ZERO RESOURCE SPOKEN AUDIO CORPUS ANALYSIS

David F. Harwath[1,2]*, Timothy J. Hazen*[1]*, and James R. Glass*[2]

[1]MIT Lincoln Laboratory, Lexington, MA
[2]MIT CSAIL, Cambridge, MA

ABSTRACT

Zero-resource speech processing involves the automatic analysis of a collection of speech data in a completely unsupervised fashion without the benefit of any transcriptions or annotations of the data. In this paper, our zero-resource system seeks to automatically discover important words, phrases and topical themes present in an audio corpus. This system employs a segmental dynamic time warping (S-DTW) algorithm for acoustic pattern discovery in conjunction with a probabilistic model which treats the topic and pseudo-word identity of each discovered pattern as hidden variables. By applying an Expectation-Maximization (EM) algorithm, our system estimates the latent probability distributions over the pseudo-words and topics associated with the discovered patterns. Using this information, we produce acoustic summaries of the dominant topical themes of the audio document collection.

Index Terms— Zero-resource speech processing, spoken term discovery, speech summarization.

1. INTRODUCTION

1.1. The Zero Resource Setting

Current state-of-the-art speech recognition systems typically rely on statistical models that require both a large amount of language specific knowledge and a sizable collection of transcribed data. These resources are required for training statistical models that map acoustic observations to phonetic units, creating pronunciation dictionaries mapping phonetic units to words, and estimating language models to provide constraints on the possible sequences of words. Recently in the speech community, there has been a push towards developing increasingly unsupervised, data-driven systems which are less reliant on linguistic expertise. One of the scenarios detailed by [6] is the zero resource learning problem: spoken audio data is available in a specific language, but transcriptions, annotations and prior knowledge for this language are all unavailable. In this scenario completely unsupervised learning techniques are required to learn the properties of the language and build models that describe the spoken audio.

In essence, the ultimate goal of zero-resource modeling is to develop completely unsupervised techniques that can learn the elements of a language's speech hierarchy solely from untranscribed audio data. This includes the set of acoustic phonetic units, the subword structures such as syllables and morphs, the lexical dictionaries of words and their pronunciations, as well as higher level information about the syntactic and semantic elements of the language. This is an extremely lofty goal, but recent research has begun to investigate solutions to sub-problems at various levels of the hierarchy.

One area of research focuses on the automatic discovery of repeated acoustic patterns in a spoken audio collection. The patterns that are found typically correspond to commonly repeated words or phrases observed in the data. Initial work in this area used a segmental dynamic time warping (S-DTW) algorithm search for repeated acoustic patterns in academic lectures [18]. Improvements to this approach were obtained when raw acoustic features were replaced with model-based posteriorgram features derived from a Gaussian mixture model [23]. Recent techniques for dramatically reducing the computational costs of the basic search have made this acoustic pattern discovery approach feasible on large corpora [13, 14, 15, 22].

Another approach is to first learn acoustic-phonetic models from the audio data. These phonetic units are then used to represent the data before performing any higher level pattern discovery. Approaches of this type include a self-organizing unit (SOU) recognition system which learns an inventory of phone-like acoustic units in an unsupervised fashion [5], a successive state splitting hidden Markov model framework for discovering sub-word acoustic units [21], and a Bayesian nonparametric acoustic segmentation framework for unsupervised acoustic model discovery [17]. Clustered patterns from a spoken term discovery system have also been used to help unsupervised learning of acoustic models [12].

Independent of the speech technology work being pursued in this area, researchers in linguistics and cognitive science have been interested in the process of language acquisition and have been developing techniques that attempt to learn words by segmenting a collection of phoneme strings. Bayesian approaches have proven to be especially successful for this task [7, 16].

The successful application of the aforementioned algorithms opens the doors for higher level semantic analysis. In [9], n-gram counts of unsupervised acoustic units were used to learn a latent topic model over spoken audio documents. In [3], vector space document modeling techniques were applied to the clustered patterns found by a spoken term discovery algorithm. In [4, 25], similar spoken term discovery algorithms were used to produce acoustic summaries of spoken audio data.

1.2. Spoken Corpus Summarization

Suppose we would like to understand the major topical themes within a collection of speech audio documents, without having to listen to each one. If text transcripts or ASR output for each document were available, topic models from Probabilistic Latent Semantic Analysis (PLSA) [11] or Latent Dirichlet Allocation (LDA) [1] could be used to generate a text summary of the corpus as in [8]. In the zero resource setting, these techniques cannot be directly applied. We instead present a method that is similar in spirit, but aims to summarize the topical themes of the corpus by extracting

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

meaningful audio snippets.

For the purpose of generating this kind of summary, we want to associate regions of the audio signal with latent topics, analogous to what is done with words in models such as PLSA and LDA. In this paper, we propose a system which:

1. Searches the audio corpus for repeated acoustic patterns, often corresponding to repetitions of the same word or phrase.
2. Uses a pair of probabilistic latent variable models to associate these acoustic patterns with latent topics and pseudo-words.
3. Summarizes the topical themes of the corpus by using the models to extract topically meaningful acoustic patterns.

In section 2, we describe the unsupervised pattern discovery stage; in section 3, we present the graphical models and the estimation procedure; in section 4, we explain how we generate acoustic summaries of the topics in the corpus; section 5 presents experimental results, and section 6 concludes.

2. SPOKEN TERM DISCOVERY

Let $D = \{d_1, d_2, \ldots, d_{|D|}\}$ be a collection of spoken audio documents. The entry point for our analysis is to apply an S-DTW like spoken term discovery algorithm to the entire collection of audio with the aim of discovering a set of low distortion match fragments. Ideally, each match fragment provides an alignment between two distinct regions of the audio signal which share the same or a similar underlying text transcription.

In our work, we utilize the fast two-pass approximate search algorithm introduced by [13]. Rather than representing the acoustic signal with posteriorgrams derived from a supervised phone classifier as in [13] or an unsupervised GMM [23], we use the Self Organizing Unit (SOU) system described in [20] and [9]. The SOU system learns a set of phone-like acoustic models in a data-driven and completely unsupervised fashion; from start to finish, our entire system requires absolutely no labeled data. For the purposes of spoken term discovery, each 10ms audio frame of an utterance is represented by an SOU posterior vector, and pairwise frame similarities are computed using the inner product between these vectors.

The output of the spoken term discovery step is a set of matches, M. Each element of M is a triple consisting of a distortion score and two matched regions of audio, $(t_1^{(a)}, t_2^{(a)})$, and $(t_1^{(b)}, t_2^{(b)})$. To remove spurious and short matches, we filter out any match with a distortion score greater than 0.5 or an average length of less than 0.5 seconds. We require a means of collapsing overlapping regions into a single interval so as to resolve when one region of audio matches multiple other regions of audio. We use a method of doing this introduced in [3] that collapses overlapping regions to the same interval whenever their fractional overlap exceeds a threshold set to 0.75. The result is a collection of intervals, where each interval consists of one or more match regions which overlap in time. For each interval i, we choose the start time, $t_1^{(i)}$, and end time, $t_2^{(i)}$ by averaging the start and end times of all regions collapsed to i. Each interval i also inherits the links associated with all match regions that overlap it; we assign i a link set, $L_i = \{l_{i,1}, l_{i,2}, \ldots, l_{i,|L_i|}\}$, where each $l \in L_i$ takes on as its value the index of some other interval j such that there exists a match in M linking a region of audio overlapping i with a region of audio overlapping j. For each interval, this yields a triple $i = (t_1^{(i)}, t_2^{(i)}, L_i)$. After this process, we are left with a set of $|I|$ intervals, $I = \{i_1, i_2, \ldots, i_{|I|}\}$, with the subset of intervals appearing in document d denoted by I_d. A visual representation of a linked spoken audio document collection is shown in Figure 1.

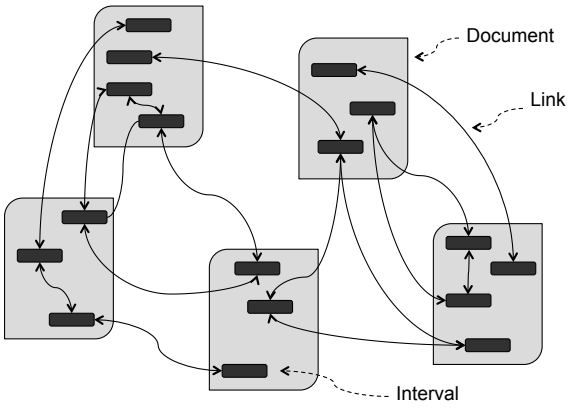

Fig. 1. An example of a linked audio document corpus.

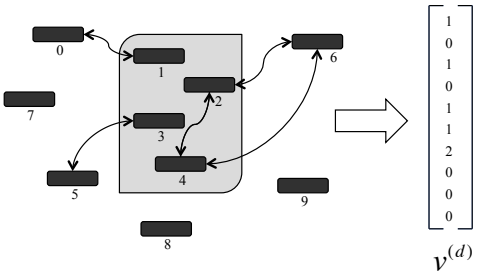

Fig. 2. Bag-of-links representation of an audio document.

3. MODELING TOPICS AND PSEUDO-WORDS

At a high level, our goal is to characterize the document collection D in terms of a set of latent topics Z, in the same spirit as algorithms such as PLSA and LDA applied to text documents. We draw inspiration from these text-based document models, but what differentiates our data from text is the fact that we do not know the word-level transcription underlying each interval of audio discovered by the spoken term discovery algorithm. In this section, we present two latent variable models which aim to capture the topical themes of a spoken audio document collection in the absence of any lexical knowledge.

3.1. PLSA on Bags-of-Links (PLSA-BoL)

This model treats each document as a bag-of-links vector $v^{(d)}$, where the j^{th} element of $v^{(d)}$ is equal to the total number of times any interval contained in d matched the j^{th} interval. That is,

$$v_j^{(d)} = \sum_{i \in I_d} \mathbf{1}_{L_i}(j) \tag{1}$$

where $\mathbf{1}_{L_i}(j) = 1$ if interval i matched the j^{th} interval and 0 otherwise. This idea for a corpus consisting of 10 match intervals is illustrated in Figure 2. We seek to model the probability of observing a link to interval l from document d using a set of latent topic variables Z:

$$\Pr(l|d) = \sum_{z \in Z} \Pr(l|z) \Pr(z|d). \tag{2}$$

The graphical model in plate notation is shown in Figure 3, and is in fact equivalent in structure to PLSA. Note that this model assumes

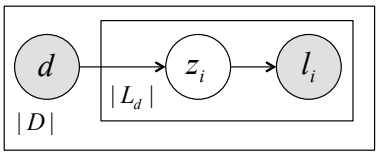

Fig. 3. The PLSA-BoL model in plate notation

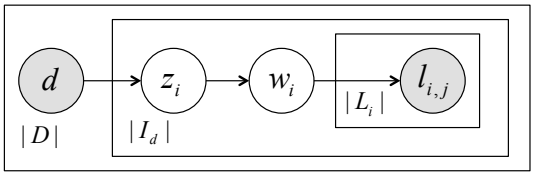

Fig. 4. The Latent Lexical and Topic Model in plate notation

that each link originating from d was generated by a different latent topic variable, even if several of these links originate from the same interval of audio within d.

The model parameters, $\Pr(z|d)$ and $\Pr(l|z)$ are estimated using the standard Expectation-Maximization update equations for PLSA [11]. We also apply TF-IDF based stop listing to the bag-of-links vectors, throwing away any interval which was linked to by more than 20% of the documents, or less than 4 other intervals. Hierarchical agglomerative clustering of the documents into $|Z|$ clusters is used in the initialization of $\Pr(l|z)$.

3.2. Latent Lexical and Topic Model (LLTM)

While the PLSA-BoL model has the capability to associate links to particular intervals of audio with latent topics, it is not able to infer which intervals of audio may be instances of the same spoken word or phrase. We consider a second model which assumes that each match interval has a latent word identity $w \in W$, where W is a fixed-size vocabulary of pseudo-words to be learned. In this model, we assume the following generative story for each link set L_i belonging to interval i in document d:

1. Draw a latent topic z from $\Pr(z|d)$.
2. Draw a latent pseudo-word w from $\Pr(w|z)$.
3. Draw a set of $|L_i|$ links to other intervals i.i.d. from $\Pr(l|w)$.

The probability of observing a link set L_i given a document can be expressed mathematically as

$$\Pr(L_i|d) = \sum_{w \in W} \sum_{z \in Z} \Pr(w|z) \Pr(z|d) \prod_{l \in L_i} \Pr(l|w). \quad (3)$$

To find a local maximum of the data likelihood surface, we employ an Expectation-Maximization algorithm. In the E-step, we estimate the joint posterior probability distribution of the pseudo-word variable w and latent topic variable z for interval i appearing in document d according to

$$\Pr(w, z|d, L_i) \propto \Pr(w|z) \Pr(z|d) \prod_{l \in L_i} \Pr(l|w). \quad (4)$$

In the M-step, we use the last estimate of this posterior to update the model parameters according to the equations

$$\Pr(l|w) \propto \sum_{d \in D} \sum_{i \in I_d} \mathbf{1}_{L_i}(l) \sum_{z \in Z} \Pr(w, z|d, L_i) \quad (5)$$

$$\Pr(w|z) \propto \sum_{d \in D} \sum_{i \in I_d} \Pr(w, z|d, L_i) \quad (6)$$

$$\Pr(z|d) \propto \sum_{i \in I_d} \sum_{w \in W} \Pr(w, z|d, L_i). \quad (7)$$

Agglomerative clustering of the documents is again used to determine the initial assignment of the topic variable associated with each interval. The $\Pr(l|w)$ and $\Pr(w|z)$ distributions are initialized by pseudo-word category assignments produced by the InfoMap graph clustering algorithm [19] applied to the graph formed by treating the intervals as nodes and their links as edges.

4. SUMMARIZING THE TOPICS

The parameters of the models presented in section 3 provide us with a means of summarizing the topical content of an audio corpus by extracting a small set of audio intervals containing words or phrases representative of the discovered latent topics. A human user could listen to these sets of audio snippets and quickly get a gist of what topics are discussed in the collection. For the purposes of ranking intervals and pseudo-words against latent topics, we use a weighted pointwise mutual information measure:

$$WPMI(x, z) = \Pr(x, z)^{\lambda} \log \left(\frac{\Pr(x, z)}{\Pr(x) \Pr(z)} \right). \quad (8)$$

Intuitively, the $\log(\cdot)$ factor is large when x and z are more likely to appear together than independently, and the $\Pr(x, z)$ factor weights this by the overall joint probability of x and z. λ acts as a tuning parameter to trade off between the factors.

To form a summary of latent topic z using the PLSA-BoL model, we rank all of the match intervals according to $WPMI(i, z)$ with $\lambda = 1$ and extract the top 10 audio intervals. Using the LLTM, we first rank the pseudo-word categories according to $WPMI(w, z)$ with $\lambda = 0.5$ to choose a representative set of 10 pseudo-words for each latent topic. We then extract the interval of audio most representative of each pseudo-word; to do this, we rank the intervals according to $\Pr(i|w) \Pr(w|d, L_i)$. Here, $\Pr(w|d, L_i)$ represents the posterior probability that interval i belongs to pseudo-word category w, while $\Pr(i|w)$ indicates how likely any other interval belonging to pseudo-word category w is to generate a link to interval i.

5. EXPERIMENTS

For our summarization experiments, we use a collection of 60 telephone calls from the English Phase 1 portion of the Fisher Corpus [2]. Each call consists of a 10-minute long telephone conversation between two speakers. At the start of each conversation the participants were prompted to discuss a particular topic. The set of calls we use spans 6 of these topic prompts, with 10 calls per prompt. As an example, the prompt for the "Anonymous Benefactor" topic is:

> "If an unknown benefactor offered each of you a million dollars - with the only stipulation being that you could never speak to your best friend again - would you take the million dollars?"

SOU posteriorgram representations for all utterances in all 60 calls were produced by an 45-unit SOU system trained on an independent 60-hour set of Fisher English data. S-DTW audio segment link detection was applied to the posteriorgram representation of all utterance pairs in the 60 call set. A total of 10,041 link pairs between 3165 unique audio intervals were discovered and used to train

Topic	Text transcripts of extracted intervals	Mapping to true topics (%)
1	minimum wage, minimum wage, minimum wage, minimum wage, . . .	Minimum Wage (99.7)
2	think computers, computers, of computers, computer, computer, computers, . . .	Computers in Education (99.9)
3	exactly, um, country, exactly, countries, um, countries, exactly, exactly	Illness (37.3), Corporate Conduct (32.2)
4	holidays, holiday, holiday is, holidays, the holidays, holidays, holiday, . . .	Holidays (83.1)
5	money, situations, situations, the more money you, friend, educational, four years, situation, situations, make money	Anonymous Benefactor (55.3)
6	weather friends, friends, friends, friends, friends, some friends, friends, kind of friends, to happen, major you know	Corporate Conduct (55.2), Anonymous Benefactor (45.3)

Table 1. Latent topic summaries generated using PLSA-BoL.

Topic	Text transcripts of extracted intervals	Mapping to true topics (%)
1	don't think, weather friends, no I, situations, the lottery, very you know, and, benefactor, don't even know who, and, economy, to happen, now um, money, so	Anonymous Benefactor (45.0), Corporate Conduct (27.3)
2	minimum wage, you, yeah I, money, out you'd be, you know, minimum wage jobs, you know people, the, economy, in New York, an hour, five dollars, he has, people working	Minimum Wage (86.1)
3	think computers, if she uses, education, more and, computers, you ah, technical ah, know the computerized, it's just, information, something that's, on there, well that that's, different things, school	Computers in Education (99.7)
4	sicker, C.E.O., stock market, exactly, without the, country, every sick, this guy, in uh in, of cold, like if you, that um, greedy, Zealand you, stomach	Illness (47.7), Corporate Conduct (43.5)
5	I really like, holidays, own holiday, holiday, equality, favorite holiday, considerate, and, the key, recognized, you, new car, keys, like, you like	Holidays (78.8)
6	is actually, I'm twenty, friend, <partial>, I've seen it done, every day, maybe ah, that and all, how, uh, best friend, that's true, increased, children, lazier and	Anonymous Benefactor (72.9)

Table 2. Latent topic summaries generated using LLTM.

our latent topic models, all of which learned a set of 6 latent topics. Additionally, the number of pseudo-word categories in the LLTM was set to 581 by the InfoMap algorithm.

Tables 1 and 2 show summaries of the latent topics learned by the PLSA-BoL and LLTM models, as represented by the text transcripts for the top scoring audio intervals extracted for each topic. Also shown is the mapping between latent topic z and the closest matching true Fisher topic label t according to $\Pr(t|z)$. To evaluate the mapping between the latent topics and the true topics, we use the normalized mutual information (NMI) measure:

$$NMI(z,t) = \frac{2 * I(z;t)}{H(z) + H(t)} \tag{9}$$

Here $I(\cdot;\cdot)$ denotes mutual information and $H(\cdot)$ denotes entropy. NMI is an information theoretic measure similar to the F-score measure used in detection problems. Its value ranges between 0 and 1, with 1 representing a perfect mapping between the true topics and the latent topics. Table 3 shows the NMI scores for a uniform random assignment of documents to latent topics, the hard agglomerative clustering used for initialization, both latent models, and a phrase-based PLSA model applied to text transcripts of the data [10].

Both models do a surprisingly good job of learning latent topics with a strong mapping to the true topics, given the fully unsupervised nature of the system. It is interesting to note that the intervals extracted by the first model tend to be repeated instances of the same word or phrase; while the extracted intervals are almost always topically relevant, the summaries are somewhat lacking in variety. The second model largely overcomes this issue. Although the LLTM summaries produced are dominated by topically indicative words and phrases, some "stop words" are also present in the summaries. Text-based topic models often utilize expertly crafted stop-lists which alleviate this problem, and finding a comparable solution in the zero resource setting is a worthy avenue for future work.

	Rand.	HAC	PLSA-BoL	LLTM	PLSA-Text
NMI	0.168	0.529	0.529	0.592	0.895

Table 3. NMI scores for the various models

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a system for completely unsupervised zero-resource learning of topics present in a collection of spoken audio documents. We have shown how the models presented in this paper can be used to produce extractive summaries of the latent topics by choosing representative snippets of audio which often correspond to topically meaningful words and phrases. Experiments conducted on a set of topic-prompted telephone calls from the Fisher Corpus have demonstrated the feasibility of the approach.

There are many avenues of future work for these methods. Parallelizing the model estimation procedure on a multicore system or a graphics processor unit using techniques similar to [22] would allow for larger audio corpora to be analyzed. The use of new unsupervised and semi-supervised acoustic models [17, 24] may prove to be useful for improving the performance of the spoken term discovery procedure. Finally, reworking the models presented in this paper into a fully Bayesian framework would expand the flexibility of the models and allow their size to automatically scale with the amount of data used.

7. ACKNOWLEDGMENTS

The authors would like to thank Herbert Gish and Man-Hung Siu of Raytheon BBN Technologies for contributing the SOU data used in our experiments.

8. REFERENCES

[1] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research* vol. 3, pp. 993-1022, 2003.

[2] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generation of speech-to-text," in *Proc. of International Conf. on Language Resources and Evaluation,* Lisbon, May 2004.

[3] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. of EMNLP,* 2010.

[4] R. Flamary, X. Anguera, and N. Oliver, "Spoken wordcloud: Clustering recurrent patterns in speech," in *Proc. of CBMI,* 2011.

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

[6] J. Glass, "Towards Unsupervised Speech Processing," Keynote, *Proc. ISSPA*, Montreal, July 2012.

[7] S. Goldwater, T. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: exploring the effects of context," *Cognition,* vol. 112 pp. 21-54, 2009.

[8] T. Hazen, "Latent topic modeling for audio corpus summarization," in *Proc. of Interspeech,* Florence, August 2011.

[9] T. Hazen, M. Siu, H. Gish, S. Lowe, and A. Chan, "Topic modeling for spoken documents using only phonetic information," in *Proc. of ASRU,* 2011.

[10] T. Hazen and F. Richardson, "Modeling multiword phrases with constrained phrase trees for improved topic modeling of conversational speech," in *IEEE Spoken Language Technology Workshop,* Miami, December 2012.

[11] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Conf. on Uncertainty in Artificial Intelligence,* Stockholm, 1999.

[12] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Proc. of Interspeech,* Florence, 2011.

[13] A. Jansen, K. Church and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Proc. of Interspeech,* Makuhari, September 2010.

[14] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. of ASRU,* 2011.

[15] A. Jansen and B. Van Durme, "Indexing raw acoustic features for scalable zero resource search," in *Proc. of Interspeech,* 2012.

[16] M. Johnson, "Unsupervised word segmentation for Sesotho using adaptor grammars," in *Proc. ACL SIG on Computational Morphology and Phonology,* 2008.

[17] C. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. of ACL,* Jeju, 2012.

[18] A. Park and J. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 16, no. 1, pp. 186-197, 2008.

[19] M. Rosvall and C.T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. of National Academy of Science, USA,* vol. 105 pp. 1118-1123, 2008.

[20] M. Siu, H. Gish, S. Lowe, A. Chan, "Unsupervised audio pattern discovery using HMM-based self-organized units," in *Proc. of Interspeech,* 2011.

[21] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proc. of ACL-08 HLT, Short Papers,* pp. 165-168, 2008.

[22] Y. Zhang, K. Adl, and J. Glass, "Fast spoken query detection using lower-bound dynamic time warping on graphical processing units," in *Proc. of ICASSP*, Kyoto, March 2012.

[23] Y. Zhang and J. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. of ICASSP,* Dallas, March 2010.

[24] Y. Zhang, R. Salakhutdinov, H. Chang, and J. Glass, "Resource configurable spoken query detection using deep Boltzmann machines," in *Proc. of ICASSP,* Kyoto, March 2012.

[25] X. Zhu, G. Penn, and F. Rudzicz, "Summarizing multiple spoken documents: finding evidence from untranscribed audio," in *Proc. of ACL,* Singapore, August 2009.