure including at least phonemes and words, and such structure are very helpful in analyzing or decoding speech[19].

In this paper we propose an approach for unsupervised discovery of structured two-level acoustic patterns including subword-like patterns and word-like patterns (concatenation of several subword-like patterns). Not only the HMMs for these patterns, the number of the subword-like patterns and the lexicon size of word-like patterns can be automatically learned from data, but more knowledge about the language such as the N-gram language model and the word-like pattern lexicon, jointly referred to as the linguistic structure in this paper, can all be obtained directly from the acoustic signals of a corpus. This is achieved by integrating a dynamic lexicon into the process of the conventional supervised HMM-training, and performing three stages of iterative optimization between the labels and the models, such that the models, parameters, and the linguistic structure can then collect knowledge from the corpus layer after layer iteratively and adjust themselves accordingly. In this way, we are able to develop semantic building blocks of the target spoken language represented by the corpus with word-like patterns and acoustic building blocks of the target spoken language with subword-like patterns.

2. PROPOSED APPROACH: CASCADED THREE STAGES OF ITERATIVE OPTIMIZATION

The goal is to find the parameter set $\theta = \{\theta^a, \theta^x, \theta^l\}$ for the linguistic structure and the word-like pattern label W given the observed acoustic feature vector sequences $\bar{O}$ for the corpus considered. The parameter set θ includes three parts: θ^a for acoustic HMMs of subword-like patterns, θ^x for lexicon of word-like patterns in terms of subword-like pattern sequences, and θ^l for N-gram word-like pattern language model. This is achieved by first finding an initial label W_0 for the observation $\bar{O}$ as in (1). In each iteration i, we train the parameters θ_i with the label W_{i-1} obtained in the previous iteration as in (2) and decode the label W_i with the obtained parameters θ_i as in (3).

$$W_0 = \text{initialization}(\bar{O}), \quad (1)$$

$$\theta_i = \arg\max_{\theta} P(\bar{O}|\theta, W_{i-1}), \quad (2)$$

$$W_i = \arg\max_{W} P(\bar{O}|\theta_i, W). \quad (3)$$

The iterations above are organized as an initialization step followed by three cascaded stages (I)(II)(III) respectively for acoustic, linguistic and lexical optimization as shown in Fig. 1. In Fig. 1, the number of iterations for each stage are I_a, I_l and I_x respectively. When the difference between W_{i-1}, W_i becomes insignificant, the process then advances to the next stage. The parameters θ_i^a are generated by EM training as in (2), while the other parameters θ_i^l or θ_i^x

are generated directly from the labels W_{i-1} obtained in the previous iteration. However, not all of θ^a, θ^x, and θ^l are used in each stage. The detailed updating procedure is depicted in Fig. 2 and will be explained shortly.

The basic idea behind the procedure in Fig. 1 is to gradually construct and update the parameters layer after layer. This prevents the parameters from being caught in local optimal situations which often happen when too many parameters are optimized at once. First, the HMM parameters for the subword-like patterns are trained alone in stage (I), because these HMMs are the primary building blocks of the whole linguistic structure and reliable estimate for their parameters is the key. With reliable enough HMMs for subword-like patterns, we then in stage (II) use N-gram parameters for word-like patterns to better decode those word-like patterns frequently appearing together while continuously updating the HMM parameters. Finally in the stage (III), we break the word-like patterns into subword-like patterns and reconstruct better word-like patterns. The number of word-like patterns in the lexicon may shrink in the iterations of the first two stages because some less frequent patterns can be absorbed by other patterns, but this number can be changed significantly in the third stage. The time alignment for the subword-like patterns are updated in all iterations when the labels W_i are decoded.

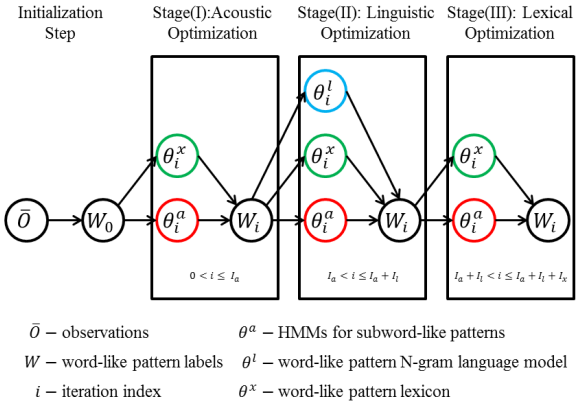

Fig. 1. Simplified diagram for the proposed initialization step followed by three stages of iterative optimization. Some dependency links have been omitted.

2.1. Initialization Step

Here we initialize the labels in a top-down fashion by first breaking each utterance into word-like segments based on the discontinuities in the energy of the MFCC features. For each word-like segment, we further divide it into subword-like segments in the following way. We perform a watershed transform on the filtered self-similarity dotplot [20] for acoustic features of each hypothesized word-like segment. Watershed transformation is able to capture the number of objects and their borders in a gray scale image [21]. So, the intersections of the diagonal entries of the dot-plot with the watershed transform object borders are taken as the boundaries between subword-like segments. An example dotplot and its watershed transform including the hypothesized subword-like segment boundaries is shown in Fig. 3.

We then extract an average representative feature vector for every hypothesized subword-like segment, and perform global k-means clustering on these representative vectors obtained from the whole corpus. The number of clusters (the initial number of subword-like

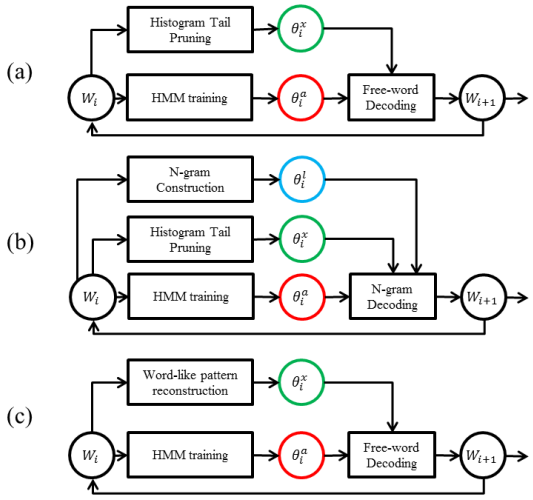

Fig. 2. Detailed diagrams for the three stages of (a)acoustic (b)linguistic and (c)lexical optimization.

patterns) is determined by the ratio of the within-cluster total scattering to the between-cluster total scattering. A subword-like pattern ID is then assigned to each cluster. A distinct sequence of consecutive subword-like patterns for word-like segments then defines a word-like pattern, and the total number of distinct word-like patterns in the corpus is the initial vocabulary size of the lexicon. The corpus is thus represented by its initial labels W_0.

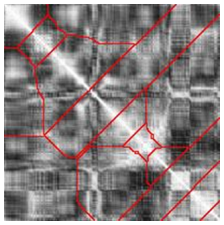

Fig. 3. An example dotplot and its watershed transform.

2.2. Stage(I):Acoustic Optimization

The process in stage(I) is shown in Fig. 2(a). In each iteration, the acoustic model set θ_i^a are the HMMs trained from the corpus based on W_i with the ML criterion. The lexicon θ_i^x is derived by collecting all word-like patterns appearing in W_i with counts exceeding a threshold. Free word decoding is then performed on the whole corpus $\bar{O}$ based on θ_i^a and θ_i^x, producing an updated label W_{i+1}. When W_i is updated to W_{i+1}, not only the HMM parameters of θ_i^a and HMM segmentation boundaries are updated, but the vocabulary size of θ_i^x may shrink when the counts of some word-like patterns become small enough.

2.3. Stage(II):Linguistic Optimization

This stage is shown in Fig. 2(b), which is very similar to the previous stage. The only difference is an N-gram language model θ_i^l for the word-like patterns is estimated from the label W_i and is used in

decoding to produce the updated labels W_{i+1}. The N-grams help produce better labels W_{i+1} especially for word-like patterns appearing together frequently.

2.4. Stage(III):Lexical Optimization

We reconstruct new word-like patterns in this step as in Fig. 2(c). This is done by breaking the word-like patterns in θ^x_{i-1} into subword-like patterns, and then reconstructing new word-like patterns based on W_i. Those segments of several consecutive subword-like patterns appearing frequent enough and with high enough right and left context variation are taken as word-like patterns. This can be achieved by constructing an efficient data structure called PAT-Tree using the labels W_i[22]. In this way, the lexicon θ^x_i can be updated significantly in each iteration. This updated lexicon θ^x_i is then used in free-word decoding to produce the labels W_{i+1}. The whole process is completed when there is no significant difference between W_i and W_{i+1}. This gives the automatically discovered linguistic structure $\theta = \{\theta^a, \theta^x, \theta^l\}$, where θ^l is trained from the final version of W_{i+1}.

3. EXPERIMENTS

3.1. Experimental Setup

The proposed approach was tested in the preliminary experiments performed on a corpus of Mandarin broadcast news collected in Taiwan in 2001 with length of 4 hours including 5034 utterances. The HMMs used for each sub-word like pattern had 13 states, each with only 1 Gaussian component. This configuration was selected due to the assumption that the subword-like patterns of interest should describe more signal trajectory variation and less acoustic variation. Signal segments with larger acoustic variation should be classified as different patterns. The final linguistic structure including all patterns, models and parameters was obtained by performing 30 iterations in each stage (I)(II)(III) in Fig. 1 on the entire corpus.

3.2. Initial Observations and Analysis

It is interesting that almost all the 208 subword-like patterns obtained here roughly correspond to Mandarin syllables (each Chinese character is pronounced as a Mandarin syllable). A global view of the exact mapping relation from the 208 subword-like patterns to the total of 399 Mandarin syllables manually labelled for the corpus is shown in Fig. 4. The Mandarin syllables on the horizontal scale of the figure have been sorted according to acoustic similarity (only a quarter of them are explicitly printed due to limited space). Every circle here represents 35 or more subword-like patterns on the vertical scale whose central feature frame belonged to the Mandarin syllable in the horizontal scale. This figure implied a very-close-to one-to-one mapping relation with some fuzziness around neighbouring syllables with similar acoustic behaviour. The 362 word-like pattern obtained corresponded to roughly 154 frequently occurring multi-syllable words and 208 monosyllables (or mono-subword-like patterns). Those words occurring not frequently enough couldn't be discovered and as a result were represented as one to several mono-subword-like patterns.

Fig. 5 further illustrates how the number of subword-like patterns, lexicon size of word-like patterns, the consistency between W_{i-1} and W_i at word-like pattern level and utterance level changed with respect to iterations. In a global perspective, lexicon size of word-like patterns dropped in the stages (I) and (II), and jumped and oscillated in stage (III). Although most word-like patterns in stage (I) did not survive by the end of stage (II), the main purpose of them was to provide some context guidance for the training of subword-like HMMs.

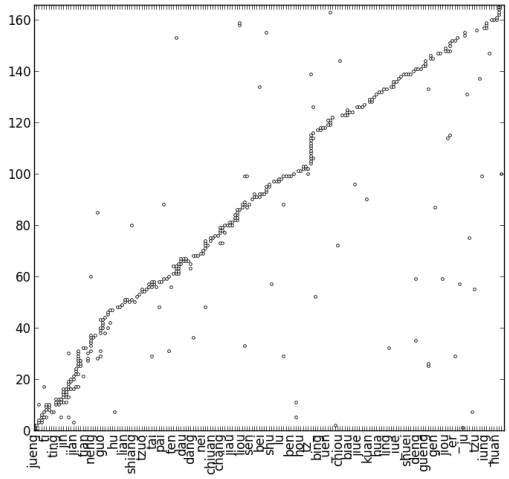

Fig. 4. Mapping relation between the discovered subword-like patterns and Mandarin syllables. Only pairs with 35 or more occurrence are shown, and the average co-occurrence mapping for all circles in the figure is 331.

3.3. Justification of the Initialization and Iterative Stages

We performed further tests with configurations slightly different from the proposed approach on a subset of 942 utterances out of the 5034 in the tested corpus. We evaluated the syllable accuracy by mapping every discovered subword-like pattern to a corresponding Mandarin syllable (as was done in Fig. 4) for each configuration considered. In the first part, we initialized W_0 with 3 different methods and then applied 50 iterations of stage (I) only. The three methods are (1) the proposed two-level top-down labelling started with word-like segments, (2) subword initialization with only watershed transform, but without higher level word-like segments, (3) same as (2) but without k-means clustering, with same number of subword-like pattern IDs randomly assigned to each subword-like segment. The main difference between methods (1)(2) was the two-level pattern structure. Method (1) brought us halfway through the proposed approach (initialization and stage (I)) producing two-level patterns, while method (2) was similar to the unsupervised initialization methods used previously with one-level patterns only [1][20]. The results are in the left half of Table 1. Although method (1) was only 1.03% better than method (2), the patterns obtained with method (1) manual auditing tests suggest that the improvement is non-trivial. This verified the word-like pattern constraints were useful in the acoustic optimization process. The random ID assignments without clustering in method (3) also offered relatively high accuracy. This implied the acoustic optimization iterations in stage (I) was quite helpful.

In the second part, we initialized W_0 with the two-layered method then applied 3 different iteration sequences: (1) $(I_a, I_l, I_x) = (30, 20, 0)$, (2) $(I_a, I_l, I_x) = (50, 0, 0)$, (3) $(I_a, I_l, I_x) = (0, 50, 0)$. Method (1) brought us halfway through the proposed approach while method (3) was actually the intuitive joint optimization

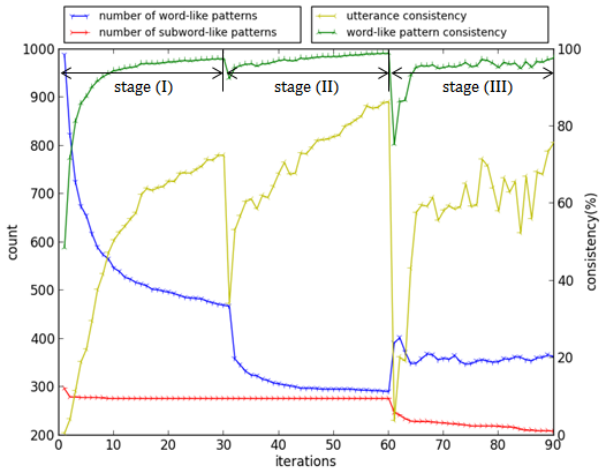

Fig. 5. Number of subword-like patterns, lexicon size for word-like patterns (left) and consistency between W_i and W_{i+1} in terms of word-like patterns and utterances (right) as functions of iterations. The transition from stage(I) to stage(II) and stage(II) to stage(III) happened at iteration 30 and 60 respectively.

(A)Initialization methods		(B)Iteration methods	
(1)Two-level	38.96%	(1)(I_a,I_l,I_x)=(30,20,0)	39.45%
(2)One-level	37.93%	(2)(I_a,I_l,I_x)=(50,0,0)	38.96%
(3)Random	35.76%	(3)(I_a,I_l,I_x)=(0,50,0)	37.08%

Table 1. ASR accuracy of unsupervised transcription translated by string replacement with most probable assignment

of both acoustic and linguistic parameters similar to previously proposed approaches [3][4]. The results are in the right half of Table 1. The proposed method (1) was 2.37% better than the joint optimization method (3). The proposed method (1) was also better than the applying method (2) alone, which implies that the transition was the source of improvement. This verified that gradually learning layer after layer yielded more reliable results. The benefits of the lexical optimization in stage (III), on the other hand, are better observed in a companion paper on semantic retrieval of spoken content also submitted to ICASSP 2013[23], since the word-like patterns carried semantics.

3.4. Spoken Term Detection

We also applied the discovered patterns on a task of spoken term detection [25]-[30] and compared to a set of Mandarin syllable models trained on a manually annotated corpus of 24.5 hours of Mandarin Broadcast News with a trigram for 72k vocabulary used in recognition. The performance of the supervised HMMs serves as an upper bound for the performance of our unsupervised HMMs. We tested the performance of the supervised and unsupervised models under the same scenario. The query set consisted of 52 name entities of countries, organizations and political leaders. For each query, we decoded their corresponding utterances in the corpus and selected the most frequent HMM sequence to represent each query (equivalent to query by one example of the best query utterance). Syllable HMMs were used for the supervised case, and subword-like pattern HMMs were used for the unsupervised case. This query HMM sequence was then compared with the HMM sequences of all utterances in the corpus for evaluation of the relevance scores for retrieval. We first computed offline the distance between each pair of two HMMs. The distance between two HMMs was defined to be the DTW-distance between the two state sequences. One state in a HMM can be matched with several states in another HMM and vice versa. The distance metric used for DTW was the KL-divergence between the two Gaussian mixtures [24]. We then calculate the distance between the query HMM sequences and corpus HMM sequences online. The distance between two HMM sequences was defined to be the sum of distances for matched pairs of models for the two sequences. Since most computation was done offline, this method was as fast as text information retrieval.

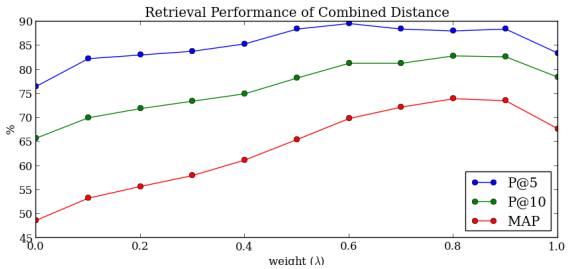

Fig. 6. The spoken term detection performance based on the weighted sum of unsupervised(left) and supervised(right) distance metrics.

We took the weighted sum of the supervised distance d_s and unsupervised distance d_u, and performed spoken term detection based on the combined distance $d_\lambda = \lambda \times d_s + (1-\lambda) \times d_u$. For detailed definitions of P@5, P@10 and MAP see [35]. The results in Fig. 6 show that reasonable detection performance was achieved for the unsupervised model on its own ($\lambda = 0$). More importantly, the combined distance can yield better results in all the three measures than using only supervised or unsupervised distances. This implies that the proposed method has successfully harvested information directly from the data that was lost during recognition with the supervised models. In other words, the proposed method not only performs reasonably well on its own, but it is also complimentary to standard supervised ASR systems.

4. CONCLUSION

This work presents an approach for unsupervised discovery of linguistic structure including two-level acoustic patterns from a corpus. The main difference from similar approaches proposed earlier [1][2][3][4][5][6][7] lies in the two-level acoustic patterns and the layer-after-layer gradual learning of the model parameters with cascaded stages of iterative optimization. Although some earlier approaches [1] also took hierarchical knowledge into consideration, our work used 13-state single Gaussian HMMs as compared to the conventional HMMs with smaller number of states and multi-Gaussian [1][2][3][4] to model the trajectories of acoustic patterns with less acoustic variation. The preliminary experiment on spoken term detection on subword-like pattern sequences indicated that the proposed system is complimentary to existing ASR systems. A more complete experiment on spoken term detection in a companion paper submitted to ICASSP 2013 [15] demonstrates how our model can outperform the segmental DTW approach. Also, the second level of word-like patterns are aimed to capture some semantic features in the acoustic signal, which can be verified in a companion paper on Semantic retrieval of spoken content also submitted to ICASSP 2013 [23].

5. REFERENCES

[1] A. Jansen and K. Church "Towards Unsupervised Training of Speaker Independent Acoustic Models" in *InterSpeech*, 2011, pp. 1693–1696.

[2] C. Lee and J. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery" in *Proc. The Association for Computer Linguistics*, 2012, vol. 1, pp. 40–49.

[3] H. Gish, M. Siu, A. Chan, and B. Belfield, "Unsupervised training of an HMM-based Speech Recognizer for Topic Classification" in *InterSpeech*, 2009, pp. 1935–1938.

[4] M. Siu, H. Gish, A. Chan, and W. Belfield, "Improved Topic Classification and Keyword Discovery using an HMM-based Speech Recognizer Trained without Supervision" in *InterSpeech*, 2010, pp. 2838–2841.

[5] M. Huijbregts, M. McLaren, and D. van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *ICASSP*, 2011, pp. 4436–4439.

[6] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *ICASSP*, 2009, pp. 4297–4300.

[7] C. Chan and L. Lee, "Unsupervised Hidden Markov Modeling of Spoken Queries for Spoken Term Detection without Speech Recognition" in *InterSpeech*, 2011, pp. 2141–2144.

[8] O. J. Rasanen, U. K. Laine, T. Altosaar "Computational language acquisition by statistical bottom-up processing," in *InterSpeech*, 2008, pp. 1980–1983.

[9] O. J. Rasanen, U. K. Laine, T. Altosaar "A noise robust method for pattern discovery in quantized time series: the concept matrix approach," in *InterSpeech*, 2009, pp. 3035–3038.

[10] O. J. Rasanen, U. K. Laine, T. Altosaar, "Self-learning vector quantization for pattern discovery from speech," in *InterSpeech*, 2009, pp. 852–855.

[11] O. J. Rasanen, Fully unsupervised word learning from continuous speech using transitional probabilities of atomic acoustic events in *InterSpeech*, 2010, pp. 2922–2925.

[12] C. Chan, "Unsupervised Spoken Term Detection with Spoken Queries " Ph.D dissertation, National Taiwan University, July, 2012.

[13] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," in *ICASSP*, 2008, pp. 3989–3992.

[14] C. Chan and L. Lee, "Integrating Frame-based and Segment-based Dynamic Time Warping for Unsupervised Spoken Term Detection with Spoken Queries" in *ICASSP*, 2011, pp. 5652–5655.

[15] C. Chan, C. Chung, Y. Kuo and L. Lee, "Toward Unsupervised Model-based Spoken Term Detection with Spoken Queries without Annotated Data" in *ICASSP*, 2013

[16] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," in *ICASSP*, 1992, vol. 1, pp. 573–576.

[17] H. Singer and M. Ostendorf, "Maximum likelihood successive state splitting," in *ICASSP*, 1997, vol. 2, pp. 601–604.

[18] B. Varadarajan and S. Khudanpur, "Automatically Learning Speaker-independent Acoustic Subword Units," in *InterSpeech*, 2008.

[19] Y.-c. Pan and L.-s. Lee, "Performance analysis for lattice-based speech indexing approaches using word and subword units," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, August 2010, pp. 1562–1574.

[20] A. Jansen, K. Church, and H. Hermansky, "Towards Spoken Term Discovery At Scale With Zero Resources" in *InterSpeech*, 2010, pp. 1676–1679.

[21] M. Couprie, G. Bertrand, "Topological gray-scale watershed transform," in *Proc. of SPIE Vision Geometry V*, 1997, vol. 3168, pp. 136–146.

[22] T. Ong and H. Chen, "Updateable PAT-Tree Approach to Chinese Key Phrase Extraction using Mutual Information: A Linguistic Foundation for Knowledge Management," in *Proc. the Second Asian Digital Library Conference*, 1999, pp. 63–84.

[23] H. Lee, Y. Li, C. Chung, and L. Lee, "Enhancing Query Expansion for Semantic Retrieval of Spoken Content with Automatically Discovered Acoustic Patterns," in *ICASSP*, 2013

[24] J. Hershey and P. Olsen, "Approximating the Kullback Liebler Divergence between Gaussain Mixture Models" in *ICASSP*, 2007, vol. 4, pp. 317–320.

[25] F. Metze, N. Rajput et al., "The spoken web search task at Mediaeval 2011," in *ICASSP*, 2012, pp. 5165–5168.

[26] 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 *ICASSP*, 2012, pp. 5157–5160.

[27] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *ICASSP*, 2006.

[28] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *InterSpeech*, 2007, pp. 2385–2388.

[29] M. Terao, T. Koshinaka, S. Ando, R. Isotani, and A. Okumura, "Open vocabulary spoken-document retrieval based on query expansion using related web documents," in *InterSpeech*, 2008, pp. 2171–2174.

[30] W. Shen, C. M. White, and T. J. Hazen, "A comparison of query-by example methods for spoken term detection," in *InterSpeech*, 2009, pp. 2143–2146.

[31] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From hmms to segment models: A unified view of stochastic modeling for speech recognition," IEEE Transactions on Speech and Audio Processing, vol. 4, pp. 360V-378, 1995.

[32] Y. Zhang and J. R. Glass, "A piecewise aggregate approximation lowerbound estimate for posteriorgram-based dynamic time warping," in *InterSpeech*, 2011, pp. 1909–1912.

[33] M.-W. Koo, C.-H. Lee, and B.-H. Juang, "Speech recognition and utterance verification based on a generalized confidence score," IEEE Transactions on Speech and Audio Processing, vol. 9, no. 8, pp. 821V-832, 2001.

[34] Y. Tsao, H. Sun, H. Li, and C.-H. Lee, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition," in *ICASSP*, 2010, pp. 4422–4425.

[35] E. Voorhees, "Overview of TREC 2006," in *TREC*, 2006.