ENHANCING AUTOMATICALLY DISCOVERED MULTI-LEVEL ACOUSTIC PATTERNS CONSIDERING CONTEXT CONSISTENCY WITH APPLICATIONS IN SPOKEN TERM DETECTION

Cheng-Tao Chung[#1]*, Wei-Ning Hsu*[*2]*, Cheng-Yi Lee*[*3]*, and Lin-Shan Lee*[#4]

Graduate Institute of Electrical Engineering, National Taiwan University[#]
Department of Electrical Engineering, National Taiwan University[*]
b97901182@gmail.com[1], mhng1580@gmail.com[2], chenyi2229@gmail.com[3], lslee@gate.sinica.edu.tw[4]

ABSTRACT

This paper presents a novel approach for enhancing the multiple sets of acoustic patterns automatically discovered from a given corpus. In a previous work it was proposed that different HMM configurations (number of states per model, number of distinct models) for the acoustic patterns form a two-dimensional space. Multiple sets of acoustic patterns automatically discovered with the HMM configurations properly located on different points over this two-dimensional space were shown to be complementary to one another, jointly capturing the characteristics of the given corpus. By representing the given corpus as sequences of acoustic patterns on different HMM sets, the pattern indices in these sequences can be relabeled considering the context consistency across the different sequences. Good improvements were observed in preliminary experiments of pattern spoken term detection (STD) performed on both TIMIT and Mandarin Broadcast News with such enhanced patterns.

Index Terms— zero-resourced speech recognition, unsupervised learning, acoustic patterns, hidden Markov models, spoken term detection

1. INTRODUCTION

Supervised training of HMMs for large vocabulary continuous speech recognition (LVCSR) relies on not only collecting huge quantities of acoustic data, but also obtaining the corresponding transcriptions. Such supervised training methods yield adequate performance in most circumstances but at high cost, and in many situations such annotated data sets are simply not available. This is why substantial effort [1][2][3][4][5][6][7] has been made for unsupervised discovery of acoustic patterns from huge quantities of acoustic data without annotation, which may be easily obtained nowadays. For some applications such as Spoken Term Detection (STD) [8][9][10][11][12] in which the goal is simply to match and find some signal segments, the extra effort of building an LVCSR system using corpora with human annotations is very often an unnecessary burden [13][14][15][16][17]. M ost effort of unsupervised discovery of acoustic patterns considered only one level of phoneme-like acoustic patterns. However, it is well known that speech signals have multi-level structures including at least phonemes and words, and such structures are very helpful in analysing or decoding speech [12]. In a previous work, we proposed to discover the hierarchical structure of two-level acoustic patterns, including subword-like and word-like patterns. A similar two-level framework was also developed recently [18]. In a more recent attempt [19], we further proposed a framework of discovering multi-level acoustic patterns with varying model granularity. The different pattern HMM configurations (number of states per model, number of distinct models) form a two-dimensional model granularity space. Different sets of acoustic patterns with HMM model configurations represented by different points properly distributed over this two-dimensional space are complementary to one another, thus jointly capture the characteristics of the corpora considered. Such a multi-level framework was shown to be very helpful in the task of unsupervised spoken term detection (STD) with spoken queries, because token matching can be performed with pattern indices on different levels of signal characteristics, and the information integration across multiple model granularities offered the improved performance.

In this work, we further propose an enhanced version of the multi-level acoustic patterns with varying model granularity by considering the context consistency for the decoded pattern sequences within each level and across different levels. In other words, the acoustic patterns discovered on different levels are no longer trained completely independently. We try to "relabel" the pattern sequence for each utterance in the training corpora considering the context consistency within and across levels. For a certain level, the context consistency may indicate that the realizations of a certain pattern should be split into two different patterns, while the realizations of another two patterns should be merged. In this way the multi-level acoustic patterns can be enhanced.

2. PROPOSED APPROACH

2.1. Pattern Discovery for a Given Model Configuration

Given an unlabeled speech corpus, it is not difficult for unsupervised discovery of the desired acoustic patterns from the corpus for a chosen hyperparameter set ψ that determines the HMM configuration (number of states per model and number of distinct models) [2][4][5][6][20]. This can be achieved by first finding an initial label ω_0 based on a set of assumed patterns for all observations in the corpus χ as in (1) [6]. Then in each iteration t the HMM parameter set θ_t^{ψ} can be trained with the label ω_{t-1} obtained in the previous iteration as in (2), and the new label ω_t can be obtained by pattern decoding with the obtained parameter set θ_t^{ψ} as in (3).

$$\omega_0 = \text{initialization}(\chi), \quad (1)$$

$$\theta_t^{\psi} = \arg\max_{\theta^{\psi}} P(\chi|\theta^{\psi}, \omega_{t-1}), \quad (2)$$

$$\omega_t = \arg\max_{\omega} P(\chi|\theta_t^{\psi}, \omega). \quad (3)$$

The training process can be repeated with enough number of iterations until a converged set of pattern HMMs is obtained.

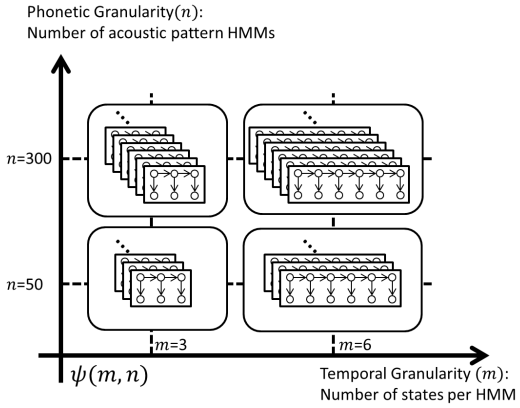

Fig. 1: Model granularity space for acoustic pattern configurations

2.2. Model Granularity Space for Multi-level Pattern Sets

The above process can be performed with many different HMM configurations, each characterized by two hyperparameters: the number of states m in each acoustic pattern HMM, and the total number of distinct acoustic patterns n during initialization, $\psi = (m, n)$. The transcription of a signal decoded with these patterns can be considered as a temporal segmentation of the signal, so the HMM length (or number of states in each HMM) m represents the temporal granularity. The set of all distinct acoustic patterns can be considered as a segmentation of the phonetic space, so the total number n of distinct acoustic patterns represents the phonetic granularity. This gives a two-dimensional representation of the acoustic pattern configurations in terms of temporal and phonetic granularities as in Fig. 1. Any point in this two-dimensional space in Fig. 1 corresponds to an acoustic pattern configuration. Note that in our previous work [19], the effect of the third dimension, the acoustic granularity which is the number of Gaussians in each state, was shown to be negligible, thus here we simply set the number of Gaussians in each state to be 4 in all cases. Although the selection of the hyperparameters can be arbitrary in this two-dimensional space, here we only select M temporal granularities and N phonetic granularities, forming a two-dimensional array of $M \times N$ hyperparameter sets in the granularity space.

2.3. Pattern Relabeling Considering Context Consistency

Context constraints successfully explored in language modeling can be used here for relabeling the acoustic patterns as shown by an example in Fig. 2. We assume the patterns 'b' and 'B' are similar without context as in Fig. 2(a). However if the context is considered, we may observe from the corpus that many realizations of pattern 'b' is preceded by pattern 'a' and followed by pattern 'c', while most realizations of pattern 'B' have different context. Therefore by relabeling all realizations of pattern 'B' which are preceded by pattern 'a' and followed by pattern 'c' as pattern 'b', the contrast between patterns 'b' and 'B' can be enhanced during the next iteration of acoustic model update as shown in Fig. 2(b) since the borderline cases have been resolved. As shown in Fig. 2(c), this relabeling includes both

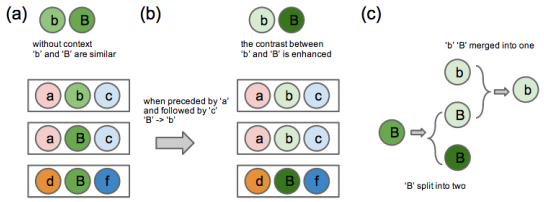

Fig. 2: Pattern relabeling considering context consistency

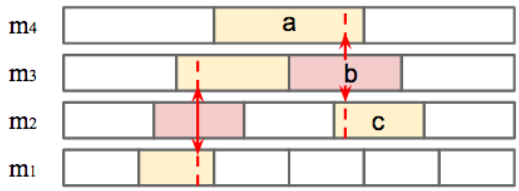

Fig. 3: Local smoothing considering granularity context

pattern splitting and merging, since the realizations of pattern 'B' are split into two patterns 'B' and 'b', while some realizations of pattern 'B' are merged into pattern 'b'. The example here considers the context in time, but can be generalized to context in model granularities as explained below.

As shown in Fig. 3, assuming an utterance is decoded into four different pattern sequences using four sets of patterns with neighboring temporal granularity $m_4 > m_3 > m_2 > m_1$, i.e., pattern HMMs with different lengths. Considering a realization of pattern 'b' of temporal granularity m_3, we find its central frame belongs to the realization of pattern 'a' of temporal granularity m_4 and the realization of pattern 'c' of temporal granularity m_2. So patterns 'a' and 'c' are taken as the context of pattern 'b' in neighboring temporal granularities. The same could be done for phonetic granularity.

2.4. Pattern Relabeling Method

Let $\omega(m_k, n_k, l)$ be the index for a decoded acoustic pattern at time l within an utterance in the corpus χ using the acoustic pattern set with the granularity $\psi(m_k, n_k)$. The relabeled pattern $\overline{\omega}(m_k, n_k, l)$ is then as in (4a) i.e., the pattern among all patterns in the set of $\psi(m_k, n_k)$ which maximizes the product of the three probabilities in (4b)(4c)(4d) evaluated with the context respectively in l, n, and m. The first probability $P_l(w)$ in (4b) for context in time l is actually the product of forward bigram and backward bigram well known in language modeling. The other two probabilities $P_n(w)$, $P_m(w)$ in (4c)(4d) are exactly the same, except n_{k-1}, n_{k+1} and m_{k-1}, m_{k+1} are the neighboring values of n and m.

$$\overline{\omega}(m_k, n_k, l) = \arg\max_w (P_l(w) P_n(w) P_m(w)), \tag{4a}$$

$$P_l(w) = P(w|\omega(m_k, n_k, l\text{-}1)) P(w|\omega(m_k, n_k, l\text{+}1)), \tag{4b}$$

$$P_n(w) = P(w|\omega(m_k, n_{k-1}, l)) P(w|\omega(m_k, n_{k+1}, l)), \tag{4c}$$

$$P_m(w) = P(w|\omega(m_{k-1}, n_k, l)) P(w|\omega(m_{k+1}, n_k, l)). \tag{4d}$$

Finer patterns and coarser patterns are drastically different in terms of perplexity; shorter patterns and longer patterns produce very different pattern sequences in terms of duration. They are complementary to each other, but we only consider the context consistency among the neighboring granularity configurations as in (4). This relabeling is performed on every decoded sequence of the $M \times N$ pattern sets considered. Katz smoothing [21] was applied to deal with unseen pattern bigrams. On the boundary of the granularity configurations or time sequences, the bigram probability is taken as 1.

2.5. Pattern Enhancement by Re-estimation after Relabeling

The relabeling in (4a) can be inserted into the recursive process of discovering the patterns in each iteration in (2)(3), as shown in (5)(6).

$$\overline{\omega}_t = \arg\max_{\overline{\omega}} P(\overline{\omega}|\omega_t), \tag{5}$$

$$\theta^{\psi}_{t+1} = \arg\max_{\theta^{\psi}} P(\chi|\theta^{\psi}, \overline{\omega}_t). \tag{6}$$

When an iteration is completed as in (2)(3), a new set of patterns is generated as in (2), with which a new set of labels is obtained as in (3). The new labels ω_t in (3) is then relabeled with (4a) based on the new labels ω_t on all different HMM sets to produce a slightly better label $\overline{\omega_t}$ as in (5). This slightly better label $\overline{\omega_t}$ is then used in (6) to generate a slightly better model set θ^{ψ}_{t+1}. Note that (6) is almost the same as (2), except here based on the slightly better label $\overline{\omega_t}$ obtained in (5). In this way the relabeling process can be repeatedly applied in every iteration, and the patterns can be enhanced by the relabeling process during the model re-estimation. Although it is theoretically possible to consider the optimization process in (3) and (5) jointly in a single step, such as maximizing the product of the two probabilities in the right hand sides of (3) and (5), practically such a joint optimization is computationally unfeasible. Therefore this is done in two separate steps here.

2.6. Spoken Term Detection

There can be various applications for the acoustic patterns presented here. In this section we summarize the way to perform spoken term detection [19]. Let $\{p_r, r = 1, 2, 3, .., n\}$ denote the n acoustic patterns in the set of ψ=(m, n). We first construct a similarity matrix S of size $n \times n$ off-line for every pattern set ψ=(m, n), for which the element $S(i, j)$ is the similarity between any two pattern HMMs p_i and p_j in the set.

$$S(i, j) = \exp(-\mathrm{KL}(i, j)/\beta. \quad (7)$$

The KL-divergence $\mathrm{KL}(i, j)$ between two pattern HMMs in (7) is defined as the symmetric KL-divergence between the states based on the variational approximation [22] summed over the states. To transform the KL divergence into a similarity measure between 0 and 1, a negative exponential was applied [23] with a scaling factor β. When β is small, similarity between distinct patterns in (7) approaches zero, so (7) approaches the delta function $\delta(i, j)$. β can be determined with a held out data set, but here we simply set it to 100.

In the on-line phase, we perform the following for each entered spoken query q and each document (utterance) d in the archive for each pattern set ψ=(m, n). Assume for a given pattern set a document d is decoded into a sequence of D acoustic patterns with indices $(d_1, d_2, ..., d_D)$ and the query q into a sequence of Q patterns with indices $(q_1, ..., q_Q)$. We thus construct a matching matrix W of size $D \times Q$ for every document-query pair, in which each entry (i, j) is the similarity between acoustic patterns with indices d_i and q_j as in (8) and shown in Fig. 4(a) for a simple example of $Q = 3$ and $D = 6$, where $S(i, j)$ is defined in (7),

$$W(i, j) = S(d_i, q_j). \quad (8)$$

It is possible to consider the N-best pattern sequences rather than the one-best sequences here by considering the posteriorgram vectors based on the N-best sequences for d, q and integrate them in the matrix W. However, previous experiments showed that the extra improvements brought in this way is almost negligible, probably because the $M \times N$ different pattern sequences based on the $M \times N$ different pattern sets can be considered as a huge lattice including many one-best paths which will be jointly considered here [19].

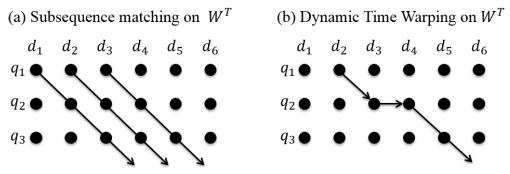

Fig. 4: The matching matrix W

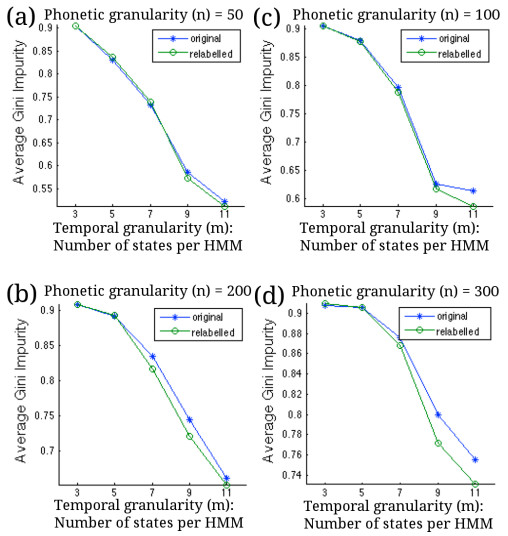

Fig. 5: Average Gini impurity for the top 20 words with the highest counts in the TIMIT training set based on the original patterns (blue) and those after relabeling (green), with different values of n and m.

For matching the sub-sequence of d with q, we sum the elements in the matrix W in (9) along the diagonal direction, generating the accumulated similarities for all sub-sequences starting at all pattern positions in d as shown in Fig. 4(a). The maximum is selected to represent the relevance between document d and query q on the pattern set ψ=(m, n) as in (9).

$$R(d, q) = \max_i \sum_{j=1}^{Q} W(i + j, j). \quad (9)$$

It is also possible to consider dynamic time warping (DTW) on the matrix W as shown in Fig. 4(b). However, previous experiments showed that the extra improvements brought in this way is almost negligible, probably because here we have jointly considered the $M \times N$ different pattern sequences based on the $M \times N$ different pattern sets (e.g. including longer /shorter patterns), so the different time-warped matching and insertion/deletion between d and q is already automatically included [19].

The $M \times N$ relevance scores $R(d, q)$ in (9) obtained with $M \times N$ pattern sets ψ=(m, n) are then averaged and the average scores are used in ranking all the documents for spoken term detection. It is also possible to learn the weights for different pattern sets to produce better results using a development set. But here we simply assume the detection is completely unsupervised without any annotation, and all pattern sets are equally weighted [19].

3. EXPERIMENTS

3.1. Purity in Pattern Sequences for known Words

In order to evaluate the quality of the acoustic patterns we discovered with varying temporal and phonetic granularities, we use the Gini impurity for the pattern sequences found for known high frequency words, since this can be evaluated for any given pattern set. Assume all the realizations of a high frequency word (e.g. the word "water") are decoded into I different pattern sequences, each occupying a percentage f_i of the realizations ($\Sigma_i f_i = 1$), we can evaluate the Gini impurity [24] for the word using the I percentages f=$\{f_i, i$=1,2,...$I\}$

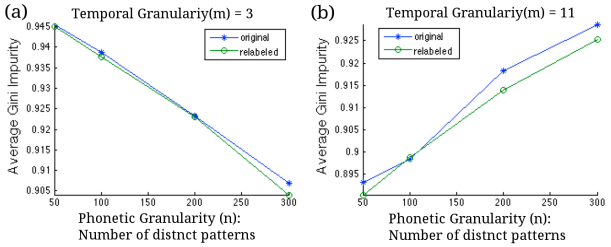

Fig. 6: Average Gini impurity for the cluster of words with occurrence counts ranging from 16 to 22 with (a) m=3, (b) m=11.

as in (10),

$$\text{Gini Impurity}(f) = \sum_{i=1}^{I} f_i(1 - f_i). \tag{10}$$

Gini impurity falls within the interval $[0, 1)$, reaches zero when all the realizations are decoded into the same pattern sequence, and becomes larger when the distribution is less pure. We trained the above different sets of patterns with m=3, 5, 7, 9, 11 and n=50, 100, 200, 300 on the TIMIT training set. Fig. 5 shows the average Gini impurity for the top 20 words with the highest occurrence counts in TIMIT training set, based on the original patterns (blue) and those after relabeling (green) for all cases considered. We see the impurity was in general high for such automatically discovered patterns because the realizations of the same phoneme produced different speakers were possibly decoded as different patterns, and the insertion/deletion inevitably increased the impurity. Although the impurity was high, the relabeling proposed here generated better patterns. We see the difference was more significant for larger m. Because the temporal variation is easily captured by models with short patterns (m=3 or 5 with high impurity) which increases the impurity, much lower impurity was achieved with longer patterns (m=9 or 11).

Another set of results for average Gini impurity for the cluster of words with occurrence counts ranging from 16 to 22 in the TIMIT training set is shown in Fig. 6 for m=3 and 11 states per HMM with varying number of distinct patterns (n). It is still quite clear that the relabeling process enhanced the patterns, and it is interesting to note that the trends for m=3 and 11 are quite different (Fig. 5(a) and (b)). As mentioned above, the temporal variation is easily captured by models with short patterns which increases the impurity (e.g. m=3 in Fig. 6(a)) so increasing the number of patterns (n) helped reduce the impurity. However, when the models are long enough (e.g. m=11 in Fig. 6(b)), larger number of patterns(n) gives more redundant patterns which caused confusion during decoding, so the impurity went up with larger n. These results indicate that the different sets of patterns of different model granularities were complementary to each other. Note that only high frequency words with enough realizations can be used or the impurity evaluation here to show the quality of the patterns. But how these patterns can be applied to spoken term detection will be shown below, for which the queries are usually low frequency words, whose impurity is difficult to evaluate.

3.2. Unsupervised Spoken Term Detection

We conducted two separate query by example spoken term detection experiments on two spoken archives. In the first experiment, the TIMIT training set was used as the spoken archive and the spoken query set consisted of 16 words randomly selected from the TIMIT testing set. In the second experiment, the spoken archive was 4.5 hours of Mandarin Broadcast News segmented into 5034 spoken documents and the spoken query set was 10 words selected from another development set. In either case, a spoken instance of a query word was randomly selected from the data set, and used as the spoken query to search for other instances in the spoken archive. The conventional 39 dimensional MFCC features were used for the HMMs. 20 sets of acoustic patterns were generated for TIMIT with m = 3, 5, 7, 9, 11 and n = 50, 100, 200, 300; 9 sets for the Mandarin Broadcast News with m = 3, 7, 13 and n = 50, 100, 300; all with 4 Gaussian mixtures per state. We compared $\overline{\omega}(m, n)$ with $\omega(m, n)$ for each (m, n) pair. We used the mean average precision (MAP) [25][26] as the performance measure, a higher value implies better performance.

Search methods(MAP)	TIMIT	Mandarin
(a) frame-based DTW on MFCC	10.16%	22.19%
(b) proposed: original patterns	26.32%	23.38%
(c) proposed: relabeled patterns	28.26%	24.50%

Table 1: Overall spoken term detection performance in mean average precision.

The MAP performance of each of the 20 pattern sets for TIMIT and 9 sets for Mandarin Broadcast News before and after relabeling is in Fig. 7(a)(b) where the performance was clearly boosted for most of the pattern sets. A paired sample t-test was used to check the MAP improvement of relabeled pattern sets, $t(28)$=3.37, p=0.0011, significant improvement was observed. Note that different from TIMIT which had many different speakers, the Mandarin Broadcast News was produced by a limited number of anchors, so MAP for each pattern set ranged between 18% to 22%, much higher than TIMIT. Although the MAP for each individual pattern set was relatively low on TIMIT (1% to 5%) in general, much better results in MAP can be obtained when all of them are jointly considered as rows (b)(c) in Table 1. Row (a) in Table 1 was the frame-based dynamic time warping (DTW) on MFCC sequences. We see the relabeled patterns achieved an MAP of 28.26% and 24.50% which is significantly better than that using the original patterns (26.32% and 23.38%). Further more, both of them significantly outperformed the baseline (10.16% and 22.19%), which proved the improvement was non-trivial.

4. CONCLUSION

In this work, we propose a method for improving the quality of multi-level acoustic patterns discovered from a target corpus. By incorporating context consistency in time and model granularity, a more consistent set of patterns can be obtained. This is verified with improved performance in spoken term detection on TIMIT and Mandarin Broadcast News.

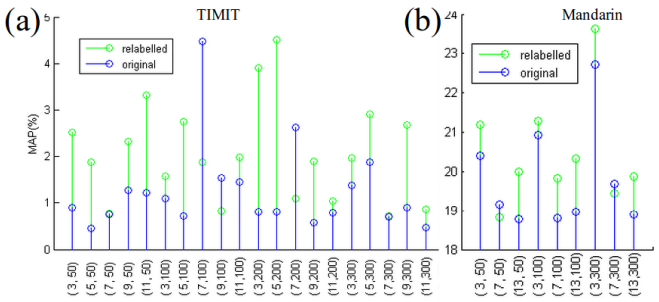

Fig. 7: mean average precision of each of the HMM sets with the granularity hyperparameters (m,n) on (a) TIMIT and (b) Mandarin Broadcast News.

5. REFERENCES

[1] Haipeng Wang, Tan Lee, Cheung-Chi Leung, Bin Ma, and Haizhou Li, "A graph-based gaussian component clustering approach to unsupervised acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[2] Aren Jansen and Kenneth Church, "Towards unsupervised training of speaker independent acoustic models.," in *INTERSPEECH*, 2011, pp. 1693–1692.

[3] Chia-ying Lee and James Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[4] Herbert Gish, Man-hung Siu, Arthur Chan, and William Belfield, "Unsupervised training of an hmm-based speech recognizer for topic classification.," in *INTERSPEECH*, 2009, pp. 1935–1938.

[5] Man-Hung Siu, Herbert Gish, Arthur Chan, and William Belfield, "Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision.," in *INTERSPEECH*, 2010, pp. 2838–2841.

[6] Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8081–8085.

[7] Chun-an Chan, Cheng-Tao Chung, Yu-Hsin Kuo, and Lin-shan Lee, "Toward unsupervised model-based spoken term detection with spoken queries without annotated data," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[8] Murat Saraclar and Richard Sproat, "Lattice-based search for spoken utterance retrieval," *Urbana*, vol. 51, pp. 61801, 2004.

[9] David RH Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A Lowe, Richard M Schwartz, and Herbert Gish, "Rapid and accurate spoken term detection.," in *INTERSPEECH*, 2007, pp. 314–317.

[10] Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan, "Vocabulary independent spoken term detection," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 615–622.

[11] Roy G Wallace, Robert J Vogt, and Sridha Sridharan, "A phonetic search approach to the 2006 nist spoken term detection evaluation," 2007.

[12] Yi-cheng Pan and Lin-shan Lee, "Performance analysis for lattice-based speech indexing approaches using words and subword units," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1562–1574, 2010.

[13] Florian Metze, Nitendra Rajput, Xavier Anguera, Marelie Davel, Guillaume Gravier, Charl Van Heerden, Gautam V Mantena, Armando Muscariello, Kishore Prahallad, Igor Szoke, et al., "The spoken web search task at mediaeval 2011," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5165–5168.

[14] Aren Jansen and Benjamin Van Durme, "Indexing raw acoustic features for scalable zero resource search.," in *INTERSPEECH*, 2012.

[15] Yaodong Zhang and James R Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[16] Marijn Huijbregts, Mitchell McLaren, and David van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4436–4439.

[17] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5157–5160.

[18] Oliver Walter, Timo Korthals, Reinhold Haeb-Umbach, and Bhiksha Raj, "A hierarchical system for word discovery exploiting dtw-based initialization," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 386–391.

[19] Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee, "Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[20] Mathias Creutz and Krista Lagus, "Unsupervised models for morpheme segmentation and morphology learning," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, no. 1, pp. 3, 2007.

[21] Stanley F Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 310–318.

[22] John R Hershey and Peder A Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–317.

[23] Marcin Marszałek and Cordelia Schmid, "Constructing category hierarchies for visual recognition," in *Computer Vision–ECCV 2008*, pp. 479–491. Springer, 2008.

[24] Leo Breiman, "Technical note: Some properties of splitting criteria," *Machine Learning*, vol. 24, no. 1, pp. 41–47, 1996.

[25] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[26] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims, "A support vector method for optimizing average precision," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 271–278.