SPEAKER VERIFICATION THROUGH LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Michael Newman, Larry Gillick, Yoshiko Ito, Don McAllaster, Barbara Peskin

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

ABSTRACT

We present a study of a speaker verification system for telephone data based on large-vocabulary speech recognition. After describing the recognition engine, we give details of the verification algorithm and draw comparisons with other systems. The system has been tested on a test set taken from the Switchboard corpus of conversational telephone speech, and we present results showing how performance varies with length of test utterance, and whether or not the training data has been transcribed. The dominant factor in performance appears to be channel or handset mismatch between training and testing data.

1. INTRODUCTION

For a number of years, Dragon Systems has advocated that large vocabulary continuous speech recognition (LVCSR) is the best starting point for voice-information applications such as topic, speaker, and language identification. In the past we have reported on a number of successful experiments [1, 2, 3]. In this paper, we present further results on speaker verification using conversational telephone data.

The motivation for our approach is that a full-scale recognizer should be able to use higher-level linguistic and phonetic knowledge to extract more information from a speech signal than a simpler system such as a gaussian mixture model. A number of other sites have also tried this approach. In particular, see [4] for a discussion and for further references.

A speaker verification system has many potential applications. For example, one might wish to set up a voice-mail system in which calls are automatically screened to prioritize messages from certain predefined callers. Alternatively, it might be part of a user-authentication system for dial-in computer access.

In this paper we give a brief overview of Dragon's recognition system, before describing the speaker verification algorithm. We present some results obtained from testing on a corpus of conversational telephone speech, and discuss where we think the challenges lie for cutting the error rate.

2. THE RECOGNITION ENGINE

The heart of Dragon's speaker verification system is a speaker-independent large vocabulary continuous speech recognizer, trained from the Switchboard corpus of spontaneous conversational dialogs recorded over long-distance telephone lines [5] (A more complete description of Dragon's recognition system is given in [6]). Over the last two years, recognition performance has improved considerably, and the word error rate of our best system has dropped from nearly 80% to under 40%. Amongst the many changes, the acoustic modeling now uses a decision tree algorithm to cluster nodes of triphones, with the output distribution for each cluster modeled as a general mixture of gaussians. In addition, the signal-processing encompasses an expanded parameter set, with improved channel normalization.

2.1 Signal Processing

The front end produces a 44-dimensional feature vector every 10ms, consisting of 8 spectral and 12 cepstral parameters, together with 12 each of first and second cepstral differences. An IMELDA transformation is applied to reduce the dimension of the feature vector from 44 to 24 [7]. To mitigate possible effects on verification arising from channel or handset variation, the signal-processing was band-limited to 3400Hz, and an amplitude-based channel normalization applied. Apart from the band-limiting, we used our standard recognition signal processing.

2.2 Acoustic and Language Models

The recognition models used for this test were built from around 60 hours of Switchboard acoustic training data, and comprised around 13,000 output distributions, each with up to 16 gaussian components. The back-off bigram language model was trained from around 1.6 million words of Switchboard text. All words occurring three or more times were retained, resulting in a vocabulary of around 10,000 words, and all in-vocabulary bigrams were kept.

Partly because of the constraints of the evaluation (such as the short length of the test utterances), and partly to keep the system simple, we made a number of compromises in the design of the system. In particular, we ran without speaker normalization [8], with gender-independent models, and with rather aggressive pruning thresholds, resulting in recognition performance somewhat inferior to the best possible. Overall, we estimate that recognition ran at approximately 20 times real time on a 200MHz SGI Indigo II, with a word error rate of around 55%.

Another measure of the recognition performance, which we believe to be more appropriate in the context of speaker

verification, is the fraction of frames of speech assigned to the correct phoneme by a time-alignment of the recognition transcript. We measured this number to be approximately 65%.

3. THE SPEAKER VERIFICATION ALGORITHM

The algorithm used here is a refinement of one previously reported in [1]. More details are given below, but the summary is as follows.

For each target speaker, we adapt a set of models to data from that speaker. This adaptation can be either supervised, requiring a transcription of the training data, or unsupervised, in which case the recognizer is used to produce the transcript.

Each test utterance is run through the recognizer, to produce a recognition transcript and a time-alignment of the hypothesized words. The time-alignment is then rescored using the target-adapted models, from which we compute a measure of how well the test data matches the target.

3.1 Scoring Models

The first requirement for a set of scoring models is that they be small enough that most of the output distributions get adapted during training. The full-scale recognition models described earlier would be far too large. Typical scoring models consist of between 130 and 1000 output distributions, i.e. a factor of between 10 and 100 smaller than the recognition models. In practice we use monophone (context-independent) models for all but the most generous of training conditions.

In a previous experiment [1], the scoring models were built entirely from the training data for the target, which required comparatively large quantities of transcribed training data, and imposed severe restrictions on the size and structure of the models. For this test we used a different procedure. We built one set of speaker-independent models (using the same acoustic training data used for building the recognition models), and the speaker-dependent models were derived from these by standard Baum-Welch adaptation on the target training data. We believe that this method is both simpler and more robust.

An important question was what IMELDA transformation to use, if any. Since the purpose of the transformation is to pick those combinations of parameters which maximize variation between phonetic classes while minimizing variation within classes (such as those due to differences between speakers), it was expected that the discarded parameters would contain a lot of information about speaker identity. For a variety of reasons, we preferred not to omit the transformation altogether: for example, one of its desirable properties is that it reduces correlations between parameters. Instead we applied the standard transformation but kept all 44 parameters. Compared to the alternative of truncating at 24 parameters, this turned out to be a considerable improvement for most training conditions.

In the past, we have tried using a special IMELDA transformation designed specifically to accentuate differences between speakers, rather than to suppress them [1]. With the current system, we were not able to demonstrate any advantage from such a specialized transformation, compared to using the standard transformation but keeping all parameters.

3.2 Training Algorithm

For training, a sample of speech data is required for each target speaker. If the text is known, we can use supervised Baum-Welch adaptation to get slightly better performance: otherwise, we first run recognition to produce a working transcription, before running unsupervised Baum-Welch adaptation. In either case, only the model means and mixture weights are adapted.

A comparison of supervised and unsupervised adaptation is given later. It appears that given enough training data, the difference is small, especially compared to other factors such as the length of the test utterances.

3.3 Scoring of Test Data

Testing consists of two stages. First, one runs recognition on each of the test utterances. For the purposes of speaker verification, the recognizer produces not just a recognition transcript, but also a detailed time-alignment of the speech data. This alignment, which we call a *segmentation*, contains information about the start and end times of each of the HMM nodes of the phonemes in the recognition transcript.

In the second stage, the segmentation is scored against the adapted models for a particular target speaker. Each frame of the test utterance is scored against the single output distribution to which it is mapped by the segmentation, except that we omit frames which have been labeled as silence. The score for the utterance is just the sum of the individual frame scores.

However, this score must be corrected in two ways. First, we subtract the equivalent score of the utterance obtained using the unadapted scoring models. Second, we define the *speaker score* as the score difference (adapted - unadapted), divided by the number of frames of speech, which is essentially a normalized log-likelihood ratio. For a given target speaker, the test utterances can then be ranked according to the speaker score.

3.4 Speed of Algorithm

The slowest step in the verification process is the recognition of the test data. However, no particular attempt was made to optimize the recognition for speed. Two possibilities would be to cut down the size of the vocabulary or make the acoustic models much smaller, both of which would greatly speed up

recognition with minimal likely effect on the speaker identification performance.

In contrast, the time taken to score the segmentations is negligible. For each frame of test data, only one output distribution needs to be computed. As a result, the system scales very efficiently with the number of target models which need to be scored.

4. COMPARISON WITH OTHER ALGORITHMS

We would like to draw particular attention to a number of aspects of our algorithm.

The first is our use of a single speaker-independent background model, built by pooling the data from a large number of training speakers. This should be contrasted with the widely-used "cohort method," which employs a large number of speaker-dependent background models built from individual speakers. We believe that building a speaker-independent model is a better way to smooth the data from the individual training speakers.

The second is our use of Bayesian adaptation to derive the target-adapted model from the speaker-independent model. One alternative would be to build each target model from scratch, using just the target training data, but for our modeling procedure this method has problems unless there is enough training data for each of the output distributions. But even for systems using simple mixture models, a case can be made that adapting to a target from speaker-independent models which cover all of acoustic space provides a graceful back-off strategy in regions of acoustic space for which no target adaptation data was seen, especially when training data is sparse.

The next two points concern the use of a segmentation derived from LVCSR. The first advantage is that one can make very close comparisons between a node of a phoneme occurring in the test data, and examples of the same node found in the adaptation training data. If we have enough training data to justify using scoring models incorporating a lot of phonetic context, the comparisons become even tighter; but even for monophone models, we are always making comparisons between examples of the same phoneme. In particular, if a specific output distribution does not get adapted during training to the target, frames mapped to that output distribution do not contribute to the speaker score: when the score of the unadapted model is subtracted, the contribution of these frames cancels exactly. As a corollary, if there is a score difference for a particular frame, we know that the model is expressing an opinion only on the basis of corresponding data seen during training.

In addition, higher-level sources of information (such as word and/or phoneme sequence constraints), which come in during the recognition stage, are implicitly being included in the speaker ID decision. However, this extra information only helps if the recognition is sufficiently accurate.

For frames which have been misrecognized, our system presumably performs as a crude mixture model. We would expect our system to have better performance than a mixture-model system on frames correctly recognized, and worse performance on the others. Overall there is a trade-off between the two, and one factor determining which method performs better is the recognition error rate during both training and testing. For the system described here, we compared forced alignments of the true and recognition transcripts, and measured that approximately 65% of frames containing speech were recognized correctly. We cannot say whether this number is "high enough," but we can promise that the percentage correct will continue to improve with time.

5. TEST RESULTS

5.1 Test Corpus

The system was tested on the 1995 NIST-administered speaker verification database, which consists of data taken from the Switchboard corpus. For each of 26 "target" speakers (15 male, 11 female) training data is specified from 4 conversation halves, and test data from around 6 more. A further 80 speakers (33 male, 47 female) are provided as "impostors," for whom test segments only are provided.

There were three training conditions for which data was specified: 10 seconds, 30 seconds, and "unlimited" use of four sides of conversation (corresponding to approximately 10 minutes of speech).

For each target speaker there were on average 36 test segments taken from 6 different message halves (in general recorded over at least two hand-sets), each segment containing a nominal 5 seconds of speech. In total, there were 936 5-second pieces for the targets and 2574 for the impostors. For longer test segments, the 5 second pieces were concatenated to provide 10 and 30 second test conditions.

For each conversation in the Switchboard corpus, a scrambled version of the phone numbers of each party is given. This provides us with a (rather unreliable) way of inferring cases of handset and channel mismatch between the target training and test messages, and allows us to get a crude idea of how results are affected by such a mismatch. Unfortunately, almost all the training messages for a given speaker originated from the same phone number, which limits our ability to explore the effect of handset variation during training.

5.2 Evaluation Results

We present here results for the "unlimited" training condition, for which we used scoring models with about 730 output distributions, with up to 16 gaussians in each mixture.

The two tables give results for supervised and unsupervised adaptation, with the rows corresponding to different lengths of test utterances. The columns show the effect of handset mismatch. (In the "combined" column, note that about 3/5 of the test utterances come from mismatched handsets.) The results are the average false-rejection rates over all targets for a false-acceptance rate of 3%. The standard error is approximately ±4%, where the dominant contribution to the uncertainty comes from the very great variation in performance between speakers.

Test	Matched	Mismatched	Combined
5	4.5	12.3	10.3
10	4.2	6.8	6.4
30	4.8	4.3	4.5

Table 1: Results for supervised training

Test	Matched	Mismatched	Combined
5	4.5	14.7	10.7
10	4.8	10.7	8.3
30	4.8	5.4	5.2

Table 2: Results for unsupervised training

We would like to make a number of comments on these numbers. Firstly, these results are computed using impostor messages of both genders. As detection of cross-gender impostors is much simpler than within-gender impostors, excluding cross-gender impostors has the effect of doubling the false rejection rate, at given false acceptance.

Secondly, it is clear that channel mismatch is the dominant contribution to the error rate. In fact, the errors are concentrated on a small number of speakers, and on specific conversation halves from those speakers. Listening to the messages from the worst offenders, there are clear audible differences between the matched and mismatched data. Bear in mind also that the message labels "matched" and "mismatched" are only approximate, and it may be that the majority of errors on the nominally matched test data may still be due to mismatch (e.g. from a different handset on the same phone line as the training data).

It also turns out that the error rate for male targets is much greater than for females[1]. To the best of our knowledge, this is simply a consequence of the greater handset variability present in this corpus for male speakers, rather than representing any intrinsic difference in the speakers' acoustics.

Thirdly, there is some advantage from knowing the true transcript of the target adaptation data, but the performance hit from moving to unsupervised adaptation is relatively small, compared to other sources of error.

Finally, because there is so much variation in performance between speakers, it is very dangerous to compare numbers obtained using different test sets.

[1] This effect has also been noted elsewhere: e.g. [9] (We thank Doug Reynolds for generously providing us with an advance copy of his paper).

6. CONCLUSIONS

We have demonstrated an effective and competitive speaker-verification system based on LVCSR, which can run with or without transcriptions of the target training data. We believe that the primary challenge in the near future will be to understand and solve the problem of channel mismatch.

7. REFERENCES

1. B. Peskin et al., "Topic and Speaker Identification via Large Vocabulary Continuous Speech Recognition," ARPA Workshop on Human Language Technology, Princeton, March 1993.
2. S. Mendoza et al., "Automatic Language Identification through Large Vocabulary Continuous Speech Recognition," Proc. ICASSP-96, Atlanta, May 1996.
3. B. Peskin et al., "Improvements in Switchboard Recognition and Topic Identification," Proc. ICASSP-96, Atlanta, May 1996.
4. J. L. Gauvain et al., "Experiments with speaker verification over the telephone," Eurospeech '95, Madrid, September 1995.
5. J. Godfrey et al., "SWITCHBOARD: Telephone Speech Corpus for Research and Development," Proc. ICASSP-92, San Francisco, March 1992.
6. B. Roth et al., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer," Proc. Spoken Language Systems Technology Workshop, Austin, January 1995.
7. M. Hunt et al., "An investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," Proc. ICASSP-91, Toronto, May 1991.
8. S. Wegmann et al., "Speaker Normalization on conversational telephone speech," Proc. ICASSP-96, Atlanta, May 1996.
9. D. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the Switchboard corpus," Proc. ICASSP-96, Atlanta, May 1996.