Available online at www.sciencedirect.com
SciVerse ScienceDirect
Computer Speech and Language xxx (2013) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery☆

Man-hung Siu*, Herbert Gish, Arthur Chan, William Belfield, Steve Lowe

Raytheon BBN Technologies, 50 Moulton Street, Cambridge, MA 02138, United States

Received 22 November 2011; received in revised form 9 May 2013; accepted 10 May 2013

Abstract

We present our approach to unsupervised training of speech recognizers. Our approach iteratively adjusts sound units that are optimized for the acoustic domain of interest. We thus enable the use of speech recognizers for applications in speech domains where transcriptions do not exist. The resulting recognizer is a state-of-the-art recognizer on the optimized units. Specifically we propose building HMM-based speech recognizers without transcribed data by formulating the HMM training as an optimization over both the parameter and transcription sequence space. Audio is then transcribed into these self-organizing units (SOUs). We describe how SOU training can be easily implemented using existing HMM recognition tools. We tested the effectiveness of SOUs on the task of topic classification on the Switchboard and Fisher corpora. On the Switchboard corpus, the unsupervised HMM-based SOU recognizer, initialized with a segmental tokenizer, performed competitively with an HMM-based phoneme recognizer trained with 1 h of transcribed data, and outperformed the Brno University of Technology (BUT) Hungarian phoneme recognizer (Schwartz et al., 2004). We also report improvements, including the use of context dependent acoustic models and lattice-based features, that together reduce the topic verification equal error rate from 12% to 7%. In addition to discussing the effectiveness of the SOU approach, we describe how we analyzed some selected SOU *n*-grams and found that they were highly correlated with keywords, demonstrating the ability of the SOU technology to discover topic relevant keywords.

Keywords: Unsupervised training; Keyword discovery; Self-learning; Speech recognition; Topic identification

1. Introduction

The training of a speech recognizer is a way of modeling a speech corpus and as such is a way of imparting structure to the speech data. We exploit this structure by working with the tokens representing sounds or sound patterns that the recognizer produces. Typically the structure is imposed by having a transcription and a pronunciation dictionary of the speech corpus. The premise of our work is that even in the absence of a transcription and a dictionary one can effectively create models for speech applications that exploit the existing sound pattern structure of speech. The

☆ This paper has been recommended for acceptance by 'A. Potamianos'.

* Corresponding author at: 10 Moulton Street, Cambridge, MA 02138, United States. Tel.: +1 617 873 2234.
E-mail addresses: msiu@bbn.com, msiu@ieee.org (M.-h. Siu), hgish@bbn.com (H. Gish), achan@bbn.com (A. Chan), wbelfiel@bbn.com (W. Belfield), slowe@bbn.com (S. Lowe).

0885-2308/$ – see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.csl.2013.05.002

approach we take results in the development of sound units that are iteratively optimized for the acoustic domain of interest using the maximum likelihood (ML) criterion. We refer to these sound units as self-organizing units or SOUs. There is a growing interest in unsupervised training methods for speech, and HMM-based approaches are not the only option. An alternative approach for the unsupervised modeling of speech was originally developed by Park and Glass (2005) and more recently by Jansen et al. (2010) which is a non-parametric approach that is related to the Dynamic Time Warping (DTW) technology for speech recognition. Just as HMM recognizers differ significantly from DTW recognizers, similar differences exist between these two approaches.

Below we discuss the difficulties for topic classification that may be encountered if a recognizer is trained using transcribed data collected at mismatched conditions. We then focus on some of the details of the SOU recognition systems. Many spoken language processing applications, such as spoken topic classification, use automatic speech recognition (ASR) as the first stage of processing to convert speech into tokens for down-stream processing. This works well when the characteristics of the classification data match that of the recognizer, which requires transcribed training data from the domain of the classification data. In some cases, transcribed speech may not be available in the specific domain, channel or language. While it is possible to use a tokenizer from a different language or channel, the system performance will be highly dependent on the level of mis-match between the tokenizer training and the classification data.

In Hazen et al. (2007), significant degradation is reported on topic classification experiments when using a Hungarian phoneme tokenizer on English data even when both the tokenizer training and test data were on telephone channels. In addition to potential language mis-match, results from robust speech recognition suggest that mis-matches in environmental conditions, such as channel or noise could also cause significant performance degradation.

An alternative is to build an acoustic tokenizer in an unsupervised fashion. Because transcribed data are no longer needed, it is possible to train the tokenizers using the domain-specific data related to classification, and thus significantly reduce the potential for mis-match. The GMM-tokenization (Zissman, 1993) is a special case of tokenization training without transcription and has been successfully applied in language identification.

In Belfield and Gish (2003), we introduced the idea of unsupervised training of a multi-frame tokenizer. We built a topic classifier using an unsupervised segmental Gaussian mixture model (SGMM) to tokenize automatically derived multi-frame segments. This mixture of segmental models was the result of clustering multi-frame segments. The multi-frame segmental units were critical because they act like phones and their *n*-grams captured keywords which were needed for topic discrimination. However, the segmental technology is limited by its inability to benefit directly from improvements made in the more traditional HMM-based speech-to-text (STT) systems. Improvements, such as vocal tract length normalization, speaker adaptation, discriminative training, among others have been shown to improve recognition significantly but will need to be "re-invented" for segment models. On the other hand, using the traditional HMM-based recognizer as tokenizer would require transcribed training data.

To benefit both from the unsupervised training in the segmental approach, and the progress made in HMM-based speech recognition technology, we propose in this paper an unsupervised HMM training approach that jointly optimizes the observation likelihood and the label sequence without any transcribed training data. Under this framework, training can be iteratively performed using existing HMM recognition tools. The HMMs can transcribe the audio into a sequence of self-organizing speech units (SOUs) using only untranscribed speech for training, and the resulting unit sequence can be used for topic identification (TID). One significant advantage of completely unsupervised training is that there will not be any mis-match between training and test, because the untranscribed test data can, if needed, be added to acoustic training. Because SOU training uses a standard HMM training framework, it benefits from improved ASR techniques such as speaker adaptation and context modeling. If some transcribed audio is available, semi-supervised training (Lamel et al., 2002; Zavaliagkos et al., 1998; Ma and Schwartz, 2008; Li et al., 2007) can be used, which is similar to our unsupervised approach with the exception that the model would be initialized using a small set of transcribed audio.

Our first set of topic identification[1] experiments were performed on a ten-topic subset from the Switchboard corpus with performance measured by the average equal error rate (EER) across the topics. The proposed approach resulted in an EER of 12.5% that outperformed the Hungarian phoneme recognizer (16.7% EER) as well as an English phoneme

[1] We are in fact doing a mix of topic verification and classification. We apply the term "identification" in a broad sense that includes both verification and classification.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

recognizer trained with 1 h of transcribed speech (14.7% EER). With the addition of context dependent acoustic models and lattice re-scoring in SOU recognition and lattice-based n-gram SVM features for TID, our TID EER was improved from 12% to 6.6%. To better understand the TID EER, we also performed an oracle experiment using the phoneme sequence from the true transcripts which gave a TID EER of 1.2%. In addition to the Switchboard corpus, we also performed TID experiments on the Fisher corpus for comparison with published TID results.

Transcribing audio into SOUs also makes it possible to perform topic keyword discovery. In Nowell and Moore (1995), a dynamic programming approach was proposed to discover topic-relevant phoneme sequences. A similar approach has also been proposed in Park and Glass (2005) and Zhang and Glass (2010) that searches for "similar" regions of speech from in-topic audio recordings. In this paper, we discuss how to discover topic-relevant SOU-ngrams via features selected by the SVM-topic classifier.

We describe the unsupervised HMM learning algorithm in Section 2. We then describe the SOU implementation, and the downstream topic classification system in Section 3. Experimental results are reported in Section 4. We then describe our work in keyword discovery in Section 5. In Section 6, we summarize the paper and describe possible further improvements.

2. Unsupervised HMM training

Denote the HMM parameters as $\theta = [\theta_{am}, \theta_{lm}]$, which include both the acoustic model parameters, θ_{am}, and the language model parameters, θ_{lm}. In typical supervised HMM training using maximum likelihood criterion, the acoustic model likelihood, $p(X|W, \theta_{am})$, and language model likelihood $p(W|\theta_{lm})$ are denoted as two separate sets of parameters because the parameters can be easily decoupled and maximized separately.[2] To simplify our notation in this paper, we merge the two maximization into a single one over θ on the joint likelihood, $p(X, W|\theta)$ such that the maximum likelihood (ML) parameter estimation finds the parameter, $\hat{\theta}_{sup}$, that maximizes the joint likelihood of observation X and the label sequence W.

Mathematically, we can express the ML parameter estimation as

$$\hat{\theta}_{sup} = arg\max_{\theta} p(X, W|\theta), \tag{1}$$

in which both the acoustic observation X and transcription W are known in training time. If one views the representation of speech as a big network of multi-state HMMs, the word sequence simply imposes constraints on the state sequence allowable for each training utterance, and the language model probability can be captured by the transition probability between certain word end and word begin nodes. 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,

$$\hat{\theta}_{unsup} = arg\max_{\theta}\max_{W} p(X, W|\theta), \tag{2}$$

$$\hat{\theta}_{unsup} = arg\max_{\theta}\max_{W} p(X|W, \theta)p(W|\theta) \tag{3}$$

The maximization over both the label sequence (language model term) and the acoustic model likelihood in Eq. (3) balances the acoustic likelihood and label sequence structure. At one possible extreme, one can use different symbols for each frame to maximize the acoustic likelihood at the expense of high entropy in the symbol sequence. At the other extreme, one can define only one symbol to maximize the language model likelihood but this will result in low acoustic likelihood.[3] The balance between these two is influenced by the choice of the initial label sequence and the complexity of the acoustic and language models.

[2] Parameter estimation using other criteria, such as discriminative training will have different formulation.
[3] We assume the number of parameters per label is fixed.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

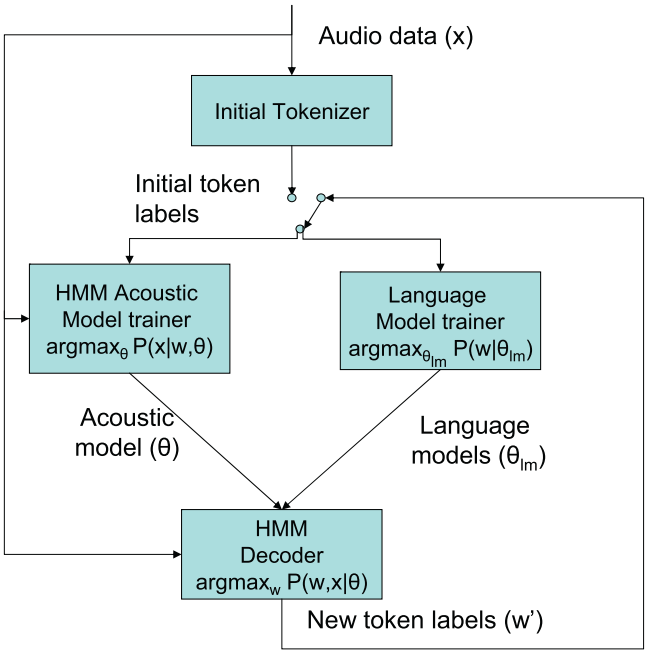

Fig. 1. Block diagram of the iterative training process.

2.1. Iterative optimization

Eq. (3) maximizes over two sets of variables, θ and W, which can be performed iteratively. At each iteration, we keep one set of variables fixed while maximizing over the other set and then alternate between them. So, at the ith iteration, the two maximization steps are:

1. Find the best parameters θ_i on the previously found label sequence W_{i-1}.

$$\theta_i = \arg\max_{\theta} p(X, W_{i-1}|\theta). \tag{4}$$

2. Find the best word sequence W_i by using the previously estimated parameters θ_i.

$$W_i = \arg\max_{W} p(X, W|\theta_i), \tag{5}$$

Comparing Eqs. (1) and (4), it is obvious that Step 1 (Eq. (4)) is simply the regular supervised HMM training (both acoustic and language model) 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, $p(W|X, \theta)$. However, it is easy to show that the same sequence also maximizes the joint likelihood $p(X, W|\theta)$ as in Eq. (5). So, Eq. (5) effectively expresses the recognition of a new transcription using the updated parameters θ_i.[4]

The iterative learning process is illustrated in Fig. 1, in which the process starts with an initial sequence of labels, followed by repeated model training and decoding. This suggests that once we have initialized the process, we can use existing recognition tools to perform unsupervised HMM training.

[4] We ignored all the approximations typically associated with practical recognition systems, such as using language modeling weights, language model smoothing, or pruning.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

Fig. 2. Block diagram of the segmental model training.

2.2. Initialization

The process can be initialized either with an initial model or an initial label sequence. Obtaining an initial model (without an initial sequence) with something like a flat start in an HMM is more difficult because there is really no information to differentiate between the units. Instead, it is easier to use another tokenizer (not necessarily an HMM) to create an initial label sequence. In many iterative maximization schemes, the quality of the initialization can have significant impact on final model quality. For the proposed unsupervised training, the initial label sequence is particularly important because it also defines the set of units to be learned. Multi-frame phone-like units are consistent with the ASR acoustic representations, and their n-grams can capture keywords that are critical for topic classification.

The use of phone-like units does not preclude us from using bigger units in recognition. Similar to supervised STT where words are mapped to a sequence of phonemes via a dictionary, one can form bigger word-like units via a mapping to a sequence of basic units.

3. Unsupervised HMM recognition and topic classification system

3.1. Training

The training process of the classification system includes three stages. The first stage creates an initial sequence for HMM training. There are multiple approaches for initializing the system with phone-like units, such as using a different recognizer trained from another language. In a sense, that is not truly unsupervised because it uses a model trained with supervision. Instead, we have taken a sequential learning approach to build a segmental tokenizer using a polynomial segmental Gaussian mixture model (SGMM). The details of the SGMM training will be described below in Section 3.1.1.

The second stage is the unsupervised HMM training that iterates between optimizing the model parameters and the label sequences. It should be noted that after initialization, only the label sequences are passed from the segmental tokenizer to the HMM trainer while other auxiliary information such as segment boundaries is ignored.

The final stage is the SVM classifier training. Again, only the label sequences are passed from the HMM recognizer.

3.1.1. SGMM tokenizer for initialization

The first component of the segmental tokenizer is an adaptive segmenter, which partitions speech at boundaries of spectral discontinuities. The quantification of discontinuity is corpus specific and is inferred by a dynamic programming algorithm with the creation of statistical models for the corpus under consideration as described in Cohen (1981). The segmentation process is immediately followed by fitting each segment with a polynomial (quadratic in our application) trajectory model, whose parameters are used to compute the pairwise distances between segments for segment clustering. The distance of a pair of segments is the total area between the trajectories of the segments under comparison. This distance is applied to a binary centroid splitting algorithm that is widely used in frame-based mixture models.

The basic idea for the segmental trajectory model is to jointly model a sequence of consecutive frames and as such model a sound instance. The process is illustrated in Fig. 2. The segmentation process cuts the audio into a sequence of variable length, multi-frame segments, each represented by a polynomial trajectory across time. The clusters of segments represent collections of sound units and any individual cluster is a collection of variants of a particular sound. We use the clusters as the basis for generating an SGMM, which is trained with the EM algorithm. Note that a segmental GMM term is a bit different from the usual Gaussian mixture model term whose mean is a fixed point in the cepstral space and is constant over time. Each term of an SGMM is a Gaussian whose mean is a vector trajectory in the cepstral feature space and varies over time to represent time varying characteristics of a sound. The mathematical basis for

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

SGMMs has been discussed previously in Gish and Ng (1993, 1996). The SGMM becomes a "speech recognizer" or tokenizer when, for a test segment, the index of the mixture term with the maximum likelihood is generated as a token.

In particular, for a segment in an utterance extending from frame i to j inclusively, we denote this segment as

$$X_i^j = [x_i, x_{i+1}, \ldots, x_j],$$

where x_t is the acoustic features for the tth frame. We define an M-component segmental Gaussian mixture model (SGMM) over the segment, X_i^j

$$f(X_i^j) = \sum_{n=1}^{M} w_n g_n(X_i^j).$$

When using SGMM as an initialization to SOU training, M would be the number of SOUs. Each component, g_n, is defined as the product of Gaussian distributions on the separate frames with a time-invariant covariance Σ_n, but the mean parameter μ_n follows a trajectory in normalized time. Specifically,

$$g_n(X_i^j) = \prod_{t=i}^{j} N(x_t; \mu_n(\tau_{t,i,j}), \Sigma_n)$$

where $\tau_{t,i,j}$ is the normalized time within the segment given by

$$\tau_{t,i,j} = \frac{t-i}{j-i}.$$

For any test segment, the symbol/token it generates in the recognition process is

$$n_{output} = arg\max_n w_n f_n(X_i^j)$$

The token sequences generated by the SGMMs are passed as the true "transcript" for HMM training.

3.1.2. SGMM alternatives for initialization

An SGMM tokenizer for initialization is one way for providing an initial transcription of the training audio in order to get the iterative training started. Alternative initializations can be achieved by using recognizers from other languages. However, while these recognizers do represent the use of transcribed audio, they can be considered as being a legacy resource. If the acoustic domain of the legacy system was similar to that of the audio at hand then it could be quite useful. An acoustic domain mismatch could reduce the utility of such a recognizer.

The SGMM is trained entirely within domain. However, it is a fairly simple model with only one Gaussian per SOU. We have looked at generalizations that include using multiple Gaussians per SOU as well as an HMM approach to initializations. This work is not yet complete and we do believe such approaches should improve performance over the current SGMM initialization.

3.1.3. HMM training

Our work uses the state-of-the-art BBN Byblos system (Matsoukas et al., 2006) to model the units derived from the SGMM. Byblos includes advanced signal processing techniques, such as vocal tract length normalization (VTLN) and heteroscedastic linear discriminative analysis (HLDA) feature transformation. Byblos training does not require time alignment. Instead, iterative alignment and model estimation are applied with the first few iterations using simpler models followed by more complex models. Because we used the phonemes (or SOUs) as words, non-crossword models are equivalent to context-independent (or mono-phone) models. To build context dependent models in Byblos, we create "phoneme" classes to drive the decision-tree based phoneme-state clustering. What "linguistic questions" can we use to drive the decision tree? In our current work, we cluster the 64 SGMM's into 16 classes to act as "phoneme classes" for decision tree context clustering of quinphones. More complex models, including cross-word models, are trained after clustering. Speaker adaptive training (SAT) is also applied to create models for unsupervised speaker adaptation.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

While discriminative training is implemented and available, our current experiments used only the maximum likelihood training.[5] Details about the Byblos training can be found in Matsoukas et al. (2006).

In addition to acoustic models, bigram and trigram language models are learned using the label sequences generated by the segmental tokenizer.

3.2. Tokenization and topic classification

With the trained acoustic models and language models, the tokenization of training (also test) audio into SOU sequences is no different from regular phoneme recognition. When an HMM is trained, many models are created and their application, as we have previously noted, is typically done with an initial set of simpler models applied to the data followed by a pass of more complex models. The multi-pass approach is for computational efficiency and is the same for an SOU system as for a phonemic system. Whether working with phonemes or SOUs, the key idea is that the sound units are capturing the syllables, words or phrases that will be important in distinguishing topics. In our earlier work (Gish et al., 2009), we tokenized the audio using only non-crossword models with bigram language models. We observed a performance improvement using the more complex cross-word quinphone model as reported in Siu et al. (2010). These quinphones are used to re-score SOU lattices generated by mono-phone models.

The step after tokenization is the SVM classifier training. For topic classification, audio cuts are tagged as in-topic or out-of-topic. Note that data used for training the classifiers do not need to be the same data used for training the HMMs but can be so if needed. We begin by tokenizing both the in-topic and out-of-topic audio into SOU sequences. We then extract from these SOU sequences the SVM classification features which are SOU n-gram statistics normalized by inverse-document-frequency (IDF). In this paper, our n-gram features are trigrams of SOUs, and a document for computing IDF is a conversation side. We then apply the more sophisticated SVM feature building techniques for generating useful long-span n-gram features (Campbell and Richardson, 2007).

The n-gram statistics can either be extracted from recognition 1-best hypotheses as in Gish et al. (2009), or more generally, from the quinphone re-scored lattices (Siu et al., 2010), which are soft counts based on the lattice posterior probabilities. However, using soft counts from lattices for classification tasks was previously introduced by other researchers (Hazen and Margolis, 2008). The use of soft counts creates a problem for estimating the IDF weighting. We experimented with soft IDF weighting as proposed in Wintrode and Kulp (2009) as compared to using a simple posterior threshold on the soft token counts. We found that the simple thresholding gave better performance.

4. Topic identification experiments

The SOU TID evaluation process is quite similar to that of a typical TID system using words. The test audio cuts are first tokenized into SOU token sequences (or lattices). Classification features are extracted and passed to the SVM topic classifier to generate TID scores. Note that instead of using the SVM as a hard decision classifier, we extracted the SVM scores, each of which is a function of the distance of the test sample to the decision boundary, as the classification scores in our computation of TID EER.

4.1. Experiment setup

In this paper, we report two sets of experiments on topic identification on two different corpora.

Our first set of experiments was performed on the Switchboard-1 corpus, which consists of telephone conversations between strangers discussing one of 70 pre-assigned topics. Each conversation is approximately 5 min long. The most frequent ten topics were selected as target topics with the remaining 60 topics denoted as the non-target topic set. A set of 96 conversation sides (4 h) was randomly selected from the non-target topic set for SOU training. The selected target topics, the number of training and test conversation sides (for TID) are listed in Table 1. TID decisions were made per conversation side. Two non-overlapping sets of target-specific impostor data were defined for each target topic. Each contains 600 tests randomly selected from the combination of the non-target topic set which is not used

[5] We have started to look at discriminative training for the SOU model which should provide greater distinction between the SOU units. Our early work has thus far only provided small gains but we believe it is still worth considering.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

Table 1
The 10 Switchboard target topics and their num. of training and test conversation sides.

Topic	# Train. conv.	# Test conv.
Car buying	85	20
Capital punishment	92	22
Recycling	102	25
Job benefits	90	21
News media	85	20
Public education	81	19
Drug testing	89	22
Exercise and fitness	81	19
Family finance	89	21
Family life	84	20

in HMM training, as well as data from the other target topics. One set was used as the negative examples in SVM classifier training and the other set was used as impostor data in classification evaluation.

The experiment started with segmentation, which was based on spectral power from 14 frequency bands with a maximum segment duration of 50 frames. The details are described in Gish et al. (2009). Vocal tract length normalized (VTLN) cepstral features and their first derivatives were generated for each segment that was then modeled with a quadratic trajectory model. The clusters were created using the k-means algorithm with initial centroids obtained using binary centroid splitting. These k-means clusters were further refined with 3 iterations of EM training. This resulted in 64 mixture components.

The SGMMs were used to tokenize the acoustic training data as the initial label for HMM training. This was followed by multiple iterations of unsupervised HMM training. Byblos used 5-state left-to-right HMMs trained with the maximum likelihood criterion. Sixty-dimensional acoustic features were generated by transforming the concatenation of 9 frames of the 13-dimensional cepstral features and its energy via an HLDA transform. We used a 512 component state-tied-mixture model (STM) for the monophone model and 64 Gaussian components for the cross-word state-cluster tied mixture (SCTM). Both speaker independent acoustic models and speaker adaptive training (SAT) models were created. During each iteration of unsupervised training, the acoustic training data were decoded using the newly learned parameters to generate new unit sequences. This decoding involved one pass of unadapted decoding followed by the HLDA adaptation, unsupervised MLLR adaptation and another pass of adapted decoding, all using only the STMs.

After the HMMs were trained, recognition was performed on all the classifier training and test data using both the STMs and the context dependent SCTMs as described in Section 3. We trained the SVM using the regression mode of the publicly available LIBSVM package (Chang and Lin, 2001).

To be able to compare performance with other published results on topic classification, we also performed larger scale experiments on the Fisher corpus similar to those set up in Hazen et al. (2007), which includes 40 topics. The topics and the amount of training and evaluation on-topic data are shown in Table 2. The data set includes a total of 244 h for topic training and 114 h for evaluation. The Fisher experiments were un-optimized in that the system is used "out-of-the-box", using exactly the same settings used in the Switchboard experiments, without any adjustment for the large data set and its variations in amount of topic training data.

4.2. Results

A number of topic classification experiments were performed with different variations of the unsupervised training as well as other benchmarks on the Switchboard experiments. The results are tabulated in Table 3.

We started with using only the SGMM (Unsup. SGMM only), which is denoted as (1) in Table 3, followed by the addition of one iteration of Byblos training (Unsup. Byblos 1 iter) denoted as (2) and five iterations of SOU training (Unsup. Byblos 5 iter (3)). These are similar to the SOU results reported in Gish et al. (2009). The most significant improvement comes from using the HMM SOUs probably because of the increased number of model parameters as well as the improved modeling techniques in HMM-based system such as speaker adaptation.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

Table 2
The 40 Fisher target topics and their num. of training and test conversation sides.

Topic	Train. hour	Test hour
Professional Sports on TV	10.0	3.47
Pets	11.6	3.4
Life Partners	11.7	4.4
Minimum Wage	11.3	3.9
Comedy	10.4	6.7
Hypothetical Situations: Perjury	3.3	2.9
Hypothetical Situations: One million dollars to leave the US	1.1	1.8
Hypothetical Situations: Opening your own business	5.4	6.9
Hypothetical Situations: Time travel	6.8	5.1
Hypothetical Situations: An anonymous benefactor	5.4	2.7
US Public School	4.8	2.7
Affirmative Action	3.0	1.1
Movies	3.5	1.8
Computer Games	2.6	0.4
Current Events	3.2	1.4
Hobbies	2.2	0.97
Smoking	2.5	1.6
Terrorism	2.8	2.8
Televised Criminal Trials	1.7	0.79
Drug Testing	1.0	0.50
Family Values	0.99	0.28
Censorship	6.9	1.2
Health and Fitness	8.6	1.9
September 11	14.6	4.0
Strikes by Professional Athletes	8.4	2.0
Airport Security	8.4	4.7
Issues in the Middle East	8.3	2.1
Foreign Relations	8.2	2.3
Education	8.7	3.5
Family	5.7	4.6
Corporate Conduct in the US	4.6	2.6
Outdoor Activities	1.7	1.4
Friends	2.9	1.4
Food	4.1	1.9
Illness	6.0	3.5
Personal Health	4.8	2.5
Reality TV	6.6	3.4
Arms Inspection in Iraq	4.0	2.3
Holidays	4.4	2.8
Bioterrorism	3.5	2.3

We test the effects of more complex modeling including context modeling as well as lattice-based features as denoted as systems (4)-(6). By using a more complex state-cluster tied-mixture acoustic model (Matsoukas et al., 2006) and trigram language model, EER is reduced to 10.7%. Adding context models in lattice rescoring further reduces the EER to 10.0%. The best performance of 6.6% EER comes from extracting *n*-gram features from the re-scored lattices.

The gain from using more complex models is consistent with our insight that SOUs can leverage progress from regular speech recognition. The particular gain in using lattice-based *n*-gram features is significant. It may be an indication that the recognized SOU sequences are not very sharp such that alternative hypotheses in the lattice contain many SOU sequences useful for TID.

To gain more insights into the relationship between SOUs and phonemes, we used dynamic programming to align the SOU sequences generated with our best SOU system and the reference phoneme sequences created from orthographic transcriptions. We found that some SOUs are good predictors of single phonemes, such as SOU "P1" to phoneme "b", SOU "P2" to phoneme "n", while other SOUs map to multiple phonemes, such SOU "P7" maps to phonemes "n", "m", and "ng" approximately uniformly.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

Table 3
Topic verification EER on 10 Switchboard topics using different phoneme/unit tokenizers.

System	Tokenizer	Class. EER
(1)	4-h unsup. SGMM only	30.2
(2)	4-h unsup. Byblos (1 iter)	16.1
(3)	4-h unsup. Byblos (5 iter)	12.5
(4)	(3) + 3-gram + SCTM	10.7
(5)	(4) + context model in lattice rescoring	10.0
(6)	(5) + lattice-based *n*-gram features	6.6
(7)	Oracle English phonemes	1.15
(8)	10-h sup. BUT Hungarian	16.7
(9)	1-h sup. Byblos English I	14.7
(10)	1-h sup. Byblos English II	6.2
(11)	2-h sup. Byblos English II	4.7
(12)	4-h sup. Byblos English II	4.4

To benchmark our performance, we performed experiments using tokenizers trained with supervision, which include an oracle experiment using the phoneme sequence derived from true test (denoted as (7)), tokenizing with a BUT Hungarian phoneme recognizer (denoted as (8)), and tokenizing with supervised training Byblos English phoneme recognizer (9–12).

The phoneme sequences in the oracle experiment were derived from a dictionary mapping using the orthographic transcription of training and test. This shows the lower bound on EER with using a perfect English phoneme recognizer. The near perfect results show the gain one can get in improving the tokenization. The BUT Hungarian model has a matching channel (i.e., telephone channel training), trained with a good amount of data (10 h) but was trained from a different language. We used the publicly available recognition software and model from (Burget and Matějka, Černocký, 2006). The 1-h English model matches the test but was trained with a small amount of data. Our first 1-h English model denoted in (9) was trained similar to the basic SOU model shown in (3) which did not use context models nor lattice-based features. Even with an order of magnitude more data, the EER from Hungarian phoneme recognizer is 2.6% absolute higher than the 1-h the basic Byblos training, probably due to the language mis-match. This is consistent with the results reported (Hazen et al., 2007) in which the Hungarian recognizer almost doubles the classification EER compared to a similarly trained English phoneme recognizer.

Given that the 1-h English model was trained with Byblos, we can also apply all the improvements to the SOU system to the English training, including the more complex models and lattice output for topic ID. The results are denoted as (8) which shows substantial improvement over the basic model by reducing the EER from 14.5% to 6.2%. This is roughly in-line with our improvement on the SOUs which improved from 12.5% EER in (3) to 6.6% in (6). Increasing the amount of training for the supervised case from 2 to 4 h sees another improvement of EER to 4.4%.

How significant is the benefit of building SOUs with an existing HMM-based recognition software? We have shown in (3)–(6) of Table 3 the benefit of using recognition features, such as the 3-gram or more complex SCTM model and lattice outputs. Another benefit is the availability of adaptation, both in speaker adaptive training (SAT) and in adaptation during decoding. In Table 4, we tabulate the results of not using adaptation in either the test, or in both training and test. Without adaptation, which is readily implemented in the recognizer, the performance of system (6) drops to 13.9% EER. In our more recent experiments, SOU training can also benefit from more training particularly when more complex models including pseudo-words are used.

Table 4
Effect of adaptation on topic verification EER on 10 Switchboard topics.

	Tokenizer	Class. EER
(1)	No adaptation in iterative training	10.3
(2)	No adaptation in both iterative training and test	13.9

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

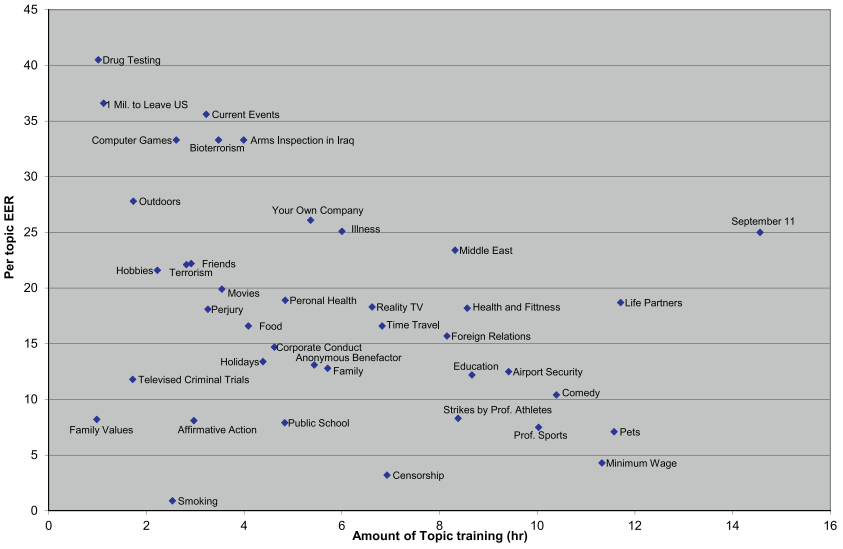

Fig. 3. Amount of in-topic classifier training data versus verification equal error rate on the Fisher corpus.

On the Fisher corpus, which has significantly more topics and variations on in-topic training, the topic verification EER is 18% and a closed set classification error rate (CER) is 45.9%. These results were obtained using the our best SGMM initialized SOU system as in system (6) in Table 3 but trained on Fisher. The 18% EER is significantly worse than the Switchboard performance but we suspect this is because of a harder topic ID setup. A comparison with results reported in Hazen et al. (2007)[6] shows that the SOU performance is significantly better than just using Hungarian phones at CER of 64.75.[7] This shows the power of unsupervised training that matches the test domain (English). On the other hand, unsupervised learning, with the same amount of data, is not as good as using a matched, English phoneme recognizer trained with hundreds of hours of data, which achieved a CER of 35.5%.

Fig. 3 shows the EER per topic against the amount of data available for each topic in the Fisher experiment. We can see the general trend of better verification performance for topics with more in-topic classifier training data. A couple of outliers can also be seen, with one showing the topic "September 11" which probably overlaps heavily with other topics such as "Terrorism". On the other hand, a distinct topic like "Smoking" has excellent performance (0.9% EER) even with limited amount of training (2.5 h).

One can also use the test data as unsupervised training data to further improve the model at the cost of making testing on new data cumbersome. This may be more useful if there is systematic mis-match between the unsupervised training data and the test, or the amount of unsupervised training data is small compared with the test data. Our experience has been the benefit of using the test data in unsupervised training is small.

5. Keyword discovery

Keyword discovery is the process of finding words that are useful in characterizing the content of an audio corpus. Thus they are content words that occur relatively frequently but at the same time, are non-uniformly distributed over the audio documents. The merit of any word being a useful keyword is often captured by such measures as TF-IDF (term frequency inverse document frequency). Other measures of keyword relevance have been based on mutual information (Levit et al., 2001).

[6] We used the exact same topic training and evaluation for closed set topic classification as in Hazen et al. (2007) but the set of impostors in open set verification task is different.

[7] This is not an exact comparison because of differences in classifiers used although the naive Bayes classifier used with the BUT phoneme sequence performed very similarly to SVM classifier as reported in Hazen et al. (2007). Better classification performance was reported by Hazen in Hazen and Margolis (2008) using improved classifier but should not affect our comparison of sequence quality.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

Table 5
Words overlaps with the long-span SOU features in different topics in Switchboard.

Topic	SOU *n*-Grams	Mapped words	Precision (%)	SOU instances
Car buying	P24 P41 P25 P23 P39	CAR(S)	35	65
Capital punishment	P58 P30 P23 P50 P30 P23	CAPITAL	83	71
Recycling	P42 P26 P41 P25 P23	RECYCL(ING/D)	84	467
Job benefits	P24 P59 P26 P33 P25	BENEFITS	36	153
News media	P18 P27 P8 P25 P38	NEWSPAPER(S)	44	158
Public education	P19 P14 P25 P22 P36	TEACHER(S)	62	319
Drug testing	P22 P37 P25 P20 P59	DRUG(S)	71	65
Exercise and fitness	P0 P62 P24 P11 P25	AEROBIC(S)	46	56
Family finance	P24 P40 P25 P21 P33	BUDGET(S)	53	88
Family life	P26 P44 P0 P42 P25	FAMILY	66	108

In our particular situation, we are working within the framework of a topic identification problem and are able to exploit the topic structure of the audio as well as SVM-based feature weighting and building algorithms in order to generate keyword candidates.

The SVM feature building described in Campbell and Richardson (2007) generates a set of long-span *n*-gram features by selectively expanding important lower order *n*-grams. The importance of any *n*-gram feature is measured by the absolute value of the corresponding SVM classifier weight. For each topic, this results in a ranked list of long-span *n*-grams, such as 5-gram and 6-gram. One question is whether these top SOU *n*-grams correspond to any topic specific keywords. By performing forced alignment on both the word transcription and SOUs, we can time align the SOU tokens with the word transcript. With this alignment, we can extract all English words that overlap with any SOU *n*-gram of interest. In Section 5.1, we discuss how the top SOU *n*-grams relate to topic keywords.

5.1. Discovering topic keywords

We will need the orthographic transcription of the test to evaluate the performance of keyword discovery. Given the transcription and an ASR model, we can align SOU sequences with the word transcription. We looked at the top 5-gram and 6-gram features in our keyword building stage for each topic as listed in the first column of Table 5. Note that one SOU instance can overlap in time with more than one word. The first column shows the topic, the second column shows the top 1 SVM *n*-gram feature while the third column shows the most frequently associated word that overlaps with these *n*-gram instances. One can view the SOU sequences as a clustering criterion for grouping audio segments, and then, the most frequently associated word can be viewed as the label of the cluster of audio segments. The fourth column shows the precision of the SOU *n*-grams, which is the percentage of these instances that contains the most frequent word. The last column shows the number of times the SOU *n*-grams occur in our corpus.

As can be seem, the keywords are always associated strongly with the topic, and the respective precisions are quite high ranging from 35% to 84%. Some of the keywords are in fact key phrases. For example, while “CAPITAL” occurs 83%, “PUNISHMENT” also occurs more than 80% within the same set of SOU sequences. “CAR” is a difficult keyword word because it is short and is a common part of other words, such as “CARD”, “CARPET” etc. While we only show the top *n*-gram per topic, other *n*-grams are also closely associated with topic keywords, with many of them being variants of the top *n*-gram that differ from the top one by one or two SOUs.

5.2. Keyword search

Keyword search, as opposed to keyword discovery, requires a pattern to search for. In an earlier work by Garcia and Gish (2006), several keyword discovery experiments were performed using a self-organizing recognizer i.e., an SGMM recognizer. In their application they had 15 min of transcribed audio which was used for learning the multi-gram mapping between letters of a language and SOUs. This mapping was used to convert words to strings of SOUs for searching purposes.

If we do not have transcribed audio then the only option is to perform query by example. Applications of using SOUs, either as components for posteriorgram (Harwath et al., 2013), or for keyword discovery (Siu et al., 2011) have

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

shown that similar SOU patterns are often associated to the same words, which would suggest that SOUs could be effective for query-by-example.

6. Conclusions and future work

In this paper, we presented our novel approach to unsupervised HMM learning using SOUs, which can be a valuable option for speech applications in domains of limited transcribed data. We presented results in topic identification that showed SOUs can be competitive with systems trained with a limited amount of in-domain data and outperform those trained from mis-matched data. We described some of the improvements in SOUs, including context dependent acoustic models and lattice-based *n*-gram features that resulted in significant TID EER reduction. We further explored the relationship between some selected SOU sequences and their corresponding English words. We showed that these SOU *n*-grams are consistently mapped to the same English keywords, and SOU *n*-grams selected as TID features were representing topic keywords. Thus, the SOU TID system can also be used for unsupervised keyword discovery.

One direction we are currently exploring is the building of bigger units during the SOU acoustic training. This is analogous to building a pseudo-word recognizer instead of phoneme recognizer which could further improve the SOU consistency and thus, improve downstream topic ID performance.

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments and suggestions.

References

Belfield, W., Gish, H., 2003. A topic classification system based on parametric trajectory mixture models. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP), pp. 1269–1272.

Burget, L., Matějka, P., Černocký, J., 2006. Discriminative training techniques for acoustic language identification. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP).

Campbell, W., Richardson, F., 2007. Discriminative feature selection using support vector machines. In: NIPS.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines, Software available at: http://www.csie.ntu.edu.tw/cjlin/libsvm

Cohen, J., 1981. Segmenting speech using dynamic programming. Journal of Acoustic Society of America, 1430–1437.

Garcia, A., Gish, H., 2006. Keyword spotting of arbitrary words using minimal speech resources. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP).

Gish, H., Ng, K., 1993. A segmental speech model with applications to word spotting. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP), pp. 447–450.

Gish, H., Ng, K., 1996. Parametric trajectory models for speech recognition. In: Proc. of the Inter. Conf. on Spoken Language Processing (ICSLP), pp. 466–469.

Gish, H., Siu, M., Belfield, W., 2009. Unsupervised training of an HMM-based speech recognition system for topic classification. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Harwath, D., Hazen, T., Glass, J.R., 2013. Zero resource spoken audio corpus analysis. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP).

Hazen, T., Margolis, A., 2008. Discriminative feature weighting using MCE training for topic identification of spoken audio recording. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP), pp. 4965–4968.

Hazen, T., Richardson, F., Margolis, A., 2007. Topic identification from audio recordings using word and phone recognition lattices. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop.

Jansen, A., Church, K., Hermansky, H., 2010. Towards spoken term discovery at scale with zero resources. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Lamel, L., Gauvain, J., Adda, G., 2002. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language 16, 115–129.

Levit, M., Gorin, A., Wright, J., 2001. Multipass algorithm for acquisition of salient acoustic morphemes. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1645–1648.

Li, H., Ma, B., Lee, C., 2007. A vector space modeling approach to spoken language identification. IEEE Transactions on Audio, Speech and Language Processing 15 (1), 271–284.

Ma, J., Schwartz, R., 2008. Unsupervised versus supervised training of acoustic models. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2374–2377.

Matsoukas, S., et al., 2006. Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. IEEE Transactions on Audio, Speech and Language Processing 14, 1541–1555.

Nowell, P., Moore, R., 1995. The application of dynamic programming techniques to non-word based topic spotting. Proc. of the European Conference on Speech Comm. and Tech. (EUROSPEECH), 1355–1358.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002

Park, A., Glass, J., 2005. Towards unsupervised pattern discovery in speech. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop, pp. 53–58.

Schwartz, P., Matejka, P., Cernocky, J., 2004. Towards lower error rates in phoneme recognition. In: Proc. of TSD.

Siu, M., Gish, H., Chan, A., Belfield, W., 2010. Improved topic classification and keyword discovery using an HMM-based speech recognizer trained without supervision. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2838–2841.

Siu, M., Gish, H., Lowe, S., Chan, A., 2011. Unsupervised audio patterns discovery using hmm-based self-organized units. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Wintrode, J., Kulp, S., 2009. Confidence-based techniques for rapid and robust topic identification of conversational telephone speech. In: Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Zavaliagkos, G., Siu, M., Colthurst, T., Billa, J., 1998. Using untranscribed training data to improve performance. In: Proc. of the Inter. Conf. on Spoken Language Processing (ICSLP).

Zhang, Y., Glass, J., 2010. Towards multi-speaker unsupervised speech pattern discovery. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP).

Zissman, M.A., 1993. Automatic language identification using Gaussian mixture and hidden Markov models. In: Proc. of the IEEE Inter. Conf. on Acoust., Speech and Signal Proc. (ICASSP), vol. 2, pp. 399–402.

Please cite this article in press as: Siu, M.-h., et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.05.002