Topic and Speaker Identification via Large Vocabulary Continuous Speech Recognition

Barbara Peskin, Larry Gillick, Yoshiko Ito, Stephen Lowe, Robert Roth, Francesco Scattone, James Baker, Janet Baker, John Bridle, Melvyn Hunt, Jeremy Orloff

Dragon Systems, Inc.
320 Nevada Street
Newton, Massachusetts 02160

ABSTRACT

In this paper we exhibit a novel approach to the problems of topic and speaker identification that makes use of a large vocabulary continuous speech recognizer. We present a theoretical framework which formulates the two tasks as complementary problems, and describe the symmetric way in which we have implemented their solution. Results of trials of the message identification systems using the Switchboard corpus of telephone conversations are reported.

1. INTRODUCTION

The task of topic identification is to select from a set of possibilities the topic that is most likely to represent the subject matter covered by a sample of speech. Similarly, speaker identification requires selecting from a list of possibilities the speaker most likely to have produced the speech. In this paper, we present a novel approach to the problems of topic and speaker identification which uses a large vocabulary continuous speech recognizer as a preprocessor of the speech messages.

The motivation for developing improved message identification systems derives in part from the increasing reliance on audio databases such as arise from voice mail, for example, and the consequent need to extract information from them. Technology that is capable of searching such a database of recorded speech and classifying material by subject matter or by speaker would have substantial value, much as text-based information retrieval technology has for textual corpora. Several approaches to the problems of topic and speaker identification have already appeared in the literature. For example, an approach to topic identification using wordspotting is described in [1] and approaches to the speaker identification problem are reported in [2] and [3].

Dragon Systems' approach to the message identification tasks depends crucially on the existence of a large vocabulary continuous speech recognition system. We view the tasks of topic and speaker identification as complementary problems: for topic identification, the speaker is irrelevant and only the subject matter is of interest; for speaker identification, the reverse is true. For efficiency of computation, in either case we first use a speaker-independent topic-independent recognizer to transcribe the speech messages. The resulting output is then scored using topic-sensitive or speaker-sensitive models.

This approach to the problem of message identification is based on the belief that the contextual information used in a full-scale recognition is invaluable in extracting reliable data from difficult speech channels. For example, unlike standard approaches to topic identification through spotting a small collection of topic-specific words, the approach via continuous speech recognition should more reliably detect keywords because of the acoustic and language model context available to the recognizer. Moreover, with large vocabulary recognition, the list of keywords is no longer limited to a small set of highly topic-specific (but generally infrequent) words, and instead can grow to include much (or even all) of the recognition vocabulary. The use of contextual information makes the message systems sufficiently robust that they are able to operate even with vocabulary sizes and noise environments that would make speech recognition extremely difficult for other applications.

To test our message identification systems, we have been using the "Switchboard" corpus of recorded telephone messages [4] collected by Texas Instruments and now available through the Linguistic Data Consortium. This collection of roughly 2500 messages includes conversations involving several hundred speakers. People who volunteered to participate in this program were prompted with a subject to discuss (chosen from a set that they had previously specified as acceptable) and were expected to talk for at least five minutes. We report results of topic identification tests involving messages on ten different topics using four and a half minutes of speech and speaker identification tests involving 24 speakers with test intervals containing as little as 10 seconds of speech.

In the next section, we describe the theoretical framework on which our message identification systems are based and discuss the dual nature of the two problems. We then describe how this theory is implemented in the current message processing systems. Preliminary tests of

the systems using the Switchboard corpus are reported in Section 4. We close with a discussion of the test results and plans for further research.

2. THEORETICAL FRAMEWORK

Our approach to the topic and speaker identification tasks is based on modelling speech as a stochastic process. For each of the two problems, we assume that a given stream of speech is generated by one of several possible stochastic sources, one corresponding to each of the possible topics or to each of the possible speakers in question. We are required to judge from the acoustic data which topic (or speaker) is the most probable source.

Standard statistical theory provides us with the optimal solution to such a classification problem. We denote the string of acoustic observations by A and introduce the random variable T to designate which stochastic model has produced the speech, where T may take on the values from 1 to n for the n possible sources. If we let p_i denote the prior probability of stochastic source i and assume that all classification errors have the same cost, then we should choose the source $T = \hat{i}$ for which

$$\hat{i} = \underset{i}{\text{argmax}}\ p_i\ P(A \mid T = i).$$

We assume, for the purposes of this work, that all prior probabilities are equal, so that the classification problem reduces simply to choosing the source i for which the conditional probability of the acoustics given the source is maximized.

In principle, to compute each of the probabilities $P(A \mid T = i)$ we would have to sum over all possible transcriptions W of the speech:

$$P(A \mid T = i) = \sum_W P(A, W \mid T = i).$$

In practice, such a collection of computations is unwieldy and so we make several simplifying approximations to limit the computational burden. First, we estimate the above sum only by its single largest term, i.e. we approximate the probability $P(A \mid T = i)$ by the joint probabiltiy of A and the single most probable word sequence $W = W^i_{\max}$. Of course, generating such an optimal word sequence is exactly what speech recognition is designed to do. Thus, for the problem of topic identification, we could imagine running n different speech recognizers, each modelling a different topic, and then compare the resulting probabilities $P(A, W^i_{\max} \mid T = i)$ corresponding to each of the n optimal transcriptions $W^i_{\max}$. Similarly, for speaker identification, we would run n different speaker-dependent recognizers, each trained on one of the possible speakers, and compare the resulting scores.

This approach, though simpler, still requires us to make many complete recognition passes across the speech sample. We further reduce the computational burden by instead producing only a single transcription of the speech to be classified, by using a recognizer whose models are both topic-independent and speaker-independent. Once this single transcription $W = W_{\max}$ is obtained, we need only compute the probabilities $P(A, W_{\max} \mid T = i)$ corresponding to each of the stochastic sources $T = i$.

Rewriting $P(A, W_{\max} \mid T = i)$ as

$$P(A \mid W_{\max}, T = i) * P(W_{\max} \mid T = i),$$

we see that the problem of computing the desired probability factors into two components. The first, $P(A \mid W, T)$, we can think of as the contribution of the acoustic model, which assigns probabilities to acoustic observations generated from a given string of words. The second factor, $P(W \mid T)$, encodes the contribution of the language model, which assigns probabilities to word strings without reference to the acoustics.

Now for the problem of topic identification, we wish to determine which of several possible topics is most likely the subject of a given sample of speech. Nothing is known about the speaker. We therefore assume that the same speaker-independent acoustic model holds for all topics; i.e. for the topic identification task, we assume that $P(A \mid W, T)$ does not depend on T. But we need n different language models $P(W \mid T = i)$, $i = 1, \ldots, n$. From the above factorization, it is then clear that in comparing scores from the different sources, only this latter term matters.

Symmetrically, for the speaker identification problem, we must choose which of several possible speakers is most likely to have produced a given sample of speech. While in practice, different speakers may well talk about different subjects and in different styles, we assume for the speaker identification task that the language model $P(W \mid T)$ is independent of T. But n different acoustic models $P(A \mid W, T = i)$ are required. Thus only the first factor matters for speaker identification.

As a result, once the speaker-independent topic-independent recognizer has generated a transcript of the speech message, the task of the topic classifier is simply to score the transcription using each of n different language models. Similarly, for speaker identification the task reduces to computing the likelihood of the acoustic data given the transcription, using each of n different acoustic models.

3. THE MESSAGE IDENTIFICATION SYSTEM

We now examine how this theory is implemented in each of the major components of Dragon's message identification system: the continuous speech recognizer, the speaker classifier, and the topic classifier.

3.1. The Speech Recognizer

In order to carry out topic and speaker identification as described above, it is necessary to have a large vocabulary continuous speech recognizer that can operate in either speaker-independent or speaker-dependent mode. Dragon's speech recognizer has been described extensively elsewhere ([5], [6]). Briefly, the recognizer is a time-synchronous hidden Markov model (HMM) based system. It makes use of a set of 32 signal-processing parameters: 1 overall amplitude term, 7 spectral parameters, 12 mel-cepstral parameters, and 12 mel-cepstral differences. Each word pronunciation is represented as a sequence of phoneme models called PICs (phonemes-in-context) designed to capture coarticulatory effects due to the preceding and succeeding phonemes. Because it is impractical to model all the triphones that could in principle arise, we model only the most common ones and back off to more generic forms when a recognition hypothesis calls for a PIC which has not been built. The PICs themselves are modelled as linear HMMs with one or more nodes, each node being specified by an output distribution and a double exponential duration distribution. We are currently modelling the output distributions of the states as tied mixtures of double exponential distributions. The recognizer employs a rapid match module which returns a short list of words that might begin in a given frame whenever the recognizer hypothesizes that a word might be ending. During recognition, a digram language model with unigram backoff is used.

We have recently begun transforming our basic set of 32 signal-processing parameters using the IMELDA transform [7], a transformation constructed via linear discriminant analysis to select directions in parameter space that are most useful in distinguishing between designated classes while reducing variation within classes. For the speaker-independent recognizer, we sought directions which maximize average variation between phonemes while being relatively insensitive to differences within the phoneme class, such as might arise from different speakers, telephone channels, etc. Since the IMELDA transform generates a new set of parameters ordered with respect to their value in discriminating classes, directions with little discriminating power between phonemes can be dropped. We use only the top 16 IMELDA parameters for speaker-independent recognition. A different IMELDA transform, in many ways dual to this one, was employed by the speaker classifier, as described below.

For speaker-independent recognition, we also normalize the average speech spectra across conversations via blind deconvolution prior to performing the IMELDA transform, in order to further reduce channel differences. A fixed number of frames are removed from the beginning and end of each speech segment before computing the average to minimize the effect of silence on the long-term speech spectrum.

Finally, we are now building separate male and female acoustic models and using the result of whichever model scores better. While in principle, one would have to perform a complete recognition pass with both sets of models and choose the better scoring, we have found that one can fairly reliably determine the model which better fits the data after recognizing only a few utterances. The remainder of the speech can then be recognized using only the better model.

3.2. The Speaker Classifier

Given the transcript generated by the speaker-independent recognizer, the job of the speaker classifier is to score the speech data using speaker-specific sets of acoustic models, assuming that the transcript provides the correct text; i.e. it must calculate the probabilities $P(A \mid W, T = i)$ discussed above. Dragon's continuous speech recognizer is capable of running in such a "scoring" mode. This step is much faster than performing a full recognition, since the recognizer only has to hypothesize different ways of mapping the speech data to the required text – a frame-by-frame phonetic labelling we refer to as a "segmentation" of the script – and need not entertain hypotheses on alternate word sequences.

In principle, the value of $P(A \mid W, T)$ should be computed as the sum over all possible segmentations of the acoustic data, but, as usual, we approximate this probability using only the largest term in the sum, corresponding to the maximum likelihood segmentation. While one could imagine letting each of the speaker-dependent models choose the segmentation that is best for them, in our current version of the speaker classifier we have chosen to compute this "best" segmentation once and for all using the same speaker-independent recognizer responsible for generating the initial transcription. This ensures that the comparison of different speakers is relative to the same alignment of the speech and may yield an actual advantage in performance, given the imprecision of our probability models.

Thus, the job of the speaker classifier reduces to scoring the speech data given both a fixed transcription

and a specified mapping of individual speech frames to PICs. To perform this scoring, we use a "matched set" of tied mixture acoustic models – a collection of speaker-dependent models each trained on speech from one of the target speakers but constructed with exactly the same collection of PICs to keep the scoring directly comparable. Running in "scoring" mode, we then produce a set of scores corresponding to the negative log likelihood of generating the acoustics given the segmentation for each of the speaker-dependent acoustic models. The speech sample is assigned to the lowest scoring model.

In constructing speaker scoring models, we derived a new "speaker sensitive" IMELDA transformation, designed to enhance differences between speakers. The transform was computed using only voiced speech segments of the test speakers (and, correspondingly, only voiced speech was used in the scoring). As is common in using the IMELDA strategy, we dropped parameters with the least discriminating power, reducing our original 32 signal-processing parameters to a new set of 24 IMELDA parameters. These were the parameters used to build the speaker scoring models. It is worth remarking that, because these parameters were constructed to emphasize differences between speakers rather than between phonemes, it was particularly important that the phoneme-level segmentation used in the scoring be set by the original recognition models.

3.3. The Topic Classifier

Once the speaker-independent recognizer has generated a transcription of the speech, the topic classifier need only score the transcript using language models trained on each of the possible topics. The current topic scoring algorithm uses a simple (unigram) multinomial probability model based on a collection of topic-dependent "keywords". Thus digrams are not used for topic scoring although they are used during recognition. For each topic, the probability of occurrence of each keyword is estimated from training material on that topic. Non-keyword members of the vocabulary are assigned to a catch-all "other" category whose probability is also estimated. Transcripts are then scored by adding in a negative log probability for every recognized word, and running totals are kept for each of the topics. The speech sample is assigned to the topic with the lowest cumulative score.

We have experimented with two different methods of keyword selection. The first method is based on computing the chi-squared statistic for homogeneity based on the number of times a given word occurs in the training data for each of the target topics. This method assumes that the number of occurrences of the word within a topic follows a binomial distribution, i.e. that there is a "natural frequency" for each word within each topic class. The words of the vocabulary can then be ranked according to the P-value resulting from this chi-squared test. Presumably, the smaller the P-value, the more useful the word should be for topic identification. Keyword lists of different lengths are obtained by selecting all words whose P-value falls below a given threshold.

Unfortunately, this method does not do a good job of excluding function words and other high frequency words, such as "uh" or "oh", which are of limited use for topic classification. Consequently, this method requires the use of a human-generated "stop list" to filter out these unwanted entries. The problem lies chiefly in the falsity of the binomial assumption: one expects a great deal of variability in the frequency of words, even among messages on the same topic, and natural variations in the occurrence rates of these very high frequency words can result in exceptionally small P-values.

The second method is designed to address this problem by explicitly modelling the variability in word frequency among conversations in the same topic instead of only variations between topics. It also uses a chi-squared test to sort the words in the vocabulary by P-value. But now for each word we construct a two-way table sorting training messages from each topic into classes based on whether the word in question occurs at a low, a moderate, or a high rate. (If the word occurs in only a small minority of messages, it becomes necessary to collapse the three categories to two.) Then we compute the P-value relative to the null hypothesis that the distribution of occurrence rates is the same for each of the topic classes. Hence this method explicitly models the variability in occurrence rates among documents in a nonparametric way. This method does seem successful at automatically excluding most function words when stringent P-value thresholds are set, and as the threshold is relaxed and the keyword lists allowed to grow, function words are slowly introduced at levels more appropriate to their utility in topic identification. Hence, this method eliminates the need for human editing of the keyword lists.

4. TESTING ON SWITCHBOARD DATA

To gauge the performance of our message classification system, we turned to the Switchboard corpus of recorded telephone conversations. The recognition task is particularly challenging for Switchboard messages, since they involve spontaneous conversational speech across noisy phone lines. This made the Switchboard corpus a particularly good platform for testing the message identification systems, allowing us to assess the ability of the

continuous speech recognizer to extract information useful to the message classifiers even when the recognition itself was bound to be highly errorful.

To create our "Switchboard" recognizer, male and female speaker-independent acoustic models were trained using a total of about 9 hours of Switchboard messages (approximately 140 message halves) from 8 male and 8 female speakers not involved in the test sets. We found that it was necessary to hand edit the training messages in order to remove such extraneous noises as cross-talk, bursts of static, and laughter. We also corrected bad transcriptions and broke up long utterances into shorter, more manageable pieces.

Models for about 4800 PICs were constructed. We chose to construct only one-node models for the Switchboard task, both to reduce the number of parameters to be estimated given the limited training data and to minimize the penalty for reducing or skipping phonemes in the often rapid speech of many Switchboard speakers. A vocabulary of 8431 words (all words occurring at least 4 times in the training data) and a digram language model were derived from a set of 935 transcribed Switchboard messages involving roughly 1.4 million words of text and covering nearly 60 different topics. Roughly a third of the language model training messages were on one of the 10 topics used for the topic identification task.

For the speaker identification trials, we used a set of 24 test speakers, 12 male and 12 female. Speaker-dependent scoring models were constructed for each of the 24 speakers using the same PIC set as for the speaker-independent recognizer. PIC models were trained using 5 to 10 hand-edited message halves (about 16 minutes of speech) from each speaker.

The speaker identification test material involved 97 message halves and included from 1 to 6 messages for each test speaker. We tested on speech segments from these messages that contained 10, 30, and 60 seconds of speech. The results of the speaker identification tests were surprisingly constant across the three duration lengths. Even for segments containing as little as 10 seconds of speech, 86 of the 97 message halves, or 88.7%, were correctly classified. When averaged equally across speakers, this gave 90.3% accuracy. The results from the three trial runs are summarized in Table 1. It is worth remarking that even the few errors that were made tended to be concentrated in a few difficult speakers; for 17 of the 24 speakers, the performance was always perfect, and for only 2 speakers was more than one message ever misclassified.

Given the insensitivity of these results to speech duration, we decided to further limit the amount of speech available to the speaker classifier. The test segments used in the speaker test were actually concatenations of smaller speech intervals, ranging in length from as little as 1.5 to as much as 50.2 seconds. We rescored using these individual fragments as the test pieces.[1] Results remained excellent. For example, when testing only the pieces of length under 3 seconds, 42 of the 46 pieces, or 91.3%, were correctly classified (90.9% when speakers were equally weighted). These pieces represented only 19 of the 24 speakers, but did include our most problematic speakers. For segments of length less than 5 seconds, 177 of the 201 pieces (88.1%, or 89.4% when the 24 speakers were equally weighted) were correctly classified.

speech interval (seconds)	weighted by message (%)	weighted by speaker (%)
10	88.7	90.3
30	88.7	90.6
60	87.6	89.9

Table 1: Speaker identification accuracy for 97-message Switchboard test.

For the topic identification task, we used a test set of 120 messages, 12 conversations on each of 10 different topics. Topics included such subjects as "air pollution", "pets", and "public education", and involved several topics (for example, "gun control" and "crime") with significant common ground. For topic identification, we planned to use the entire speech message, but for uniformity all messages were truncated after 5 minutes and the first 30 seconds of each was removed because of concern that this initial segment might be artificially rich in keywords.

Keywords were selected from the same training messages used for constructing the recognizer's language model. This collection yielded just over 30 messages on each of the ten topics, for a total of about 50,000 words of training text per topic. Because this is relatively little for estimating reliable word frequencies, word counts for each topic were heavily smoothed using counts from all other topics. We found that it was best to use a 5-to-1 smoothing ratio; i.e. data specific to the topic were counted five times as heavily as data from the other nine topics.

Keyword lists of lengths ranging from about 200 words to nearly 5000 were generated using the second method of keyword selection. We also tried using the entire 8431-

[1] The initial speaker-independent recognition and segmentation were not, however, re-run so that such decisions as gender determination were inherited from the larger test.

word recognition vocabulary as the "keyword" list. The results of the initial runs, given in the second column of Table 2, were disappointing: performance fell between 70% and 75% in all cases.

#keywords	original (%)	recalibrated (%)
203	70.0	71.7
1127	71.7	85.0
2655	72.5	87.5
4658	74.2	87.5
8431	72.5	88.3

Table 2: Topic identification accuracy for 120-message Switchboard test.

It is worth noting that, as it was designed to, the new keyword selection routine succeeded in automatically excluding virtually all function words from the 203-word list. For comparison, we also ran some keyword lists selected using our original method and filtered through a human-generated "stop list". The performance was similar: for example, a list of 211 keywords resulted in an accuracy of 67.5%.

The problem for the topic classifier was that scores for messages from different topics were not generally comparable due to differences in the acoustic confusability of the keywords. When tested on the true transcripts of the speech messages, the topic classifier did extremely well, missing only 2 or 3 messages out of the 120 with any of the keyword lists. Unfortunately, when run on the recognized transcriptions, some topics (most notably "pets", with its preponderance of monosyllabic keywords) never received competitive scores.

In principle, this problem could be corrected by estimating keyword frequencies not from true transcriptions of training data but from their recognized counterparts. Unfortunately, this is a fairly expensive approach, requiring that the full training corpus be run through the recognizer. Instead, we took a more expedient course. In the process of evaluating our Switchboard recognizer, we had run recognition on over a hundred messages on topics other than the ten used in the topic identification test. For each of these off-topic messages, we computed scores based on each of the test topic language models to estimate the (per word) handicap that each test topic should receive. When the 120 test messages were rescored using this adjustment, the results improved dramatically for all but the smallest list (where the keywords were too sparse for scores to be adequately estimated). The improved results are given in the last column of Table 2.

5. CONCLUSIONS

As the Switchboard testing demonstrates, message identification via large vocabulary continuous speech recognition is a successful strategy even in challenging speech environments. Although the quality of the recognition as measured by word accuracy rates was very low for this task – only 22% of the words were correctly transcribed – the recognizer was still able to extract sufficient information to reliably identify speech messages. This supports our belief in the advantages of using articulatory and language model context.

We were surprised not to find a more pronounced benefit from using large numbers of keywords for the topic identification task. Our prior experience had indicated that there were small but significant gains as the number of keywords grew and, although such a pattern is perhaps suggested by the results in Table 2, the gains (beyond those in the recalibration estimates) are too small to be considered significant. It is possible that with better modelling of keyword frequencies or by introducing acoustic distinctiveness as a keyword selection criterion, such improvements might be realized.

Given the strong performance of both of our identification systems, we also look forward to exploring how much we can restrict the amount of training and testing material and still maintain the quality of our results.

References

1. R.C. Rose, E.I. Chang, and R.P. Lipmann, "Techniques for Information Retrieval from Voice Messages," *Proc. ICASSP-91*, Toronto, May 1991.
2. A.L. Higgins and L.G. Bahler, "Text Independent Speaker Verification by Discriminator Counting," *Proc. ICASSP-91*, Toronto, May 1991.
3. L.P. Netsch and G.R. Doddington, "Speaker Verification Using Temporal Decorrelation Post-Processing," *Proc. ICASSP-92*, San Francisco, March 1992.
4. J.J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. ICASSP-92*, San Francisco, March 1992.
5. J.K. Baker *et al.*, "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems," *Proc. DARPA Speech and Natural Language Workshop*, Harriman, New York, February 1992.
6. R. Roth *et al.*, "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data," *Proc. ICASSP-93*, Minneapolis, Minnesota, April 1993.
7. M.J. Hunt, D.C. Bateman, S.M. Richardson, and A. Piau, "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," *Proc. ICASSP-91*, Toronto, May 1991.