Topic Modeling for Spoken Documents Using Only Phonetic Information

Timothy J. Hazen[#1], Man-Hung Siu[*], Herbert Gish[*], Steve Lowe [*], Arthur Chan[*]

[#] *MIT Lincoln Laboratory*
Lexington, Massachusetts, USA

[*] *Raytheon BBN Technologies*
Cambridge, Massachusetts, USA

***Abstract*—This paper explores both supervised and unsupervised topic modeling for spoken audio documents using only phonetic information. In cases where word-based recognition is unavailable or infeasible, phonetic information can be used to indirectly learn and capture information provided by topically relevant lexical items. In some situations, a lack of transcribed data can prevent supervised training of a same-language phonetic recognition system. In these cases, phonetic recognition can use cross-language models or self-organizing units (SOUs) learned in a completely unsupervised fashion. This paper presents recent improvements in topic modeling using only phonetic information. We present new results using recently developed techniques for discriminative training for topic identification used in conjunction with recent improvements in SOU learning. A preliminary examination of the use of unsupervised latent topic modeling for unsupervised discovery of topics and topically relevant lexical items from phonetic information is also presented.**

I. Introduction

While processing based on word-based automatic speech recognition (ASR) outputs represents a dominant paradigm for spoken language processing (SLP), such ASR systems typically require large amounts of transcribed training data to yield accurate results. However, not all SLP applications require word-level ASR outputs to perform adequately. Various applications exist that have successfully used only phonetic-level information including spoken term detection [1], automatic document retrieval [2], topic identification [3], and automatic language identification [4].

For some tasks, it is not even necessary to possess a phonetic recognizer in the language(s) of interest. It has been shown that automatic language identification (LID) can be performed based on the outputs of multiple language dependent phonetic recognizers [4] or on the output of a single universal phonetic recognizer [5]. For accurate LID it is often sufficient to have training data marked only with language labels without requiring any additional transcriptions of the training data.

In this paper we examine the problems of topic identification (topic ID) and unsupervised topic modeling and lexical discovery. In the topic ID task, like the LID task, our goal is to train an accurate classification system using training data which only contains topic labels for each audio document without the benefit of any manually generated word transcripts. In the unsupervised topic modeling and lexical discovery task, we seek to automatically learn lexical and topical information from a collection of audio data that is completely unlabeled.

Various previous studies have investigated the problem of topic ID using only phonetic speech recognition outputs [6], [7], [8], [9]. Our own previous work in this area has not only demonstrated successful topic ID based on the output of a same-language phonetic recognition system, but also the viability of topic ID using a cross-language phonetic recognizer, i.e., a recognizer trained for a completely different language [3], or topic ID using parametric trajectory mixture models learned directly from the acoustic signal [10]. In more recent work, we have shown that topic ID can also be applied to the output of a phonetic tokenizer constructed from self-organizing units (SOUs), i.e., phone-like units learned in a completely unsupervised fashion from untranscribed acoustic data [11], [12]. These studies have verified the viability of topic ID based purely on phonetic information, even for situations where no transcriptions are available to train a same-language phonetic recognizer in a supervised fashion.

In this paper, we extend our previous investigations of topic modeling techniques that use only phonetic information. In a set of topic ID experiments, we compare the use of a same-language phonetic recognizer, a cross-language phonetic recognizer, and SOU phonetic tokenizers. We also introduce the use of new discriminative training techniques that have yielded improvement on word-based topic ID to the phonetically-based topic ID problem. Additionally, we examine new modeling approaches for training a phonetic tokenizer based on SOUs. Our experiments will show that topic ID performance using our latest SOU system is approaching the accuracy levels achievable when using a same-language phonetic recognizer trained with full supervision. This paper also presents preliminary investigations we have conducted in the area of completely unsupervised lexical and topic discovery from unannotated audio data.

[1]This work was sponsored by the Air Force Research Laboratory 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

II. Phonetically-Based Topic ID

A. Phonetic Recognition and Feature Extraction

The first stage of our topic ID system is the application of automatic phonetic recognition to each segment of audio in each audio document. The phonetic recognizer generates a lattice of phonetic hypotheses for each audio segment. Pruning is employed during recognition to remove the highly unlikely phonetic hypotheses from the recognition lattices. The posterior probability of occurrence for each phonetic hypothesis is computed for every arc in every lattice. From these posterior lattices a posterior probability for any phonetic sequence through the lattice can also be estimated. A total estimated occurrence count for each phonetic n-gram sequence over the course of the entire audio document is then computed by summing the individual posterior scores over all instances of each n-gram sequence over all lattices for that document. In our work we generally use estimated counts of triphone sequences as the set of features for topic ID.

Mathematically, every audio document d is expressed as:

$$d = \{c_1, c_2, \ldots, c_{N_V}\} \tag{1}$$

Here each c_f is the occurrence count of a specific feature f, where f is a specific phonetic n-gram from the full set V of N_V unique phonetic n-gram features used by the system.

B. Topic Classification

1) Support Vector Machines: For classification we use a support vector machine (SVM) classifier in conjunction with a minimum classification error (MCE) training approach for discriminatively learning feature weights. Though we will summarize our classification approach below, full details for our training algorithms can be found in [13].

In our SVM approach, each document d is represented by a feature vector $\vec{x}$. Each vector $\vec{x}$ is created using *term frequency - log likelihood ratio* (TF-LLR) normalization of the underlying feature counts [14]. This can be expressed as:

$$x_f = \frac{P(f|d)}{\sqrt{P(f)}} \tag{2}$$

Here $P(f|d)$ is estimated from document d as follows:

$$P(f|d) = \frac{c_f}{\sum_i c_i} \tag{3}$$

$P(f)$ is estimated from the full collection of training documents using MAP estimation as follows:

$$P(f) = \frac{N_f + 1}{N_F + N_V} \tag{4}$$

Here, N_f is the estimated occurrence count of feature f over the entire training set, N_F is the sum of estimated occurrence counts over all features in the training set, and N_V is the number of unique features in the feature vocabulary V.

To create a multi-class SVM for N_T different topics, a one-vs.-rest SVM classifier is independently learned for each topic t expressed by the following scoring function:

$$S(\vec{x}, t) = -b_t + a_t \sum_{\forall i} \omega_{i,t} K(\vec{v}_i, \vec{x}) \tag{5}$$

Here, each vector $\vec{v}_i$ is a unique training vector or *support vector* from the set of training documents. Each $\omega_{i,t}$ is a learned support vector weight for training vector i for topic t. The function $K(\vec{v}, \vec{x})$ is an SVM *kernel function* for comparing the vectors $\vec{v}$ and $\vec{x}$. Our system uses a linear kernel function which simply computes the dot product between between the two component vectors:

$$K(\vec{v}, \vec{x}) = \vec{v} \cdot \vec{x} \tag{6}$$

The a_t and b_t values in Eqn. 5 represent class specific scale and offset values. In the multi-class SVM scenario, the settings of the a_t and b_t over all topics t are jointly calibrated to optimize the multi-class classification error rate. This calibration process is performed using an MCE calibration procedure described in [13].

2) MCE Feature Weight Training: The SVM classifier described above is a linear classifier that defines one linear separating hyperplane per class in the vector space. For each individual topic t, the binary SVM can be expressed as:

$$S(\vec{x}, t) = \vec{r}_t \cdot \vec{x} - b_t \tag{7}$$

Here, $\vec{r}_t$ is a linear projection vector and b_t is a score offset. The set of projection vectors $\vec{r}_t$ over all t can be concatenated to form a matrix R, and the score offsets b_t can be concatenated to form a vector $\vec{b}$, thus creating a multi-class linear classifier expressed as:

$$\vec{s} = \begin{bmatrix} S(\vec{x}, t_1) \\ \vdots \\ S(\vec{x}, t_{N_T}) \end{bmatrix} = R\vec{x} - \vec{b} \tag{8}$$

This linear classifier can be augmented to include a discriminatively trained set of feature weights as follows:

$$\vec{s} = R(\vec{\lambda} * \vec{x}) - \vec{b} \tag{9}$$

Here, the $*$ operator performs a term-wise multiplication of vectors $\vec{\lambda}$ and $\vec{x}$ where each element λ_f of $\vec{\lambda}$ is a feature weight applied to the corresponding feature x_f. In our previous work we have demonstrated how these feature weights can be learned with a leave-one-out version of minimum classification error (MCE) training to improve the performance of a trained multi-class SVM [13].

III. Self Organizing Units (SOUs)

A. Unsupervised HMM

For supervised training of an HMM for speech recognition, its parameters are typically specified by the acoustic model parameters, θ_{am}, and the language model parameters, θ_{lm}. For notational conveniences in this paper, we group them as a single parameter set $\theta = [\theta_{am}, \theta_{lm}]$. Then, the ML parameter estimation finds the parameter set, $\hat{\theta}_{sup}$, that maximizes the joint likelihood, $p(A, W|\theta)$, of acoustic observation sequence A and the label sequence W. That is,

$$\hat{\theta}_{sup} = \arg\max_{\theta} p(A, W|\theta). \tag{10}$$

In the case of unsupervised training in which the label sequence W is not known, we maximize the joint likelihood by searching not only over the model parameters but also all possible label sequences. That is, W becomes a variable to be optimized. The unsupervised ML parameter estimation becomes,

$$\hat{\theta}_{unsup} = \arg\max_{\theta}\max_{W} p(A, W|\theta), \quad (11)$$

$$= \arg\max_{\theta}\max_{W} p(A|W, \theta)p(W|\theta) \quad (12)$$

The maximization over both the label sequence and the acoustic model likelihood in Eqn. 12 balances the acoustic likelihood and label sequence structure.

Eqn. 11 maximizes over two sets of variables, θ and W, which can be performed using iterative maximization. At each iteration, one set of variables is fixed while the other set is maximized. Then we alternate between them. So, at the i-th iteration, the two maximization steps are:

1) find the best parameter set θ_i on the previously found label sequence W_{i-1}.

$$\theta_i = \arg\max_{\theta} p(A, W_{i-1}|\theta). \quad (13)$$

2) find the best word sequence W_i by using the previously estimated parameter set θ_i.

$$W_i = \arg\max_{W} p(A, W|\theta_i), \quad (14)$$

Comparing Eqns. 10 and 13, it is obvious that Step 1 (Eqn. 13) is simply the regular supervised HMM training (both acoustic and language models) using the newly obtained transcription W_{i-1} as reference. Finding the best word sequence in the second step would suggest a Viterbi recognition pass. Although recognition is usually viewed as finding the most likely label sequence over the posterior probability, $p(W|X, \theta)$, it is easy to show that the same sequence also maximizes the joint likelihood $p(X, W|\theta)$ as in Eqn. 14. So, Eqn. 14 expresses the recognition of a new transcription using the updated parameters θ_i.

B. Initialization

We explored two different SOU initializations: phoneme recognition based, and unsupervised segmental Gaussian mixture model (SGMM) based. For phoneme recognition based initialization, an off-the-shelf phoneme recognizer (potentially from a different language) transcribes the training data with the resulting recognition outputs used as the initial transcription for HMM training. For SGMM-based initialization, audio is first segmented based on its spectral discontinuities which are learned without supervision from the audio signal [15]. It is followed by fitting each audio segment with a polynomial (quadratic) trajectory in the cepstral space. The audio segments are then grouped into clusters of similar acoustics based on the distance between their polynomial trajectory parameters [16], [17]. The distance measure currently used on a pair of segments is the area between their polynomial trajectories. These segment clusters represent collections of sound units. Any individual cluster is a collection of variants of a particular sound and forms the basis for generating a SGMM with each mixture component representing a segment cluster. The SGMM is trained with the EM algorithm. The SGMM becomes a speech tokenizer when, for an audio segment, it returns the mixture index by which the segment likelihood is maximized. After building the SGMM, it is used as a tokenizer for the training segments. These segment labels form the initial transcription for HMM training.

C. SOU HMM Training

We use the state-of-the-art BBN Byblos recognizer [18] for HMM training. Byblos includes advanced signal processing techniques, such as Vocal Tract Length Normalization (VTLN), Heteroscedastic Linear Discriminant Analysis (HLDA) feature transformation, context-dependent triphone and quinphone models, multi-pass recognition, speaker adaptive training etc. Byblos uses "flat start" HMM training that does not require token time marks. Instead, iterative alignment and model estimation are carried out. While discriminative training is part of Byblos, our current experiments used only the maximum likelihood training. Details about the Byblos training can be found in [18]. In addition to acoustic models, initial bigram and trigram language models are constructed using the label sequences generated by the segmental tokenizer. Following Eqns. 13 and 14, we iteratively maximize the model likelihood and find the best label sequence.

Non-crossword models trained with SOU as words does not capture any contexts but these are used in our first two passes of tokenization. This limits the resolution of the acoustic models in these early passes. To remedy this, we generate multi-unit compound words by combining frequent bi- and tri-units into compound words during each iteration of the HMM training. This effectively builds up long multi-unit SOUs progressively.

D. Tokenization

With the trained acoustic models and language models, the tokenization of audio into SOU sequences is no different from regular phoneme recognition. To create context dependent models in Byblos, we need to create "phoneme" classes to drive the decision-tree based phoneme-state clustering. To produce "linguistic questions" to drive the decision tree-based state clustering, We cluster the 64 SGMM's into 16 classes to act as "phoneme classes". Because we used the phonemes (or SOUs) as words, true context modeling occurs only in cross-word models. To use context-based model, recognition lattices are re-scored using cross-word models.

E. Different SOU Systems

For our experiments, we have built 3 SOU systems using different initializations and amounts of training data. The first two systems used single SOUs as units and were trained with 15 hours of data from the Fisher Corpus [19]. The first system, denoted as SGMM-15, was initialized with SGMM. The second system, denoted as BUT-15, was initialized with

recognized Hungarian phonemes generated by a phonetic recognition system created at the Brno University of Technology (BUT) [20]. The third system, denoted as SGMM-120, was initialized with SGMM and trained with 120 hours of Fisher data. In addition to more data, a set of 1106 frequent multi-SOU sequences up to 6 units long, which we refer to as pseudo-words, were added into the dictionary in SGMM-120. Our preliminary experiments have shown that the use of pseudo-words are very effective for exploiting additional training data.

Other than the amounts of training data and compound word units, all three systems were built with the same configurations. In training, 5 iterations of SOU training were performed. Multi-pass recognition was performed on all systems with the final lattices generated with unsupervised speaker adapted, non-crossword models and rescored by crossword models.

IV. Topic ID Experiments

A. Corpus

Our topic ID experiments were performed on data from the English Phase 1 portion of the Fisher Corpus [19]. During the Fisher data collection, two participants were connected over the telephone network and instructed to discuss a specific topic for 10 minutes. Data was collected from a set of 40 different prompted topics. For our experiments, a set of 1374 topic-labeled conversations serve as the topic ID training set, and an independent set of 1372 conversations are used for the topic ID test set. There is no overlap between these sets and the Fisher data used for SOU training.

B. Phonetic Recognition Systems

Our experiments used five different phonetic recognition systems for processing the audio data. The baseline English system used the BUT phonetic recognizer trained on 10-hours of telephone speech from the Switchboard cellular corpus. Within the lattices produced on the topic ID training set, this recognizer produced 86,407 unique triphones for use in the topic ID feature set. For our cross-language experiment, we used a Hungarian version of the BUT recognizer trained on 10 hours of read Hungarian telephone speech. This system yielded 161,442 unique triphone features for topic ID.

The primary focus of our experiments is the use of the three SOU tokenizers discussed in Section III-E. The SOU SGMM-15 system used 64 base phone units yielding 187,854 triphone sequences for its topic feature set. The SOU BUT-15 system used 44 base phone units and yielded 38,071 unique triphones during recognition.

The SOU SGMM-120 system used 64 base units in conjunction with 1106 pseudo-words (which were treated as individual units when computing n-gram counts). As a result, SGMM-120 yielded over 18 million unique trigram sequences (with most of these features being longer than 3 base units in length). To reduce the number of features to a more manageable number we only computed bigram counts over the pseudo-word units produced by the SGMM-120 system. Even when only using bigrams over the pseudo-words, the number of unique features produced by SGMM-120 system was 890,540.

TABLE I
Topic ID results for using a standard SVM approach and an SVM with MCE feature weighting (FW) as applied to phonetic features extracted from five different phonetic recognition systems.

Phonetic Recognizer	# n-gram Features	Classification Error Rate	
		SVM	SVM+FW
BUT Hungarian	161,442	45.7	40.9
SOU SGMM-15	187,854	42.1	33.7
SOU BUT-15	38,071	37.5	30.7
SOU SGMM-120	890,540	35.1	26.7
BUT English	86,407	19.9	17.9

C. Results

Table I shows our topic ID results using the five recognizers described above. Results are presented using both a standard multi-class SVM approach, and the SVM approach that incorporates our MCE feature weight (FW) training. There are two primary observations that can be made about the results in Table I. First, MCE feature weight training produces relative reductions in error rate of between 10% (for the BUT English system) and 24% (for the SOU SGMM-120 system) over the baseline SVM. Second, the SOU systems all perform better than the cross-language BUT Hungarian system. This demonstrates that untranscribed in-language data can be used effectively to produce models trained in an unsupervised fashion. In fact, the SOU SGMM-120 system using pseudo-words has closed over 60% of the gap between the cross-language BUT Hungarian system and the BUT English system (which had transcribed data available for supervised training).

V. Unsupervised Lexical and Topical Discovery

A. Background

The results we reported in Section IV-C are on a traditional topic ID problem with supervised learning which requires a manual labeling of the topics present in the training data. A related problem is that of completely unsupervised learning of lexical items and topics. In recent years there have been several research efforts geared toward the unsupervised discovery of lexical items directly from acoustic data without the benefit of any lexical and topical annotation. The goal is to automatically discover repeated acoustic patterns corresponding to lexical items within an audio corpus. These include efforts to discover acoustic patterns within various representations of the acoustic signal including frame-based mel-frequency scale cepstral coefficients (MFCCs) [21], frame-based posteriorgrams generated form Gaussian mixture models (GMMs) [22], self-organizing units [12], or frame-based posteriorgrams generated from an English phonetic recognizer [23].

One technical challenge for pattern discovery is learning to distinguish patterns that correspond to topically important content words from patterns that corresponds to words or word sequences that possess little or no topical relevance (e.g., “excuse me”, “you know”, etc.). In [12], it was shown that long

SOU sequences that correspond to important topical words can be discovered when information about the topic labels is available. In [24], common multi-phone sequences were discovered automatically as a pre-processing stage for clustering documents into a topic hierarchy based on the counts of discovered sequences. Our goal is to *jointly* discover both the topics and the important words or phrases corresponding to these topics in an unsupervised fashion.

B. PLSA Modeling

We have performed a pilot study to examine the utility of applying probabilistic latent semantic analysis (PLSA) [25] to the output of a phonetic recognizer. The end goal is to help guide existing lexical discovery techniques towards the speech regions most likely to contain topically relevant phonetic sequences. PLSA learns a set of N_Z latent topics $Z = \{z_1, \ldots, z_{N_Z}\}$ each possessing a probabilistic model $P(f|z)$ for generating features. PLSA also learns a model $P(z|d)$ for mapping each document d to a distribution of topics $z \in Z$.

C. PLSA Experimental Conditions

Our pilot experiment was conducted using the output of the BUT English phonetic recognizer applied to the 1374 document Fisher training set. Each document is represented by the raw estimated occurrence counts of triphones. PLSA was applied to the training set feature vectors to learn a set of latent topic models in an unsupervised fashion. Our pilot experiment sets the number of latent topics to be 40 (the actual number of prompted topics in the Fisher Corpus).

Word-based PLSA models are often aided by manually crafted word stop-lists, i.e., a list of non-content-bearing words (e.g., articles, conjunctions, etc.) that are ignored during processing. Statistical stop-listing can be applied in the phonetic modeling approach to achieve the same effect. In our experiments, PLSA modeling ignores both rare triphone features (those with occurrence counts of 5 or less in the data set) and extremely common triphone features (those estimated to have appeared in $> 50\%$ of the documents). This stop-listing removed 1754 triphones with high document-frequency and 37008 triphones with low term-frequency, leaving 47646 triphone features for PLSA modeling.

D. Learning Topics

When PLSA is performed on word-level transcriptions of the same Fisher data, it learns latent topics that align well with prompted topics [26]. One goal of our pilot study was to see whether or not PLSA could similarly learn meaningful latent topics using only triphone counts. In contrast to our word-based experiments, our PLSA experiments yielded more mixed results. An alignment of the learned PLSA models with the true prompted topic labels demonstrated that only 13 of the 40 learned PLSA topics were strongly associated with actual Fisher topics. The remaining 27 learned latent topics do not show any strong association with actual Fisher topics.

Table II shows 12 PLSA topics that align closely with actual Fisher topics. Each topic is represented by its top 3 most distinctive triphones based on the weighted point-wise mutual information (WPMI) measure. The WPMI between a feature f and a PLSA topic z is expressed as:

$$wpmi(f, z) = P(f, z) \log \frac{P(f, z)}{P(f)P(z)} \tag{15}$$

The percentage value next to each Fisher topic shows the percentage of the corresponding latent topic model that is associated with documents from that Fisher topic. Thus, 72.0% of the top latent topic model in Table II is associated with documents labeled with the "Holidays" topics. In each case, the top scoring triphones in these latent models are easily associated with words that are distinctive to the matching Fisher topic. In two cases, the concatenation of three triphones correspond exactly to title word for the associated Fisher topic (i.e, "Movies" and "Perjury").

TABLE II
A COLLECTION OF 12 PLSA TOPICS LEARNED FROM PHONETIC TRIPHONE COUNTS WITH STRONG ASSOCIATION WITH ACTUAL FISHER TOPICS. EACH TOPIC IS REPRESENTED BY ITS 3 MOST DISTINCTIVE TRIPHONE SEQUENCES (LISTED WITH THEIR ASSOCIATED WORDS).

12 PLSA Topics	
Distinctive Triphones (Associated Words)	Associated Fisher Topics (% Overlap with Associated Topic)
hh:oh:l hh:ao:l l:d:ey (holiday, holidays)	Holidays (72.0%)
w:oh:ch r:iy:aw iy:aw:l (watch, reality)	Reality TV (70.1%)
p:er:jh jh:r:iy er:jh:r (perjury)	Perjury (67.5%)
ax:d:aa ax:d:ao d:d:ao (a dog)	Pets (58.1%)
m:uw:v uw:v:iy v:iy:z (movie, movies)	Movies (55.1%)
w:ey:jh m:w:ey ow:w:ey (minimum wage)	Minimum Wage (53.1%)
w:iy:hh p:ae:t sh:iy:z (we have, pet, she's)	Pets (51.5%)
f:ae:m ae:m:l s:ae:m (family)	Family (50.8%)
uw:dx:axr m:p:y p:y:uw (computer, computers)	Computers in Education (50.0%)
w:oh:ch s:p:ao b:ao:l (watch, sports, ball)	Sports on TV (44.3%) Strikes by Athletes (16.6%)
uw:l:z t:iy:ch uw:l:s (schools, teachers)	US Public Schools (36.6%) Censorship in Schools (29.6%)
s:m:ow m:ow:k ow:k:ix (smoke, smoking)	Personal Habits (29.0%) Smoking (27.8%)

E. Discovering Important Phonetic Sequences

A global ranking of the topical importance of phonetic sequences over an entire corpus can also be obtained by aggregating the WPMI measure of a feature f over all latent topics z to yield a *total topical information* (TTI) measure as follows:

$$tti(f) = \sum_{\forall z} wpmi(f, z) \tag{16}$$

Table III shows 17 of the top 20 triphone sequences as ranked by the TTI measure and the words for which they

TABLE III
MAPPING OF TRIPHONES SEQUENCES WITH HIGH TTI RANKINGS TO ASSOCIATED WORDS AND THEIR TTI RANKINGS DETERMINED FROM A REFERENCE PLSA MODEL LEARNED FROM THE TEXT TRANSCRIPTS.

Triphones (TTI Rankings)	Associated Words (Reference TTI Rankings)
s:m:ow(1), m:ow:k(2)	smoking(20), smoke(26)
m:p:y(10), p:y:uw(7), y:uw:dx(17), uw:dx:axr(3)	computer(6), computers(11)
f:ae:m(4), ae:m:l(9)	family(1)
w:oh:ch(5), w:ah:ch(15)	watch(3)
m:w:ey(8), w:ey:jh(6)	minimum(4), wage(5)
b:ao:l(11)	baseball(39), football(40)
s:p:ao(12)	sports(14), sport(98)
m:uw:v(18), uw:v:iy(13)	movie(10), movies(21)
k:ey:sh(16)	education(66)

are predominantly associated. Each of the reference words is also shown with a reference TTI ranking obtained from a 40-topic PLSA model learned directly from the text transcripts of the same data. In this table we observed that triphone sequences from 7 of the top 11 (and also 9 of the top 21) TTI ranked reference words from the text transcripts are present in the top 20 ranked triphone sequences. This indicates that PLSA applied to the phonetic recognition outputs exhibits the capability to identify at least some of the triphone sequences corresponding to topically important words.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have extended our previous work by incorporating MCE-based feature weight training techniques for SVM classifiers to the phonetic-based topic identification problem. Significant gains in topic ID performance were observed using the MCE approach. We have also demonstrated the effectiveness of new methods for training tokenizers based on self-organizing units (SOUs). Our experiments have demonstrated that learned SOUs can be used effectively for topic ID and their performance can be significantly improved when pseudo-words are exploited.

In this paper, we have also examined the feasibility of jointly learning topics and topically important phonetic sequences in an unsupervised fashion. Our initial studies demonstrate some promising results in this area. Because of the success of our SOU-based topic ID experiment, we believe we can extend our current work by similarly applying latent topic modeling to SOUs. Additionally, we plan to extend our method for identifying topically important phonetic sequences by tying the TTI measure back to the recognition lattices. This would allow us to generate a topical importance measure that is directly associated with regions of the original acoustic data. Using this measure, lexical discovery methods (such as segment dynamic time warping) could be directed to efficiently focus their acoustic sequence matching on specific acoustic regions that are believed to contain the most topically relevant words.

REFERENCES

[1] R. Wallace, R. Vogt and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Proc. Interspeech*, Antwerp, August 2007.

[2] K. Ng, *Subword-based Approaches for Spoken Document Retrieval*, Ph.D. Thesis, MIT, February 2000.

[3] T. Hazen, F. Richardson and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Proc. IEEE ASRU Workshop*, Kyoto, December 2007.

[4] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[5] T. Hazen, "Segment-based automatic language identification," *Jour. of the Acoustical Society of America*, vol. 101, no. 4, pp. 2323–2331, 1997.

[6] J. Wright, M. Carey, and E. Parris, "Statistical models for topic identification using phoneme substrings," in *Proc. ICASSP*, Atlanta, May 1996.

[7] R. Kuhn, P. Nowell, and C. Drouin, "Approaches to phoneme-based topic spotting: An experimental comparison," in *Proc. ICASSP*, Munich, April 1997.

[8] E. Nöth, S. Harbeck, H. Niemann, and V. Warnke, "A frame and segment based approach for topic spotting," in *Proc. Eurospeech*, Rhodes, September 1997.

[9] M. W. Theunissen, K. Scheffler, and J. A. du Preez, "Phoneme-based topic spotting on the Switchboard Corpus," in *Proc. Eurospeech*, Aalborg, September 2001.

[10] W. Belfield and H. Gish, "A topic classification system based on parametric trajectory mixture models," in *Proc. Interspeech*, Geneva, September 2003.

[11] H. Gish, M.-H. Siu, A. Chan and W. Belfield, "Unsupervised training of an HMM-based speech recognition system for topic classification," in *Proc. Interspeech*, Brighton, UK, September 2009.

[12] M.-H. Siu, H. Gish, A. Chan and W. Belfield, "Improved topic classification and keyword discovery using an HMM-based speech recognizer trained without supervision," in *Proc. Interspeech*, Makuhari, September 2010.

[13] T. Hazen, "MCE training techniques for topic identification of spoken audio documents," to appear in *IEEE Trans. on Audio, Speech and Language Processing*, 2011.

[14] W. Campbell, *et al*, "Phonetic speaker recognition with support vector machines", *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. Saul and B. Schölkopf", MIT Press, Cambridge, MA, USA, 2004.

[15] J. Cohen, "Segmenting speech using dynamic programming", *Jour. of the Acoustical Society of America*, vol. 69, no. 5, pp. 1430–1437, 1981.

[16] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. ICASSP*, Minneapolis, April 1993.

[17] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *Proc. ICSLP*, pp. 466–469, Philadelphia, October 1996.

[18] S. Matsoukas et al, "Advances in transcript of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1541–1555, 2006.

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

[20] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Proc. Int. Conf. on Text, Speech and Dialogue*, Brno, September 2004.

[21] 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.

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

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

[24] C. Cerisara, "Automatic discovery of topics and acoustic morphemes from speech," *Computer Speech and Language*, vol. 23, no. 2, pp. 220-239, 2009.

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

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