Language Identification via Large Vocabulary Speaker Independent Continuous Speech Recognition

Steve Lowe, Anne Demedts, Larry Gillick, Mark Mandel, Barbara Peskin

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

ABSTRACT

The goal of this study is to evaluate the potential for using large vocabulary continuous speech recognition as an engine for automatically classifying utterances according to the language being spoken. The problem of language identification is often thought of as being separate from the problem of speech recognition. But in this paper, as in Dragon's earlier work on topic and speaker identification, we explore a unifying approach to all three message classification problems based on the underlying stochastic process which gives rise to speech. We discuss the theoretical framework upon which our message classification systems are built and report on a series of experiments in which this theory is tested, using large vocabulary continuous speech recognition to distinguish English from Spanish.

1. INTRODUCTION

In this paper we describe preliminary work being conducted at Dragon Systems exploring the use of large vocabulary continuous speech recognition as an engine for automatically classifying spoken utterances by language. Several approaches to the problem of language identification have already appeared in the literature, but they generally address the problem as quite separate from the problem of speech recognition. For example, LIMSI [1] has reported results in French-English discrimination via phone recognition and a number of sites, such as OGI [2], have performed language classification by using broad phonetic labels and analyzing sets of phonological "features".

Our approach to the problem of language identification grows naturally out of our model for the underlying stochastic process giving rise to speech. In earlier papers ([3], [4]) we have described our unified approach to the problems of topic and speaker identification via large vocabulary continuous speech recognition and demonstrated the success of this strategy even in classifying speech data in domains where the recognition task is far too difficult to obtain accurate transcriptions. We believe that the contextual information – both acoustic and language model – available in full-scale large vocabulary continuous speech recognition is invaluable in extracting reliable data from difficult speech channels. We now examine how this same framework supports work on the problem of language identification.

In the next section we describe the theoretical foundations upon which our message classification systems are based and discuss some simplifying approximations introduced in their implementation. We then describe our initial testing of English-Spanish discrimination, primarily work with microphone data using our Wall Street Journal speech recognition system, but also work we are now beginning in language identification on telephone speech. Finally, we discuss some lessons learned from these early explorations and suggest plans for future work.

2. THEORETICAL FRAMEWORK

We briefly review the theoretical background described in our earlier papers [3] and [4]. Our approach to the message classification problem – for topic, speaker, or language identification – is based on modelling speech as a stochastic process. We assume that a given stream of speech is generated by one of several possible stochastic sources, one corresponding to each of the languages (or topics or speakers) in question. We are faced with the problem of deciding, based on the acoustic data alone, which is the true source of the speech.

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 speech 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 so to limit the computational burden we introduce a simplifying approximation. Instead of computing the full probability $P(A \mid T = i)$, we approximate the sum by its largest term: the joint probability 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, we could imagine running n different speech recognizers, one trained in each of the n languages, and then compare the resulting probabilities $P(A, W^i_{\max} \mid T = i)$ corresponding to each of the n optimal transcriptions $W^i_{\max}$. The speech would then be assigned to the language whose recognizer produced the best score.

This approach still requires us to make multiple recognition passes across the test speech, one pass for each stochastic source. In the cases of topic and speaker identification studied earlier, we were able to further limit the demand on the recognizer by producing a single "best" transcription $W = W_{\max}$, using a speaker-independent topic-independent recognizer, to approximate the optimal transcriptions produced by each of the stochastic sources $T = i$. The corresponding probabilities $P(A, W_{\max} \mid T = i)$ were then computed by rescoring this "best" transcription using either topic-specific language models in the case of topic identification, or speaker-specific acoustic models for speaker identification. (See the above-cited articles for further details.)

For the problem of language identification, we do not have the option of obtaining a single "language-independent" transcription: the transcription depends inextricably on the language we are recognizing. Thus it would appear that in this case we are forced to run several recognition passes on each test utterance, one for each language in question, or at the very least perform the recognition using a recognizer capable of running several sets of models/languages simultaneously and allow the best performing language to threshold hypotheses from poorer performing ones.

We are currently working to develop parallel recognizers trained on telephone-quality speech in a number of languages which should allow us to perform exactly this experiment. This effort is described in more detail below. While this development effort is under way, we are exploring the possibility of performing two-language discrimination using a single recognizer trained in one of the languages. There are several ways of using the theory above to construct a one-recognizer test. Using two recognizers, one trained in each language, we would estimate $P(A \mid T = i)$ for each of the two languages and then, as described above, assign the speech sample to the recognizer producing the best score. Alternatively, we could look at the log likelihood ratio

$$S = \log \frac{P(A \mid T = 1)}{P(A \mid T = 2)},$$

and make the assignment based on a threshold $S = S_0$, assigning the sample to language #1 if $S > S_0$, and to language #2 otherwise. With only one recognizer, trained, say, in language #1, we could simply impose a threshold on $\log P(A \mid T = 1)$ alone, assigning the speech sample to language #1 if the score was good enough and to language #2 otherwise. This naive solution suffers from a number of problems, most significantly that the recognition score depends on many variables unrelated to the language – such as speaker, channel, or phonetic content – that are not properly controlled for without the normalizing effect of the denominator in the likelihood ratio.

In the experiments described below, we have explored the possibility of controlling for these confounding factors in the acoustics by introducing a normalization based on the acoustics of individual speech frames. In Dragon's speech recognition system the acoustics for each frame are represented in terms of a feature vector and the recognizer's acoustic models consist in part of probability distributions on the occurrence of these feature vectors. We refer to these models, the output distributions for nodes of our hidden Markov models, as PELs (for "phonetic elements"). In normal speech, the PEL sequences we expect to see are constrained by the phonemic sequences within the words in the recognizer's vocabulary, but as a group the PELs should provide good coverage of that region of acoustic parameter space where speech data lie. To normalize the recognition scores for the one-recognizer tests we compute a second score using, for each speech frame, the probability corresponding to whichever PEL model – unconstrained by word-level hypotheses – best matches the acoustics in that frame. The product of these frame-by-frame probabilities provides a second score (referred to below as the "maximal acoustic score") that can be used as the denominator in the log likelihood ratio above. Presumably, when the speech being recognized is in the language of the recognizer, this optimal frame-by-frame PEL sequence should be reasonably close to the true PEL sequence, but when the language is different the maximal acoustic score should be far better than the score produced by the recognizer.

This normalization using best-matching PELs captures some sense of how well the acoustic signal fits the recognizer's models independent of constraints imposed by the words in the language we are recognizing. Thus, we expect it to help minimize sources of variability unrelated to differences between languages. However, it may be that different languages cover somewhat different parts of acoustic parameter space and, as we shall see below, this normalization may also have the undesirable side effect of normalizing away this language-rich information as well.

The scores produced by the language identification system are negative log probabilities, normalized by the number of frames in the utterance. Thus, in practice, the log likelihood ratio translates to a simple difference of recognizer (or recognizer and maximal acoustic) scores.

3. INITIAL EXPERIMENTS

Our approach to language identification depends crucially on the existence of a large vocabulary continuous speech recognition system, so in order to test the feasibility of our language identification strategy, we turned to our primary LVCSR system, the Wall Street Journal recognizer developed under the ARPA SLS program. This recognition system has been described extensively elsewhere (see, for example, [5] and [6]). We review its basic properties here.

The recognizer is a time-synchronous hidden Markov model based system. It makes use of a basic set of 32 signal-processing parameters: 1 overall amplitude term, 7 spectral parameters, 12 mel-cepstral parameters, and 12 mel-cepstral differences. Our standard practice is to employ an 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 speaker-independent recognition we choose directions which maximize the average variation between phonemes while being relatively insensitive to differences within the phoneme class, such as might arise from different speakers, 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 used only the top 16 IMELDA parameters for speaker-independent recognition, divided into four 4-parameter streams. For speaker-independent recognition, we also normalize the average speech spectra across utterances via blind deconvolution prior to performing the IMELDA transform, in order to further reduce channel 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 – the PELs referred to above – and a double exponential duration distribution. The output distributions of the states were modelled as tied mixtures of Gaussian distributions. The recognizer used for our language identification work was trained from the standard WSJ0 SI-12 training speakers (using 7200 sentences in all, totalling about 16 hours of speech data). Because Dragon's in-house recordings are made at 12 kHz, rather than the WSJ standard of 16 kHz, the training data was first down-sampled to 12 kHz before training the models. For these experiments, the standard WSJ 20K vocabulary and digram language model (based on about 40 million words of newspaper text) were used.

For the language identification test material, three bilingual Dragon employees each recorded 20 English sentences taken from a current issue of the Wall Street Journal. For Spanish data, they read 20 Spanish sentences taken from the financial section of a current issue of America Economia, a Spanish language news magazine. The resulting test corpus thus consisted of 60 English and 60 Spanish utterances, averaging about 8 seconds in length and recorded on a Shure SM-10 microphone at a 12 kHz sample rate.

Using the simple (unnormalized) one-recognizer strategy described above, we obtained an 83% probability of detection at the equal error point (i.e. the point where the probability of detection equals the probability of false alarm). After rescoring using the maximal acoustic score normalization, this figure improved to 95%. It is also worth noting that using the maximal acoustic score alone we obtained a result of 68%. Such a non-speech-based strategy is similar in spirit to approaches to language identification using subword acoustic features rather than full speech recognition. The results are summarized in the first line of Table 1.

Inspired by the success of this initial trial on read speech data, we next turned to an assessment of performance on spontaneous telephone speech, using as test material speech drawn from the OGI corpus [8] of recorded telephone messages. This multi-lingual corpus contains "evoked monologues" from 90 speakers in each of ten languages. For our in-house testing, we selected 10 Spanish

	R	$R-M$	M
original	83%	95%	68%
no IMELDA	82%	90%	73%

Table 1: English-Spanish discrimination using the Wall Street Journal recognizer. The figures give the probability of detection at the equal error point for the recognizer score R, the maximal acoustic score M, and the normalized recognition score $R - M$.

and 10 English calls from among the designated OGI training material. We used the "story-bt" segments of these calls, which run up to 50 seconds in length. Prior to testing, these were broken at pauses into shorter segments using an automatic "acoustic chopper". This resulted in 102 English segments and 104 Spanish segments, each less than about 10 seconds in length.

For our first foray into language discrimination on telephone speech, we used the same SWITCHBOARD speech recognition system used in our topic and speaker identification work. This recognizer was trained – and for topic and speaker identification, tested – on telephone conversations from the SWITCHBOARD corpus [9], collected by TI and now available through the Linguistic Data Consortium. Details of the recognizer are given in [3] and [4]; it is similar in structure to our Wall Street Journal recognizer, but was trained on only about 9 hours of conversational telephone speech. The recognition performance even on SWITCHBOARD data is very weak, although it is still capable of extracting sufficient information to achieve good topic and speaker identification performance. When used for language identification on OGI utterances the results were disappointing: it was unable to perform at anything better than chance levels, even aided by the acoustic normalization scoring.

Dragon Systems is currently engaged in an effort to collect telephone data in a number of languages using an "evoked monologue" format similar to that used for the OGI corpus. Our first collection efforts focussed on Spanish data collection and, using about 3 hours of our own Spanish data and an additional 15 minutes of OGI Spanish training material, we built a rudimentary recognition system for Spanish telephone speech. It has a 5K vocabulary and a digram language model trained from 30 million words of Spanish newswire data.

This new Spanish recognizer achieved a 72% probability of detection at the equal error point on the OGI test data when using the simple (unnormalized) recognition scores. In this case, unlike for the Wall Street Journal and SWITCHBOARD experiments, there was no advantage to using acoustic normalization techniques. Instead, using the maximal acoustic score for normalization actually *degraded* performance: probability of detection dropped to only 66% at the equal error point. Interestingly, for this system, the maximal acoustic score alone did as well as the regular recognition score: 74% probability of detection at the equal error point.

We conjectured that this behavior might be due at least in part to the fact that the Spanish recognizer, unlike the Wall Street Journal and SWITCHBOARD recognizers, did not employ our usual speaker-independent IMELDA transformation. Recall that this transform is designed to emphasize differences between phoneme classes while minimizing differences within each class and so may well be suppressing language-informative phonetic distinctions. Acoustic normalization may help to overcome this deficiency, but may be unnecessary – or even counterproductive – with non-IMELDA models. To test this hypothesis, we re-ran the WSJ language identification test, but this time with models trained without the IMELDA transform. The results are reported in the second line of Table 1. Without IMELDA, the recognition itself was somewhat less accurate, but the language identification performance using the recognizer scores was essentially unchanged. However, as expected, the performance of the maximal acoustic score alone improved without IMELDA, even though that performance remained well below that of the full recognizer, and there was a corresponding drop in the normalized score performance.

4. DISCUSSION AND FUTURE WORK

As the initial Wall Street Journal trials indicate, large vocabulary continuous speech recognition is clearly a successful strategy for language discrimination on high-quality microphone speech. Unlike some other trials of language identification on read speech, the WSJ test was designed to control for such confounding factors as speaker and channel differences. The chief drawbacks of the test were its small size and the possible bias introduced by a recognizer so tuned to the Wall Street Journal grammar (despite our best efforts to choose Spanish data in a matched domain), but despite these objections the evidence for an LVCSR approach to language identification is very strong.

The performance in the much harder domain of spontaneous telephone speech is more difficult to interpret. The preliminary testing described above differed in so many respects from the read speech experiments that it is hard to tease apart the effects without further experimentation. We look forward to exploring the roles of

several components, for example:

- the use of IMELDA transformations for the speech models — As suggested by the experiments above, the use of a speaker-independent IMELDA transformation, while unquestionably improving the recognition performance, may be removing important clues about language differences. To take advantage of the recognition boost without sacrificing language-rich information, the best strategy may be to perform an initial recognition pass using IMELDA models to generate a transcription, but to score the transcript using non-IMELDA models – a two-pass strategy similar to that used in our speaker identification work. Indeed, we may want to use a "language sensitive" IMELDA transformation – i.e. one which chooses directions in acoustic parameter space most useful in distinguishing languages rather than phonemes – for the scoring pass, in much the same way that we employed a "speaker sensitive" IMELDA in our work on speaker classification.

- acoustic normalization of recognition scores — In the case of one-recognizer tests, it seems important to have some way of controlling for such sources of variation as speaker differences, which should play no role in language discrimination. The acoustic normalization described above is a simple attempt to achieve this. However, like the speaker-independent IMELDA, it may also be removing important information about which regions of acoustic space different languages inhabit.

- recognition quality — It is our experience from our topic and speaker work that small improvements in recognition performance can yield enormous gains in classification tasks. However, in order to take advantage of the contextual information available in large vocabulary CSR, the recognition must exceed a certain minimal level of performance. We believe that none of our telephone speech recognition systems yet achieve the recognition levels needed to demonstrate the advantage of large vocabulary CSR as a language classification engine, but that this minimal level is well within reach.

We are now focussing attention on the task of improving our telephone speech recognizers. The Spanish recognizer was constructed in under a week's time when the Spanish telephone data became available and could still profit from such simple measures as further training iterations. We also hope to introduce into our telephone recognition systems such improved features as phonetically-tied mixture modelling, now used routinely in our microphone speech recognizers. The task of recognizing natural speech appears to be much more difficult than recognizing read speech and may require new techniques to address the problems of speaking rate, word contraction or fragmentation, and non-speech events.

Dragon's telephone data collection effort is also continuing. We hope to have at least five hours of recorded telephone speech in each of seven languages by the end of 1994 with further collections scheduled for next year. This will allow us to create parallel recognition systems in a number of languages and finally run a two- (or n-) recognizer test of language identification. In particular, we will be collecting English telephone data and look forward to building a new English telephone speech recognizer more directly analogous to our current Spanish system. It should be interesting to see how this new English recognizer and the SWITCHBOARD recognizer perform on each other's data.

We believe that these improvements should allow us to achieve the strong language identification performance we anticipate, based on our earlier work on topic and speaker identification.

References

1. J.-L. Gauvain and L.F. Lamel, "Identification of Non-Linguistic Speech Features," *Proc. ARPA HLT Workshop*, Princeton, March 1993.
2. Y.K. Muthusamy and R.A. Cole, "Automatic Segmentation and Identification of Ten Languages Using Telephone Speech," *Proc. Intl. Conf. on Spoken Language Processing 92*, Banff, October 1992.
3. B. Peskin *et al.*, "Topic and Speaker Identification via Large Vocabulary Continuous Speech Recognition," *Proc. ARPA HLT Workshop*, Princeton, March 1993.
4. L. Gillick *et al.*, "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification Using Telephone Speech," *Proc. ICASSP-93*, Minneapolis, Minnesota, April 1993.
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.
8. Y.K. Muthusamy, R.A. Cole, and B.T. Oshika "The OGI Multi-Language Telephone Speech Corpus," *Proc. Intl. Conf. on Spoken Language Processing 92*, Banff, October 1992.
9. J.J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. ICASSP-92*, San Francisco, March 1992.