LANGUAGE IDENTIFICATION USING PHONEME RECOGNITION AND PHONOTACTIC LANGUAGE MODELING*

Marc A. Zissman
Lincoln Laboratory, Massachusetts Institute of Technology
244 Wood Street
Lexington, MA 02173-9108
USA
Voice: (617) 981-2547
Fax: (617) 981-0186
E-mail: MAZ@SST.LL.MIT.EDU

ABSTRACT

A language identification technique using multiple single-language phoneme recognizers followed by n-gram language models yielded top performance at the March 1994 NIST language identification evaluation. Since the NIST evaluation, work has been aimed at further improving performance by using the acoustic likelihoods emitted from gender-dependent phoneme recognizers to weight the phonotactic likelihoods output from gender-dependent language models. We have investigated the effect of restricting processing to the most highly discriminating n-grams, and we have also added explicit duration modeling at the phonotactic level. On the OGI Multi-language Telephone Speech Corpus, accuracy on an 11-language, closed-set, language identification task has risen to 89% on 45-s utterances and 79% on 10-s utterances. Two-language classification accuracy is 98% and 95% for the 45-s and 10-s utterances, respectively. Finally, we have started to apply these same techniques to the problem of dialect identification.

1. INTRODUCTION

This paper describes ongoing work at M.I.T. Lincoln Laboratory to research and develop high performance language identification (LID) and dialect identification (DID) systems. After conducting a number of studies over the past few years in which we compared the performance of a variety of different LID approaches [12, 13] using the Oregon Graduate Institute (OGI) Multi-language Telephone Speech Corpus [9] (described in Section 2), we have more recently focused our efforts on what we call the PRLM-P system, which stands for **P**honeme **R**ecognition followed by **L**anguage **M**odeling performed in **P**arallel. The baseline version of PRLM-P, which was introduced last year [13] and is reviewed in Section 3, comprises a bank of single-language phoneme recognizers followed by phonotactically motivated, n-gram language models. In Section 4, the use of gender-dependent phoneme recognizers and n-gram language models is discussed. Their use improves performance significantly at the cost of increased computational complexity. Next, in Section 5, a scheme for explicitly modeling phoneme duration in the language modeling is described. Section 6 reports on an unsuccessful attempt to base the LID decision on the presence or absence of a few, highly discriminating phonemes. Our first foray into dialect identification is the subject of Section 7. Finally, some conclusions are drawn and plans for future work are proposed in Section 8.

*THIS WORK WAS SPONSORED BY THE DEPARTMENT OF THE AIR FORCE. THE VIEWS EXPRESSED ARE THOSE OF THE AUTHORS AND DO NOT REFLECT THE OFFICIAL POLICY OR POSITION OF THE U.S. GOVERNMENT.

Although not discussed directly in the present paper, applications for LID and DID systems fall into two main categories: pre-processing for machine understanding systems and pre-processing for human listeners. For example, an LID system might be used to select which elements of a bank of real-time, language-dependent, speech recognizers should be activated. As speech recognition systems proliferate at locations frequented by speakers of many languages (e.g. hotel lobbies, international airports), the LID system would be used as a pre-processor to determine which speech recognition models should be loaded and run. Alternatively, LID might be used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language. As reported in a study by Muthusamy [10], it can be difficult for humans to identify languages in which they are not fluent. Furthermore, Muthusamy reports anecdotely that delays on the order of several minutes can be incurred as human "front-ends" in commercial speech translation services attempt to determine the language of a speech utterance [8].

2. CORPUS

The Oregon Graduate Institute Multi-language Telephone Speech Corpus [9] has been used at a wide variety of sites to evaluate LID systems. The training segment of the corpus as used in the experiments described herein contains speech collected from at least 70 speakers for each of 11 languages. The speech, which was collected over long-distance telephone channels, comprises responses to a series of prompts with each speaker speaking for 1-2 minutes. Testing is carried out according to the U.S. National Institute of Standards and Technology (NIST) March 1994 specification:[1]

"45-s" utterance testing: LID is performed on a set of "stories" that are roughly 45 seconds in duration. These utterances are the responses to the prompt asking the speaker to speak about any topic of his choice.

"10-s" utterance testing: LID is performed on a set of 10-second cuts from the same utterances used in "45-s" testing.

3. BASELINE PRLM-P LID SYSTEM

Our baseline LID system is a parallel bank of phoneme recognizers followed by n-gram phonotactic language models (PRLM-P) [13]. Discussed below are the basic strategy of the algorithm and its performance.

3.1. Algorithm

Figure 1 shows a block diagram of the baseline PRLM-P system as it was used in the March 1994 NIST evaluation. HMM-based phoneme recognizers were trained using a phonetically labeled subset of the OGI training speech in each

[1] Contact Dr. Alvin F. Martin at NIST for details.

0-7803-2431-5/95 $4.00 © 1995 IEEE

	11L		Eng. vs. L		L vs. L'	
System	45-s	10-s	45-s	10-s	45-s	10-s
Baseline PRLM-P	79.7	69.9	96.1	94.4	94.8	91.8
PRLM-P + Gender	82.4	72.6	96.9	94.9	96.0	92.9
PRLM-P + Gender + Duration	88.8	78.9	97.5	96.0	97.9	94.9
(standard deviation)	3	2	2	1	1	1

Table 1. PRLM-P performance results using March 1994 NIST guidelines. "11L" refers to 11-alternative, forced-choice classification, "Eng. vs. L" refers to an average of the ten two-alternative, forced-choice experiments with English and one other language, and "L vs. L'" refers to an average of the 55 two-alternative, forced-choice experiments using each pair of languages. Each row of results is described in the text. The final row shows the standard deviations assuming a binomial distribution.

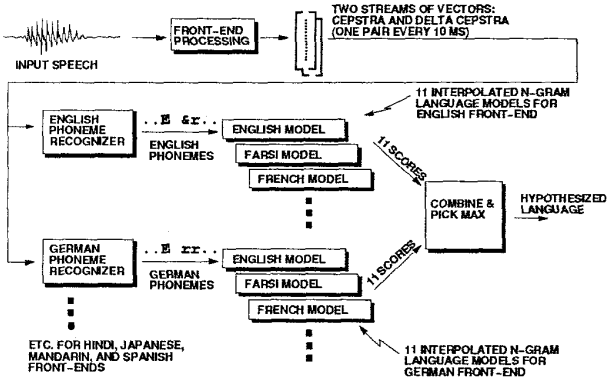

Figure 1. Baseline PRLM-P block diagram.

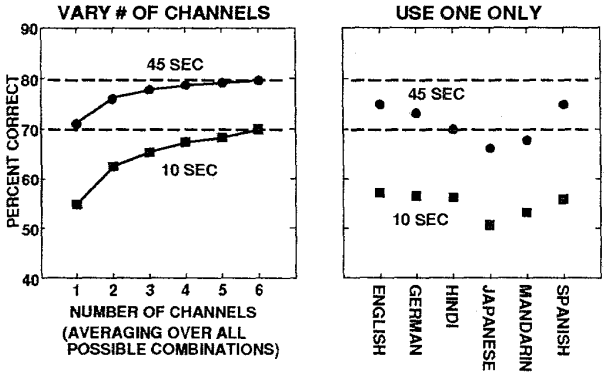

Figure 2. Using fewer than six front-ends.

of six languages: English, German, Hindi, Japanese, Mandarin, and Spanish. Each phoneme recognizer takes as input a stream of mel-weighted cepstra and delta cepstra computed from the incoming digitized speech and produces a stream of phoneme symbols as output. Interpolated n-gram language models [4] designed to capture the phonotactic statistics of each language are created by passing the training speech for each of the 11 OGI languages through each of the six front-end phoneme recognizers and recording the unigram and bigram counts. During recognition, the test utterances are passed through each of the phoneme recognizers, after which the likelihoods of the resulting phoneme sequences are calculated according to each of the language models. The final likelihood scores for each language for each utterance are calculated as the average of the individual log likelihoods emanating from the corresponding language models associated with each channel. Using multiple channels broadens our overall front-end phoneme coverage, making our composite of front-ends more language independent. It also provides multiple streams of phones that are somewhat independent of each other.

Note that though we can only build front-end phoneme recognizers in languages for which we have orthographically or phonetically transcribed speech, we can use the PRLM-P system to perform LID even on languages for which no orthographically or phonetically transcribed speech is available. Our system is different from those of Hazen [3] and Tucker [11] in that we use parallel, language-dependent front-end recognizers rather than a single front-end recognizer. We differ from Lamel [6] in that we use primarily phonotactic scores, rather than acoustic scores, for making the LID decision.

3.2. Performance

The first row of Table 1 shows the results of evaluating the PRLM-P system according to the March 1994 NIST guidelines. This was our first pass through the evaluation data, so there was no possibility of tuning the system to specific speakers or messages. For all such closed-set, forced-choice LID experiments, the Lincoln PRLM-P system was the top performing system in the official March 1994 NIST evaluation [8].

Further analysis of our March 1994 results was performed to determine the effect of reducing the number of front-end phoneme recognizers. The results on the eleven language classification task are shown in Figure 2. The left panel shows that reducing the number of channels generally reduces performance more quickly for the 10-s utterances than the 45-s utterances. The right panel shows that using only one channel, no matter which one it is, greatly reduces performance.

4. USING GENDER-DEPENDENT CHANNELS

The use of gender-dependent acoustic models is a well-known technique for improving speech recognition performance. For LID, we hoped that gender-dependent phoneme recognizers would produce a more reliable tokenization of the input speech relative to their gender-independent counterparts; therefore, n-gram analysis might prove more effective.

The general idea of employing gender-dependent channels for LID is to make a soft determination regarding the gender of the speaker of a message and then to use the confidence of that determination to weight the phonotactic evidence from gender-dependent channels. A block diagram is shown in Figure 3. During training, three phoneme recognizers per front-end language are trained: one from male speech, one from female speech, and one from combined male and female speech. Next, for each language to be identified, three interpolated n-gram language models are trained, one for each of the front-ends. The language models associated with the male phoneme recognizer are trained only on male messages, the female language models only on

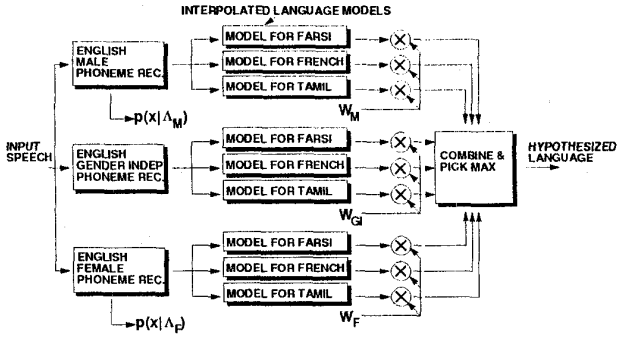

Figure 3. Example of gender-dependent processing for a channel with an English front-end. The acoustic likelihoods $p(x|\Lambda_M)$ and $p(x|\Lambda_F)$ are used to compute the weights W_M, W_F, and W_{GI}.

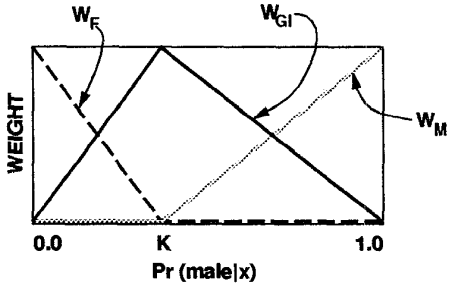

Figure 4. Weight functions.

female messages, and the combined models on both male and female messages.

During recognition, an unknown message x is processed by all three front-ends. The acoustic likelihood scores emanating from the male front-end and from the female front-end are used to compute the a posteriori probability that the message is male as

$$Pr(\text{male}|x) = \frac{p(x|\Lambda_M)}{p(x|\Lambda_M) + p(x|\Lambda_F)} \quad (1)$$

where $p(x|\Lambda_M)$ is the likelihood of the best state sequence given the male HMMs, Λ_M, and $p(x|\Lambda_F)$ is the likelihood of the best state sequence given the female HMMs, Λ_F. Observing empirically that the cutoff between male and female messages is not absolutely distinct and does not always occur exactly at $Pr(\text{male}|x) = 0.5$, $Pr(\text{male}|x)$ is used to calculate three weights:

$$W_M = \begin{cases} \frac{Pr(\text{male}|x)-K}{1-K} & \text{if } Pr(\text{male}|x) \geq K \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$W_F = \begin{cases} \frac{K-Pr(\text{male}|x)}{K} & \text{if } Pr(\text{male}|x) < K \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$W_{GI} = \begin{cases} 1-W_M & \text{if } Pr(\text{male}|x) \geq K \\ 1-W_F & \text{if } Pr(\text{male}|x) < K \end{cases} \quad (4)$$

where W_M is the weight for the male channel, W_F is the weight for the female channel, W_{GI} is the weight for the gender-independent channel, and K is a constant set empirically during training (typically ranging from 0.30 to 0.70). The weight functions are shown graphically in Figure 4. The W's are used to weight the phonotactic language model

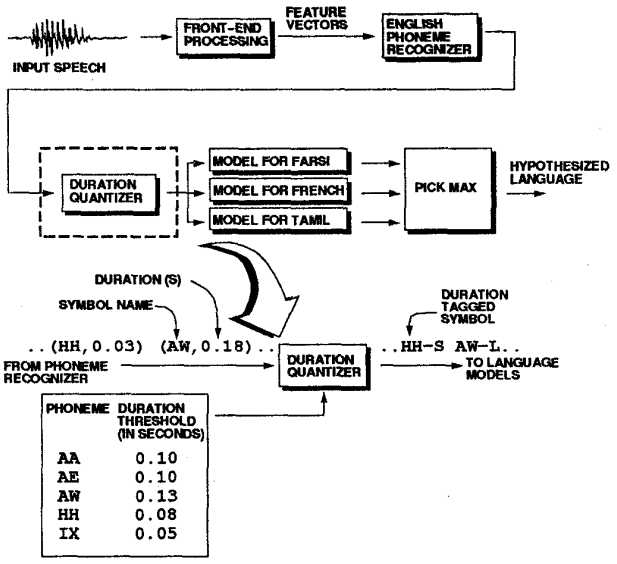

Figure 5. Approach to duration tagging (as suggested by Mistretta).

scores as follows:

$$p(x|l) = W_M p(x|\lambda_l^M) + W_F p(x|\lambda_l^F) + W_{GI} p(x|\lambda_l^{GI}), \quad (5)$$

where λ_l^M is the interpolated n-gram language model trained by passing male language l speech through the male phoneme recognizer, λ_l^F is the interpolated n-gram language model trained by passing female language l speech through the female phoneme recognizer, and λ_l^{GI} is the interpolated n-gram language model trained by passing both male and female language l speech through the gender-independent phoneme recognizer.[2]

The second row of results in Table 1 shows the performance of a PRLM-P system with gender dependent channels. This new system has 16 channels: three each for English, German, Japanese, Mandarin, and Spanish, and one for Hindi, as there was insufficient female speech to train gender-dependent front-ends for Hindi. As shown in the table, use of gender-dependent front-ends together with gender-independent front-ends results in an improvement in LID performance.

5. DURATION MODELING

On advice from Bill Mistretta of Lockheed-Sanders [7], we have begun to use phoneme duration information output from the front-end phoneme recognizers explicitly in the language ID process. Our version of the Mistretta approach for using duration information is shown in Figure 5. The training data for all languages are passed through each of the front-end phoneme recognizers. A histogram of durations for each phoneme emitted from each recognizer is compiled and the average duration determined. A -L suffix is appended to all phonemes having duration longer than the average duration for that phoneme, and a -S suffix is appended to all phonemes having duration shorter than the average duration. This modified sequence of phoneme symbols is then used in place of the original sequence for training the interpolated language models. During recognition,

[2] We could certainly use a simpler algorithm for making the gender ID decision, but the phoneme recognizer acoustic likelihoods are already being calculated as part of the phone recognition process; hence, we get them for free in our system.

the same procedure is applied to the output symbols from the phoneme recognizer using the means determined during training as thresholds. As shown in the third row of Table 1, this simple technique for modeling duration combined with the use of gender-dependent front-ends further improves LID performance.

6. KEYPHONES

It has been suggested (e.g. [2], [1]) that performance of LID systems might be improved by focusing on the "best" sounds, where a sound is "good" for the purposes of LID if it can both be reliably identified and its expected rate of occurrence is different among the languages to be identified. Presumably, the "bad" sounds, i.e. those that either cannot be identified reliably or that do not occur with different rates in different languages, only confuse the LID process. In the context of our PRLM system, we call such "good" sounds "keyphones."[3] Keyphones are those phonemes whose statistics are found to be most dissimilar from one language to the next, as measured by passing training data in each language through a front-end phoneme recognizer. Specifically, we use a keyphone goodness measure related to the difference in the phoneme's measured probability of occurrence from language to language normalized by the standard deviation of those measurements.

Using a PRLM-P system with a single English front-end, Vietnamese vs. Korean LID was performed. These two languages were chosen because they form one of the most difficult language pairs. Results on 10-s test utterances showed that filtering out all but the five best keyphones resulted in LID performance equivalent to using all 42 phonemes. Similar results were obtained for English vs. German LID (another difficult pair). Though it was possible to reduce by an order of magnitude the number of symbols scored by the language model without eroding performance, we were never able to improve performance by ignoring poorly discriminating phonemes.

7. DIALECT IDENTIFICATION

More recently, attention has focused on applying LID techniques to the problem of dialect identification. Using the "dialect region" labels of the TIMIT database, we attempted to train an LID system to recognize the difference between "New England" and "Southern" American English. As all TIMIT speech is phonetically labeled, it seemed most appropriate to to use a parallel phoneme recognition (PPR) LID system as first proposed for LID by Lamel [5, 13]. In this approach, a phoneme recognizer is created for each class (i.e. each dialect region in this case) during training. During recognition, the acoustic likelihoods along the most likely phoneme paths are calculated and compared. Using such a system, we were able to classify correctly the dialect 71% of the time using test utterances that were eight sentences in duration. It is important to note that these tests on read sentences really demonstrate accent ID rather than dialect ID, because the speakers were not free to choose the words they spoke. As such, the results may understate the potential performance of PPR (or other LID system) on the dialect ID problem.

8. CONCLUSIONS

This paper has reported on the progress that has been made during the past year at improving the Lincoln LID system. Starting with a system that demonstrated top performance at a government-sponsored evaluation, a number of enhancements have been made, including the use of gender-dependent channels and explicit duration modeling. Basing the LID decision only on the statistics of the phonemes best able to discriminate between languages did not improve performance. Finally, we ran a simple dialect identification experiment using a system that had been developed for LID.

We still await the availability of larger, standardized, multi-language speech corpora. Our hope is that these new corpora will allow us to train and test systems that model language dependencies more accurately than is possible with our current phoneme recognizers and interpolated n-gram language models.

[3] Others have called such units "monophones," meaning a phone that is specific to one language vs. "polyphones," meaning a phone that is present in many languages. The term "monophone," however, is already widely associated with context-independent subword units.

9. ACKNOWLEDGEMENTS

The author wishes to thank Bill Mistretta and Dave Morgan of Lockheed Sanders for introducing him to their method of duration modeling described in Section 5.

REFERENCES

[1] O. Andersen, P. Dalsgaard, and W. Barry. On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four European languages. In *ICASSP '94 Proceedings*, volume 1, pages 121–124, April 1994.

[2] K. M. Berkling, T. Arai, and E. Barnard. Analysis of phoneme-based features for language identification. In *ICASSP '94 Proceedings*, volume 1, pages 289–292, April 1994.

[3] T. J. Hazen and V. W. Zue. Recent improvements in an approach to segment-based automatic language identification. In *ICSLP '94 Proceedings*, 1994.

[4] F. Jelinek. Self-organized language modeling for speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in speech recognition*, pages 450–506. Morgan Kaufmann, Palo Alto, CA, 1990.

[5] L. F. Lamel and J.-L. Gauvain. Cross-lingual experiments with phone recognition. In *ICASSP '93 Proceedings*, volume 2, pages 507–510, April 1993.

[6] L. F. Lamel and J. L. Gauvain. Language identification using phone-based acoustic likelihoods. In *ICASSP '94 Proceedings*, volume 1, pages 293–296, April 1994.

[7] W. Mistretta. Private Communication.

[8] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, October 1994.

[9] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *ICSLP '92 Proceedings*, volume 2, pages 895–898, October 1992.

[10] Y. K. Muthusamy, N. Jain, and R. A. Cole. Perceptual benchmarks for automatic language identification. In *ICASSP '94 Proceedings*, volume 1, pages 333–336, April 1994.

[11] R. C. F. Tucker, M. J. Carey, and E. S. Parris. Automatic language identification using sub-word models. In *ICASSP '94 Proceedings*, volume 1, pages 301–304, April 1994.

[12] M. A. Zissman. Automatic language identification using Gaussian mixture and hidden Markov models. In *ICASSP '93 Proceedings*, volume 2, pages 399–402, April 1993.

[13] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *ICASSP '94 Proceedings*, volume 1, pages 305–308, April 1994.