TOPIC IDENTIFICATION FROM AUDIO RECORDINGS USING WORD AND PHONE RECOGNITION LATTICES

Timothy J. Hazen, Fred Richardson and Anna Margolis

MIT Lincoln Laboratory
Lexington, Massachusetts, USA

ABSTRACT

In this paper, we investigate the problem of topic identification from audio documents using features extracted from speech recognition lattices. We are particularly interested in the difficult case where the training material is minimally annotated with only topic labels. Under this scenario, the lexical knowledge that is useful for topic identification may not be available, and automatic methods for extracting linguistic knowledge useful for distinguishing between topics must be relied upon. Towards this goal we investigate the problem of topic identification on conversational telephone speech from the Fisher corpus under a variety of increasingly difficult constraints. We contrast the performance of systems that have knowledge of the lexical units present in the audio data, against systems that rely entirely on phonetic processing.

Index Terms— Audio document processing, topic identification, topic spotting.

1. INTRODUCTION

As new technologies increase our ability to create, disseminate, and locate media, the need for automatic processing of these media also increases. Spoken audio data in particular is a media which could benefit greatly from automatic processing. Because audio data is notoriously difficult to "browse", automated methods for extracting and distilling useful information from a large collection of audio documents would enable users to more efficiently locate the specific content of their interest. One specific task of interest is automatic topic identification (or topic ID), for which the goal of a system is to identify the topic(s) of each audio file in its collection. A variant of the topic identification problem is the topic detection (or topic spotting) problem, for which a system must detect which audio files in its collection pertain to a specific topic.

Topic identification has been widely studied in both the text processing and speech processing communities. The most common approach to topic identification for audio documents is to apply word-based automatic speech recognition to the audio, and then process the resulting recognized word strings using traditional text-based topic identification techniques [1]. This approach has proven to work effectively for tasks in which reasonably accurate speech recognition performance is achievable (e.g. news broadcasts) [2]. Of course, speech recognition errors can degrade topic identification performance, and this degradation becomes more severe as the accuracy of the speech recognizer decreases.

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.

Despite previous successes, existing speech recognition systems may not perform well enough to support accurate topic identification for some tasks. Two common reasons for the inadequacy of a speech recognition system are (1) a severe mismatch between the data used to train the recognizer and the unseen data on which it is applied, and (2) a dearth of training data that is well matched to the conditions in which the recognizer is used. One problem that could manifest itself, for example, is a mismatch between the vocabulary employed by the recognizer and the topic-specific vocabulary used in the data of interest. In the most extreme case, a recognition system may not even be available in the language of the data of interest.

When training a topic identification system, one would ideally possess a large corpus of transcribed data to help train both a speech recognition system and a topic identification module. Unfortunately, manual transcription of data is both costly and time-consuming. To alleviate this cost, one could resort to a more rapid manual annotation of available data in which audio content is only labeled by topic and full lexical transcription is not performed. In this case, the determination of relevant lexical items for topic identification can not be determined from manual transcriptions, but instead must be deduced somehow from the acoustics of the speech signal. Towards this end, several previous studies have investigated the use of phonetic speech recognizers (instead of word recognizers) in the development of topic identification systems [3, 4, 5, 6].

In this paper, we empirically contrast topic identification systems using word-based speech recognition vs. phone-based speech recognition. Furthermore, we investigate a variety of methods for improving the performance of both word- and phone-based topic identification. We begin by investigating a traditional Naïve Bayes formulation of the problem. Within this formulation we examine a variety of feature selection techniques required to optimize performance of the approach. We also investigate a support vector machine (SVM) approach which has previously been successfully applied to the problems of speaker and language identification [7].

2. EXPERIMENTAL TASK DESCRIPTION

2.1. Corpus

For the data set for our experiments we have used the English Phase 1 portion of the Fisher Corpus [8, 9]. This corpus consists of 5851 recorded telephone conversations. During data collection, two people were connected over the telephone network and given instructions to discuss a specific topic for 10 minutes. Data was collected from a set of 40 different topics. The topics were varied and included relatively distinct topics (e.g. "Movies", "Hobbies", "Education", etc.) as well as topics covering similar subject areas (e.g. "Issues in Middle East", "Arms Inspections in Iraq", "Foreign Relations"). Fixed prompts designed to elicit discussion on the topics

were played to the participants at the start of each call. For example, the prompt for the "Foreign Relations" topic was:

> *Do either of you consider any other countries to be a threat to U.S. safety? If so, which countries and why?*

From this corpus we conduct two basic types of experiments:

1. Closed set topic identification (i.e., identify the topic from the closed set of 40 topics).
2. Topic detection (i.e., specify if an audio document is or is not about a specified topic)

For our experiments the corpus was subdivided into four subsets:

1. Recognizer training set (3104 calls; 553 hours)
2. Topic ID training set (1375 calls 244 hours)
3. Topic ID development test set (686 calls; 112 hrs)
4. Topic ID evaluation test set (686 calls; 114 hrs)

There is no overlap in speakers across the sets.

For our experiments we can treat each conversation as an independent audio document. This yields a total of 686 separate audio documents for each of the development and evaluation test sets. Because each side of the conversation was recorded into independent channels within its audio file, we can further break each conversation into two separate call side documents yielding a total of $686 \times 2 = 1372$ audio documents in each of the development and evaluation test sets. In our experiments we provide results on a *whole call* and/or a *call side* basis. Individual calls are further subdivided into individual audio segments, which are typically a few seconds in length, for processing by a speech recognition system.

2.2. Speech Recognition Systems

2.2.1. Overview

In our topic ID experiments, the first stage of processing is to apply automatic speech recognition (ASR) to each segment of audio in each audio document. The ASR system is used to generate a network, or *lattice*, of speech recognition hypotheses for each audio segment. In this work we explore the use of both word-based and phone-based speech recognition. Within each lattice we can compute the posterior probability of any hypothesized word (or sequence of phones for a phone-based system), and an *expected count* for each word can be computed by summing the posterior scores over all instances of that word in the lattice.

2.2.2. Word-Based Speech Recognition

For word-based ASR we have used the MIT SUMMIT speech recognition system [10]. The system's acoustic models were trained using a standard maximum-likelihood approach on the full 553 hour recognition training set specified above without any form of speaker normalization or adaptation. For language modeling, the system uses a basic trigram language model with a 31.5K word vocabulary trained using the transcripts of the recognizer training set. Because this recognizer applies very basic modeling techniques with no adaptation, the system performs recognition faster than real time (on a current workstation) but word error rates can be high (typically over 40%).

2.2.3. Phone-Based Speech Recognition

For phonetic recognition we use a phonetic ASR system developed at the Brno University of Technology (BUT) [11]. Two versions of the system were trained, one which uses an English phone set and one which uses a Hungarian phone set. The English recognizer was trained using 10 hours from the Switchboard Cellular Phase 1 conversational telephone speech corpus [9]. This training was seeded with phonetic time alignments generated by the BBN ASR system [12]. The Hungarian recognizer was trained using the Hungarian portion of the SPEECH-DAT corpus [13]. This corpus contains read speech collected over the Hungarian telephone network.

3. PROBABILISTIC TOPIC IDENTIFICATION

3.1. The Naïve Bayes Formulation

In a probabilistic approach to topic identification, the goal is to determine the likelihood of an audio document being of topic t given the string of spoken words W. Here, each known topic t is an element of a set of topics T. This can be expressed mathematically, and expanded via Bayes rule, as follows:

$$P(t|W) = P(t)\frac{P(W|t)}{P(W)} = P(t)\frac{P(w_1,\ldots,w_N|t)}{P(w_1,\ldots,w_N)} \quad (1)$$

Here, the word string W is expanded into its underlying sequence of N words, $w_1,\ldots,w_N$. In the Naïve Bayes approach to the problem, statistical independence is assumed between each of the individual words in W. Under this assumption, the posterior of t given W is approximated as:

$$P(t|W) \approx P(t)\prod_{i=1}^{N}\frac{P(w_i|t)}{P(w_i)} \quad (2)$$

The expression above assumes a sequence of N individual words. This expression can alternatively be represented using a counting interpretation instead as follows:

$$P(t|W) \approx P(t)\prod_{\forall w\in V}\left(\frac{P(w|t)}{P(w)}\right)^{C(w|W)} \quad (3)$$

In this interpretation, the occurrence count $C(w|W)$ within W of each word w in the system's vocabulary V is used to exponentially scale the score contribution of that word. Under this interpretation non-integer values of the counts $C(w|W)$ are allowed, thus providing the system the ability to incorporate word posterior estimates from a lattice generated by a recognition system. This Naïve Bayes approach is utilized as our baseline system.

3.2. Parameter Estimation

The likelihood functions in our probabilistic systems are all estimated from training materials using maximum *a posteriori* probability (MAP) estimation. For example, the prior probability of the topic, $P(t)$, obtained using MAP estimation is expressed as follows:

$$P_{map}(t) = \frac{N_D}{N_D+\alpha_T N_T}P_{ml}(t) + \frac{\alpha_T N_T}{N_D+\alpha_T N_T}\frac{1}{N_T} \quad (4)$$

In this expression, N_D is the total number of documents in the training set and N_T is the number of distinct topics used for classification. MAP estimation results in an interpolation of the maximum likelihood (ML) estimate, $P_{ml}(t)$, with a prior distribution for $P(t)$ which is assumed to be uniform (i.e. $1/N_T$). The rate of the interpolation is controlled by the smoothing parameter α_T (which is typically determined empirically). As the number of training documents, N_D, increases, the MAP estimate moves away from the prior uniform distribution towards the ML estimate. The ML estimate is simply expressed as:

$$P_{ml}(t) = \frac{N_{D|t}}{N_D} \quad (5)$$

Here, $N_{D|t}$ is the number of documents in the training set belonging to topic t. When combining Equations 4 and 5, the MAP estimate used to model $P(t)$ reduces to:

$$P_{map}(t) = \frac{N_{D|t} + \alpha_T}{N_D + \alpha_T N_T} \quad (6)$$

A MAP estimate for the prior likelihood of word w, $P(w)$, can be constructed in the same fashion, and is expressed as:

$$P_{map}(w) = \frac{N_w + \alpha_W}{N_W + \alpha_W N_V} \quad (7)$$

In this expression, N_V is the number of unique words in the vocabulary used for topic identification, N_w is the number of occurrences of the specific word w in the training corpus, N_W is the total count of all words from the N_V word vocabulary in the training corpus, and α_W is a MAP estimation smoothing parameter.

The MAP estimate for $P(w|t)$ is also built in a similar fashion. However, instead of using a uniform distribution, the distribution $P_{map}(w)$ described in Equation 7 can be used as the prior distribution against which the ML estimate $P_{ml}(w|t)$ is interpolated. In this case the MAP estimate is expressed as:

$$P_{map}(w|t) = \frac{N_{w|t} + \alpha_{W|T} N_V P_{map}(w)}{N_{W|t} + \alpha_{W|T} N_V} \quad (8)$$

In this expression, $N_{w|t}$ is the number of times word w occurs in training documents of topic t, $N_{W|t}$ is the total number of words the training documents of topic t, and $\alpha_{W|T}$ is the MAP estimation smoothing parameter.

Equations 6, 7, and 8 each contain an α smoothing parameter. Appropriate settings for these parameters must be learned empirically. In our case, preliminary experiments revealed that performance is not highly sensitive to the value of these terms and near-optimal performance on development test data was achieved by simply setting all of the α terms to a value of 1. Thus, we set each α term to a value of 1 for all experiments discussed later in this paper.

3.3. Feature Selection

In word-based topic identification, it is typically the case that a small number of topic specific content words contribute heavily to the determination of the topic, while many non-content words (i.e., articles, prepositions, auxiliary verbs, etc.) contribute nothing to the decision. For this reason, probabilistic approaches to topic identification typically employ a feature selection process in which only a subset of words from the full vocabulary of the system are used when performing the probabilistic scoring. The use of a "stop list" of common topic-independent words that should be ignored is a common practice in topic identification systems. Techniques for selecting useful topic specific words using the χ^2 statistic or the information gain measure have also been employed in previous work [14].

3.3.1. Topic Posterior Estimation

Two of the feature selection approaches we examine require an estimate of the posterior probability $P(t|w)$. While this measure can be estimated indirectly using Bayes Rules (as is done in Sections 3.1 and 3.2), our feature selection techniques perform better when $P(t|w)$ is estimated directly using the following MAP estimate:

$$P_{map}(t|w) = \frac{N_{w|t} + \alpha_{T|W} N_T P_{map}(t)}{N_w + \alpha_{T|W} N_T} \quad (9)$$

Here, $\alpha_{T|W}$ is a MAP smoothing parameter set to one.

3.3.2. The Information Gain Measure

The information gain measure for a word w can be defined mathematically as:

$$IG(w) = H(t) - H(t|w) \quad (10)$$

where $H(t)$ is the entropy of the prior distribution, $P(t)$, and $H(t|w)$ is the entropy of the conditional distribution $P(t|w)$ when the specific word w is observed. This expression expands to:

$$IG(w) = \sum_{\forall t \in T} -P(t)\log P(t) + P(t|w)\log P(t|w) \quad (11)$$

One potential problem with the information gain measure is that it favors words that are predominantly present in only one topic. In this case the information gain measure may only select words from the topics which are distinctly different from the other topics (and hence already easy to distinguish). In cases where several topics may be similar, the words that could predict these topics may appear frequently across these similar topics, which in turn would increase the conditional entropy of these words. Under these circumstances, the most predictive words for these topics may not be selected as features when using the information gain criteria.

3.3.3. The χ^2 Statistic

The χ^2 statistic is used for testing the independence of two variables t and w from their observed co-occurrence counts. It is defined as follows: let A be the number of times word w occurs in topic t, B the number of times w occurs outside of topic t, C the total number of words in topic t that aren't w, and D the total number of words outside of topic t that aren't w. Let N_W be the total number of word occurrences in the training set. Then:

$$\chi^2(t,w) = \frac{N_W(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (12)$$

This statistic uses raw counts and does not depend on smoothed probability estimates. In [14], this measure was converted into a "global" selection measure for a word w by averaging its χ^2 statistic over all classes. We use it here as a per-class selection measure: for each topic t, we select the N words with the highest χ^2 values. These words are the least likely to be independent of topic t, i.e., their presence (or absence) in a document is likely to give information about the document belonging to topic t.

3.3.4. The Maximum Posterior Probability Measure

Another approach we have investigated is to select the top N words per topic which maximize the posterior probability of the topic, i.e. the words which maximize the value of $P(t|w)$ as estimated in the Equation 9. Experiments on our development set show this feature selection approach generally outperforms the other approaches mentioned above. For example, Figure 1 shows the performance of the three measures discussed above on individual call sides within the topic ID development test set when applied to the manual text transcriptions of the data. Optimal performance was achieved when selecting the top 25 words per topic using the maximum posterior probability measure. To illustrate the feature selection, Table 1 shows the top 10 words which maximize $P(t|w)$ for five specific topics in the topic ID training set. Feature selection in the remaining Naïve Bayes experiments in this paper all use this maximum posterior probability approach.

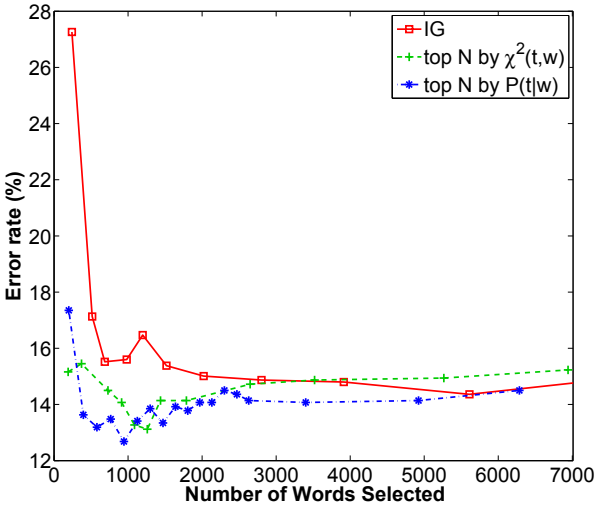

Fig. 1. Topic identification on call sides within the topic ID development test set when using the manual text transcriptions. Performance of the information gain, χ^2, and maximum posterior probability measures is shown as the number of selected features is varied.

Topics				
Pets	Professional Sports on TV	Airport Security	Arms Inspections in Iraq	U.S. Foreign Relations
dog	football	airport	inspections	threat
dogs	hockey	security	disarming	Korea
cat	Olympics	plane	oil	countries
pet	sport	flights	weapons	nuclear
cats	basketball	flight	inspectors	China
pets	sports	shoes	arms	threats
animals	golf	flown	Saddam	nations
fish	team	safer	U.N.	Iraq
goldfish	soccer	airports	voting	Iran
animal	baseball	fly	destruction	allies

Table 1. Examples of top-10 words w which maximize $P(t|w)$ for five specific topics.

3.4. Processing Phonetic Strings

In our formulation of the topic identification problem, we have assumed that documents can be expressed as a collection of words and their associated counts. However, for some applications knowledge of the important lexical keywords for a task may be incomplete or unavailable. Under these situations, it is possible to instead perform topic identification based on phonetic strings (or phonetic lattices) while retaining the same probabilistic framework. In this case, the word string W is not comprised of words, but rather a set of units derived from the proposed sequence of phonetic units. A common approach to deriving word-like units from phonetic strings is to extract n-grams from the phonetic string. For example, the word string "I like dogs" would be represented by the phonetic string:

ay l ay k d ao g z

This phonetic string can be converted into the following sequence of phonetic trigram units:

ay:l:ay l:ay:k ay:k:d k:d:ao d:ao:g ao:g:z

Topics				
Pets	Professional Sports on TV	Airport Security	Arms Inspections in Iraq	U.S. Foreign Relations
p:ae:t	w:oh:ch	r:p:ao	w:eh:p	ch:ay:n
ax:d:ao	s:b:ao	ch:eh:k	hh:ao:s	w:eh:p
d:oh:g	g:ey:m	ei:r:p	w:iy:sh	th:r:eh
d:d:ao	s:p:ao	r:p:w	axr:oh:k	r:eh:t
d:ao:ix	ey:s:b	axr:p:ao	axr:dh:ei	th:r:ae
axr:d:ao	oh:ch:s	iy:r:p	p:aw:r	ay:r:ae
t:d:ao	w:ay:ch	iy:r:dx	w:ae:p	r:ae:k
p:eh:ae	w:aa:ch	ch:ae:k	axr:ae:k	ah:n:ch
d:ow:ao	hh:oh:k	s:ey:f	v:axr:dh	n:ch:r
d:oh:ix	oh:k:iy	r:p:l	r:ae:k	uw:ae:s

Table 2. Examples of top-10 phonetic trigrams w recognized by the BUT English phonetic recognizer which maximize $P(t|w)$ for five specific topics.

The same feature selection and probabilistic scoring mechanisms used for processing words can similarly be used with phonetic n-grams. For example, the feature selection mechanism could learn that the phonetic trigram unit d:ao:g (corresponding to the word *dog*) is a useful feature for predicting the topic "Pets".

To provide an example of phonetic n-gram feature selection, Table 2 displays the top 10 trigrams extracted from outputs of our English phonetic recognizer for five topics. The feature selection mechanism identifies trigrams which can typically be mapped to distinctive words from that topic. For example, in the "U.S. Foreign Relations" topic, the trigram *ch:ay:n* comes from the word *China*, and the trigrams *th:r:eh* and *r:eh:t* come from the word *threat*. Trigrams containing common phonetic errors within the distinctive words are also evident, e.g., the trigram *th:r:ae* is a substitution error for the correct trigram *th:r:eh* within the word *threat*.

4. SVM TOPIC IDENTIFICATION

As an alternative to the Naïve Bayes approach, support vector machines (SVMs) have also been successfully applied to the topic identification problem for text processing applications [15]. In the speech processing arena, SVMs have also proven effective for the speaker identification and language identification tasks [7].

In this work, our SVM architecture is based upon a previous phonetic n-gram approach to speaker identification [16]. As in our Naïve Bayes approach, a document is represented as a sequence of words (or phonetic n-gram tokens), $W = w_1, \cdots, w_N$. For each audio document, the relative frequency of each unique word w_j in W is expressed as follows:

$$P(w_j|W) = \frac{C(w_j|W)}{|W|} \tag{13}$$

Here $C(w_j|W)$ is the occurrence count of unique word w_j in W and $|W|$ is the total number of words in W. These relative frequencies are mapped to a sparse vector representation with one dimension per vocabulary item. The vector's entries have the following form:

$$P(w_j|W)\sqrt{\frac{1}{P(w_j)}} \tag{14}$$

Here the prior probability $P(w_j)$ is estimated (via maximum likelihood) from all documents across all topics in the topic identification training data.

Next, a kernel function for comparing two word sequences W and V, is defined as follows:

$$K(W,V) = \sum_{\forall j} \frac{P(w_j|W)P(w_j|V)}{P(w_j)} \quad (15)$$

Intuitively, the kernel in Equation (15) expresses a high degree of similarity between W and V (via a large inner product) if the frequencies of the same word are similar between the two sequences. If individual words are not present in one of the sequences, then this will reduce similarity because one of the probabilities in the numerator of Equation (15) will be zero. The denominator, $p(w_j)$, insures that words with large probabilities do not dominate the score. The kernel can alternatively be viewed as a linearization of the log-likelihood ratio (see [16] for more details).

Incorporating the kernel in Equation (15) into an SVM system is straightforward. Our system uses the SVMTorch package [17]. Training is performed with a one-versus-all strategy. For each target topic, we pool the vectors from all audio documents from all non-target topics then train an SVM to distinguish the target topic from the non-target topics. In aggregate, one topic-specific SVM classifier is created for each topic.

5. EXPERIMENTAL RESULTS

5.1. Topic Identification: Words vs. Phones

In our first set of experiments, we explore topic identification conducted under five different, and increasingly difficult, constraints:

1. Using the human generated text transcripts for each call.
2. Using phonetic strings generated by phonetic forced alignment of the transcripts (as generated by the MIT SUMMIT recognizer).
3. Using word lattices automatically generated by the MIT SUMMIT word recognition system.
4. Using phonetic lattices generated by the BUT English phonetic recognizer.
5. Using phonetic lattices generated by the BUT Hungarian phonetic recognizer.

For each constraint, a Naïve Bayes classifier was trained using the data from the topic ID training subset. System parameters, such as the number features used by the classifier under each condition, were optimized on the development test set. Results for closed set topic identification on the evaluation test set using the Naïve Bayes system are shown in Table 3.

In examining the results in Table 3, several observations can be made. First when comparing the result using the human generated word transcriptions vs. the result obtained from phone trigrams extracted from the forced alignments, we observe that there is little difference in performance between the two constraints. The word-based system performs slightly better on the individual call sides, but the phone-based system performs slightly better when analyzing the whole call. This suggests that lexical knowledge is not necessary to perform topic identification provided accurate phonetic transcriptions are available.

Next, when we examine the performance of the system using the human generated transcripts vs. the system using the lattices from the automatic word recognizer, we see a modest degradation in performance from the system using the automatic word recognizer. This indicates that the high error rate of the word recognition system does harm performance, but not dramatically so.

Experimental Conditions		Topic ID Error Rate(%)	
Features Extracted	Feature Selection	Call Sides	Whole Call
Words from Transcription	25 words per topic	12.4	8.2
English Phones from Forced Alignment	100 3-grams per topic	12.8	7.6
Words from ASR Lattices	100 words per topic	16.8	9.6
English Phones from ASR Lattices	100 3-grams per topic	35.3	22.9
Hungarian Phones from ASR Lattices	100 3-grams per topic	64.7	52.9

Table 3. Closed-set topic identification using a Naïve Bayes classifier under a variety of experimental conditions.

Finally, the last two rows of the table show the performance using phonetic lattices. Here, a more substantial degradation is observed. This indicates that the lack of lexical knowledge combined with imperfect phonetic recognition contributes to a significant drop in topic identification accuracy. It is interesting to note, however, that despite the significant relative drop in performance from the English phonetic recognizer to the Hungarian phonetic recognizer, the system using the Hungarian recognizer is still able to identify the topic of an English call nearly 50% of the time. By contrast, the accuracy of always selecting the most frequent topic in the training data is only 4%. This discriminative ability exists despite the severe mismatch in both the language and speaking style between the recognizer's training material and the test data.

5.2. Topic Identification: System Comparisons

To explore the capabilities of different classification systems, score normalization techniques, and fusion techniques, we have conducted additional experiments using the phonetic lattices generated by the BUT English phonetic recognizer. In these experiments we explore the use of both the Naïve Bayes System and the SVM system. Table 4 shows a wide range of results over different system configurations. These results are discussed below.

In our topic identification work we have discovered score normalization issues comparable to those experienced in the fields of speaker identification and language identification. In particular, the range of the scores across different audio documents can vary dramatically. To compensate for these issues we have explored two score normalization techniques: (1) Test Normalization (or T-norm), and (2) a "back-end" (BE) normalization technique that applies a linear discriminant analysis transform to the vector of scores followed by a Gaussian model classifier. A description of T-norm can be found in [18]. A description of our back-end classifier can be found in [19]. It is important to note that the back-end normalization technique requires training on scores generated from our development test set. For both of the normalization techniques used, the final normalized scores, $S(W,t_i)$, for a word sequence W for each topic t_i are further converted into a log-likelihood ratio form using this expression:

$$S_{llr}(W,t_i) = \log \frac{\exp(S(W,t_i))}{\frac{1}{N_T-1}\sum_{j\neq i}\exp(S(W,t_j))} \quad (16)$$

T-norm can also be applied first and then followed by normalization by the back-end classifier (as expressed as "T-Norm+BE" in Table 4). Using this normalization sequence reduces the closed set

Classifier	Normalization	Fusion	CER(%)	EER(%)
Bayes	None	N/A	35.3	11.07
Bayes	T-Norm	N/A	35.3	10.80
Bayes	BE	N/A	38.9	12.90
Bayes	T-Norm + BE	N/A	34.1	9.77
SVM	None	N/A	37.1	12.24
SVM	T-Norm	N/A	37.1	10.50
SVM	BE	N/A	34.9	10.57
SVM	T-Norm + BE	N/A	35.2	10.13
Bayes+SVM	None	Linear	34.0	10.28
Bayes+SVM	None	BE	32.4	9.08
Bayes+SVM	T-Norm	Linear	32.3	9.25
Bayes+SVM	T-Norm	BE	31.2	8.82

Table 4. Closed-set topic ID classification error rates (CER) and topic detection equal error rates (EER) using the BUT English recognizer phone lattices with various classifiers, score normalization techniques, and score fusion techniques.

topic classification error of the Naïve Bayes system from 35.3% to 34.1%, and the SVM system from 37.1% to 35.2%.

The results of the Naïve Bayes and SVM systems can also be fused. The simplest fusion is a linear combination of the score vectors of the two systems using an equal weighting for each system. Our back-end system can also be used as the fusion mechanism by providing it with the concatenated vector of topic scores from each of the two systems. Fusion of the T-normed scores from the two systems using the back-end classifier reduces the classification error rate to 31.2%. The represents an 8.5% relative error rate reduction from the best Naïve Bayes system result of 34.1%.

Table 4 also shows results for topic detection, i.e., specifying whether a test file is (or is not) from a specific topic. Results are reported in terms of the system equal error rate (i.e., the point in a detection error trade-off curve where a topic detector is equally likely to falsely reject a document belonging to the target topic as it is to falsely accept a document that does not belong to the target topic.) In the topic detection case, the best Naïve Bayes system achieves an equal error rate (EER) of 9.77%. The best fused system achieves an EER of 8.82%, which represents a 10% relative reduction in EER over the best Naïve Bayes system.

6. SUMMARY

In this paper we have examined the task of topic identification for audio documents. We have examined the problem under a variety of constraints under which systems may be deployed. Under the best case scenario, a well-trained word-based ASR system for the topic domains of interest is available. In this case, only minor degradations in topic identification from a comparable text-based system can be expected. For the scenario in which knowledge of the lexical units useful for predicting the topics of interest is unknown, we have attacked the problem by performing pattern selection of phonetic sequences obtained from automatic phonetic transcriptions of the training material. Experiments showed that phonetic sequences corresponding to topic specific words can be learned automatically and exploited to perform reasonably accurate topic detection.

7. REFERENCES

[1] J. McDonough, *et al*, "Approaches to topic identification on the Switchboard corpus," in *Proc. of ICASSP*, Adelaide, Australia, April 1994.

[2] R. Schwartz, T. Imai, L. Nguyen, and J. Makhoul, "A maximum likelihood model for topic classification of Broadcast News," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997.

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

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

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

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

[7] W. Campbell, *et al*, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210-229, April 2006.

[8] 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, Portugal, May 2004.

[9] Available from: http://www.ldc.upenn.edu/

[10] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 137-152, 2003.

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

[12] S. Matsoukas, *et al*, "The 2004 BBN 1xRT recognition systems for English broadcast news and conversational telephone speech," in *Proc. of Interspeech*, Lisbon, Portugal, Sep. 2005.

[13] Available from: http://www.fee.vutbr.cz/SPEECHDAT-E

[14] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. of Int. Conf. on Machine Learning (ICML)*, Nashville, TN, USA, July 1997.

[15] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. of European Conf. on Machine Learning (ECML)*, Chemnitz, Germany, April 1998.

[16] 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.

[17] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, nos. 1-3, pp. 42-54, 2000.

[19] W. Campbell, *et al*, "Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation," in Proc. of the IEEE Odyssey Speaker and Language Recognition Workshop, San Juan, Puerto Rico, June 2006.