AUTOMATIC LANGUAGE IDENTIFICATION OF TELEPHONE SPEECH MESSAGES USING PHONEME RECOGNITION AND N-GRAM MODELING*

Marc A. Zissman and Elliot Singer
Lincoln Laboratory, Massachusetts Institute of Technology
244 Wood Street
Lexington, MA 02173–9108
USA
(617) 981-2547
maz@sst.ll.mit.edu , es@sst.ll.mit.edu

ABSTRACT

This paper compares the performance of four approaches to automatic language identification (LID) of telephone speech messages: Gaussian mixture model classification (GMM), language-independent phoneme recognition followed by language-dependent language modeling (PRLM), parallel PRLM (PRLM-P), and language-dependent parallel phoneme recognition (PPR). These approaches span a wide range of training requirements and levels of recognition complexity. All approaches were tested on the development test subset of the OGI multi-language telephone speech corpus. Generally, system performance was directly related to system complexity, with PRLM-P and PPR performing best. On 45 second test utterances, average two language, closed-set, forced-choice classification performance reached 94.5% correct. The best 10 language, closed-set, forced-choice performance was 79.2% correct.

1. INTRODUCTION

This paper compares the performance of four approaches to automatic language identification (LID) of telephone speech messages: Gaussian mixture model classification (GMM), language-independent phoneme recognition followed by language-dependent language modeling (PRLM), parallel PRLM (PRLM-P), and language-dependent parallel phoneme recognition (PPR). Generally, we were interested in studying performance as a function of required training data and system complexity. The rest of this paper is organized as follows: Section 2 describes each LID approach in detail, Section 3 reviews the organization of the OGI multi-language speech corpus, Section 4 reports LID results, and Section 5 discusses the implications of this work and suggests future research directions.

2. ALGORITHMS

This section describes the four LID algorithms.

2.1. Gaussian Mixture Model (GMM) Classification

A GMM LID system, having been studied previously both by us and other sites [1, 2], served as a baseline for this study. A multi-variate GMM density, $p(\vec{x}|\lambda)$, is a weighted sum of uni-modal multi-variate Gaussian densities, i.e.

$$p(\vec{x}|\lambda) = \sum_{k=1}^{N} p_k b_k(\vec{x}), \qquad (1)$$

where λ is the set of model parameters

$$\lambda = \{p_k, \vec{\mu_k}, \Sigma_k\}, \qquad (2)$$

k is the mixture index ($1 \leq k \leq N$), the p_k's are the mixture weights, and the b_k's are the multi-variate Gaussian densities defined by the means $\vec{\mu_k}$ and variances Σ_k. Two GMMs are created for each language l, one for cepstral feature vectors and one for first-order difference ("delta") cepstral feature vectors, as follows:

- From training speech spoken in language l, two independent feature vector streams are extracted: centisecond mel-scale cepstra ($c_1 - c_{12}$) and delta cepstra ($\Delta c_0 - \Delta c_{12}$). An adaptive threshold energy-based silence detector is used to remove silence. RASTA is applied to help remove effects of the telephone channel [3].
- Each stream of feature vectors is clustered, using K-means, producing N cluster centers for each stream. $N = 40$ is used in this study.
- Using the cluster centers as initial estimates for the Gaussian mixture centers, multiple iterations of the estimate-maximize (E-M) algorithm are run, producing, for each stream, a more likely set of mean vectors, diagonal covariance matrices, and mixture weights [4].

During recognition, an unknown speech utterance is classified by first converting the PCM waveform to feature vectors, including removing silence and applying RASTA, and then calculating the log likelihood that the language l model produced the unknown speech utterance, where the log likelihood, $\mathcal{L}$, is defined as

$$\mathcal{L}(\{\vec{x_t}, \vec{y_t}\}|\lambda_l^C, \lambda_l^{DC}) = \sum_{t=1}^{T} \left[\log p(\vec{x_t}|\lambda_l^C) + \log p(\vec{y_t}|\lambda_l^{DC})\right], \qquad (3)$$

where λ_l^C is the ceptral GMM for language l, λ_l^{DC} is the delta cepstral GMM for language l, $\vec{x_t}$ is the cepstral feature vector at time t, and $\vec{y_t}$ is the delta cepstral feature vector at time t. Implicit in this equation are the assumptions that the observations $\{\vec{x_t}\}$ are statistically independent of each other, the observations $\{\vec{y_t}\}$ are statistically independent of each other, and the two streams are jointly statistically independent of each other. The maximum likelihood classifier hypothesizes $\hat{l}$ as the language of the unknown utterance, where

$$\hat{l} = \arg\max_{l} \mathcal{L}\left(\{\vec{x_t}, \vec{y_t}\}|\lambda_l^C, \lambda_l^{DC}\right) \qquad (4)$$

GMM is very simple to train, as it requires neither an orthographic nor phonetic labeling of the training speech. GMM maximum likelihood recognition is also very simple, e.g. a C implementation of a two language classifier can be run easily in real-time on a Sun SPARCstation-10.

*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.

0-7803-1775-0/94 $3.00 © 1994 IEEE

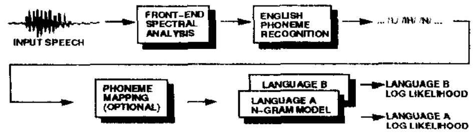
INPUT SPEECH
FRONT-END SPECTRAL ANALYSIS
ENGLISH PHONEME RECOGNITION
... /U/ /IH/ /N/ ...
PHONEME MAPPING (OPTIONAL)
LANGUAGE B
LANGUAGE A N-GRAM MODEL
LANGUAGE B LOG LIKELIHOOD
LANGUAGE A LOG LIKELIHOOD

Figure 1. PRLM Block Diagram

2.2. Phoneme Recognition followed by Language Modeling (PRLM)

The second language ID approach is novel, though a similar strategy was developed independently at another site [5]. This new technique, which employs an English phoneme recognizer followed by an N-gram analyzer as shown in Figure 1, is motivated by a desire to model speech sequence information and is a compromise between:

- Modeling the sequence information using hidden Markov models (HMMs) trained from unlabeled speech. Such systems have been shown to perform no better than static classification [1, 2].
- Employing language-dependent phoneme recognizers trained from labeled speech.[1] Such a system is the subject of Section 2.4.

2.2.1. *English Phoneme Recognition*

For PRLM, an English phoneme recognizer is trained either on the entire[2] training set of the NTIMIT telephone speech corpus (3.1 hours of read, labeled, telephone speech recorded using single microphone) [6] or on CREDITCARD excerpts from the SWITCHBOARD corpus (3.8 hours of spontaneous, labeled, telephone speech recorded using many microphones) [7]. The phoneme recognizer, implemented using the commercially available Hidden Markov Model Toolkit (HTK) available from Entropic Research Laboratory, is a network of English monophones (48 for NTIMIT, 42 for SWITCHBOARD), where each phoneme model contains three emitting states.

The observation streams are the same mel-weighted, RASTA processed, silence removed, centisecond cepstra and delta-cepstra vectors used in the GMM system. The probability density for each state for each stream is modeled as a 6-component Gaussian mixture density. Models are trained using the forward-backward algorithm. Recognition is performed via a Viterbi search, using a fully connected null-grammar network of monophones. Phoneme recognition, which dominates the PRLM processing time, takes about 1.5x real-time on a Sun SPARCstation-10 (i.e. a 10 second utterance takes about 15 seconds to process).

2.2.2. *Language Model*

With an English phoneme recognizer in hand, a language model can be trained for each language l by running training speech for language l into the phoneme recognizer and computing a set of n-gram histograms. An interpolated n-gram language model [8] is used to approximate the probability of an n-gram as the weighted sum of the probabilities of the n-gram, the (n-1)-gram, etc. An example for a bigram model is

$$\hat{P}(w_t|w_{t-1}) = \alpha_2 P(w_t|w_{t-1}) + \alpha_1 P(w_t) + \alpha_0 P_0. \qquad (5)$$

where w_{t-1} and w_t are consecutive symbols observed in the phoneme stream. The P's are ratios of counts observed in the training data, e.g. $P(w_t|w_{t-1}) = C(w_{t-1}, w_t)/C(w_{t-1})$. The α's can be estimated iteratively using the E-M algorithm so as to minimize perplexity, or they can be set by hand. During recognition, the test utterances are first passed through the English phoneme recognizer, producing a phoneme sequence, $W = \{w_1, w_2, ...\}$. The log likelihood, $\mathcal{L}$, that the interpolated bigram language model for language l, λ_l^{BG}, produced the phoneme sequence W, is

$$\mathcal{L}(W|\lambda_l^{BG}) = \sum_{t=1}^{T} \log \hat{P}(w_t|w_{t-1}, \lambda_l^{BG}) \qquad (6)$$

The maximum likelihood classifier decision rule is used, which hypothesizes that $\hat{l}$ is the language of the unknown utterance, where

$$\hat{l} = \arg\max_{l} \mathcal{L}(W|\lambda_l^{BG}). \qquad (7)$$

2.2.3. *Phoneme Mapping: Fineness vs. Accuracy*

It was speculated that mapping the phones from fine classes to a smaller number of broader classes might improve performance. While the fine phone information would be lost, it was thought that the higher accuracy obtained with broad class vs. fine class phoneme recognition might more than compensate for the loss of information. Using the NTIMIT trained recognizer, two different mappings of the 48 fine classes were investigated: a five class mapping (vowel, sonorant, fricative, stop, silence), and a 12 class mapping (high/low vowels, diphthongs, liquids, semi-vowels, nasals, closures, voiced/unvoiced stops, voiced/unvoiced fricatives).

2.3. Parallel PRLM (PRLM-P)

If labeled training speech is available for more than one language, *but not necessarily for any of the languages to be recognized,* one can create the parallel PRLM system shown in Figure 2. Considering the figure, one might use labeled English speech to train the front-end for one PRLM system (L=English) and use labeled French speech to train a front-end for a separate, PRLM system (M=French). The two PRLM systems could be trained and then tested in parallel, say on Spanish vs. Japanese (A=Spanish, B=Japanese) with the final log likelihood scores for each language A and B calculated as the sum of the individual scores from the L-based and M-based PRLM systems. Note that this approach extends easily to any number of parallel PRLM systems. One is only limited by the number of languages for which labeled training speech is available.

2.4. Parallel Phoneme Recognition (PPR)

Parallel phoneme recognition has been shown to be an effective LID technique both for high quality speech [9] and telephone speech [10]. PPR systems require labeled speech *for every language to be recognized*; therefore, it is more difficult to implement a PPR system than any of the other systems already discussed. A block diagram of the PPR system studied is shown in Figure 3.

The language-dependent phoneme recognizers in the PPR language ID system have the same configuration as the English recognizer used in PRLM with a few exceptions. First, the interpolated language model is an integral part of the recognizer in the PPR system, whereas it is a post processor in the PRLM system. During recognition, the inter-phoneme transition probability between two phoneme models i and j is

$$a_{ij} = s \log \hat{P}(j|i) \qquad (8)$$

where s is the grammar scale factor, and the $\hat{P}$'s are interpolated bigram probabilities derived from the training

[1] In this paper, "labeled" speech means speech waveforms with either (1) time-aligned, phoneme labels or (2) orthographic transcriptions plus a pronunciation dictionary that maps words to phoneme sequences.

[2] Except for the shibboleth sentences.

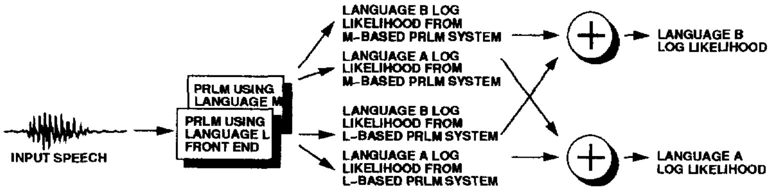
INPUT SPEECH
PRLM USING LANGUAGE M
PRLM USING LANGUAGE L FRONT END
LANGUAGE B LOG LIKELIHOOD FROM M-BASED PRLM SYSTEM
LANGUAGE A LOG LIKELIHOOD FROM M-BASED PRLM SYSTEM
LANGUAGE B LOG LIKELIHOOD FROM L-BASED PRLM SYSTEM
LANGUAGE A LOG LIKELIHOOD FROM L-BASED PRLM SYSTEM
+
+
LANGUAGE B LOG LIKELIHOOD
LANGUAGE A LOG LIKELIHOOD

Figure 2. PRLM-P Block Diagram

	English/Japanese		English/Spanish		Japanese/Spanish		Average	
	Whole	10-sec	Whole	10-sec	Whole	10-sec	Whole	10-sec
GMM	82.9	84.2	82.9	83.7	64.7	64.0	76.8	77.3
PRLM (NTIMIT)	88.6	88.3	88.6	82.9	91.2	75.7	89.4	82.3
PRLM (SWITCHBOARD)	94.3	87.5	97.1	84.6	88.2	78.4	93.2	83.5
PRLM-P	91.4	90.0	97.1	88.0	94.1	90.1	94.2	89.4
PPR	94.3	92.5	97.1	92.3	85.3	87.4	92.2	90.7

Table 1. English/Japanese/Spanish Language Pair Results (% correct)

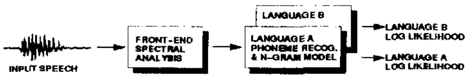
INPUT SPEECH
FRONT-END SPECTRAL ANALYSIS
LANGUAGE B
LANGUAGE A PHONEME RECOG. & N-GRAM MODEL
LANGUAGE B LOG LIKELIHOOD
LANGUAGE A LOG LIKELIHOOD

Figure 3. PPR Block Diagram

data. Based on prelininary testing, $s = 3$ was used in these experiments. Another difference between our PRLM and PPR phoneme recognizers is that whereas the PRLM recognizer uses only monophones, the PPR recognizers use the monophones of each language plus the 100 most commonly occurring right context-dependent diphones.

PPR LID is performed by Viterbi decoding the test utterance once for each language-dependent phoneme recognizer. Each phoneme recognizer finds the most likely path of the test utterance through the recognizer and calculates the log likelihood score (normalized by length) for that best path. A modified maximum likelihood criterion is used to hypothesize the most likely language, in which a recognizer dependent bias is subtracted from each log likelihood score prior to applying the maximum likelihood decision rule. The recognizer dependent bias is set to the average of the log likelihoods for all messages processed by the recognizer. The PPR recognizer for each language runs at about 2x real-time on a Sun SPARCstation-10.

3. CORPUS

The Oregon Graduate Institute Telephone Speech Corpus (OGI-TS) was used to evaluate the performance of each of the LID approaches outlined above [11]. This corpus contains 50 training messages, 20 development test messages, and 20 evaluation test messages for each of 10 languages. Each message was spoken by a unique speaker over a telephone channel and comprises responses to ten prompts. Messages are typically about 1-2 minutes in duration.

The GMMs are trained on the training speaker responses to each of the six free-text prompts. As the PPR models require labeled training data, PPR training is limited to that part of the OGI corpus that is labeled — currently the 45 second long "story-before-the-tone" (story-bt) utterances. Because the forward-backward algorithm would likely have trouble aligning phoneme models against 45 second long utterances, shorter, hand endpointed segments of the story-bt utterances are used. This also results in less heavy reliance on the OGI supplied phoneme start and end times. When PRLM and PPR systems are compared, PRLM models are trained on the same short utterances as the PPR models; otherwise, PRLM models are trained on the story-bt utterances in their entirety. Testing is carried out according to the NIST April 1993 specification:

"whole" utterance testing: LID is performed on a set of 45-second story-bt utterances spoken by the development test speakers.

"10-sec" utterance testing: LID is performed on a set of 10-second cuts from the same story-bt utterances used in "whole" testing.

Though phonetic labeling of at least five OGI-TS languages is underway at OGI, only English, Japanese, and Spanish had been labeled sufficiently at the time of our experiments. As the PPR system requires these labels for training, experiments comparing GMM, PRLM, PRLM-P and PPR are limited to these three languages. Additional experiments comparing GMM, PRLM and PRLM-P use all ten languages. All PRLM-P experiments use parallel English, Japanese, and Spanish PRLM systems whose front-ends are trained on the OGI training data.

4. RESULTS

In this section, results of LID performance are reported, followed by some phoneme recognition results for recognizers used in the PPR system.

4.1. LID Results

Results of the English/Japanese/Spanish experiments are shown in Table 1. Averages are computed with equal-weighting per language pair. Standard deviations on the averages are approximately 4% (104 trials) and 2% (348 trials) for the whole and 10-sec utterances, respectively, assuming binomial distributions. Generally, the results show that PRLM-P and PPR perform about equally, which is not surprising as the only difference between the two systems for these three languages is the manner in which the language model is applied. For the whole messages, SWITCHBOARD-based PRLM performs about as well as PRLM-P and PPR. The results shown in Table 1 are obtained when the α's for the PRLM and PPR interpolated language models are set by hand. For PRLM, $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$, whereas for PPR, $\alpha_1 = 0.0$ and $\alpha_2 = 1.0$. Setting the α's this way results in slightly better performance than setting them automatically using the E-M algorithm.

Some additional experiments were run comparing PRLM, PRLM-P and GMM using all 10 languages of the OGI-TS

	10L		ENG vs. L		L vs. L'	
	Whole	10-sec	Whole	10-sec	Whole	10-sec
GMM	53.4	49.7	81.3	83.5	80.0	79.4
PRLM (NTIMIT)	67.4	47.3	88.3	81.7	90.0	83.8
PRLM (SWITCHBOARD)	71.9	53.7	94.7	88.0	92.0	85.8
PRLM-P	79.2	63.0	92.0	88.5	94.5	89.9
σ	3	2	2	1	1	1

Table 2. Full 10 Language Results (% correct)

	Accuracy %	Number of monophones	# phone classes
English	41.9	51	39
Japanese	55.5	27	25
Spanish	54.9	38	34

Table 3. PPR Phoneme Recognition Results

corpus. The first two columns of Table 2 show 10 language, closed-set, forced-choice results. Next, two language, closed-set, forced-choice average results for English vs. each of the other nine languages are presented. The final two columns show two language, closed-set, forced-choice results averaged over all of the 45 language pairs. Approximate standard deviations (σ) are shown in the bottom row. Generally, PRLM-P performs best.

4.2. Phoneme Recognition Results

Within language phoneme recognition performance of the PPR recognizers is shown in Table 3. The results are presented in terms of accuracy, thereby accounting for substitution, deletion, and insertion errors. Note that for each language, the number of equivalence classes is less than the number of monophones. The 10-sec utterances were used to evaluate phoneme recognition performance.

4.3. Other Results

Although phoneme class mapping results are not shown in the tables. PRLM was found to perform best with the finest phoneme classes and worst with the broadest classes, which is consistent with previously reported results [10]. Additionally, single PRLM systems using OGI trained phone recognizers in each of the three languages perform comparably to the SWITCHBOARD-based PRLM system. Combining the PRLM outputs to form the PRLM-P system improves performance (as shown in the tables).

5. DISCUSSION

This paper has compared the performance of four approaches to automatic language identification (LID) of telephone speech messages. As labeling foreign language speech can be difficult, it is encouraging that best performance is obtained with PRLM-P, a system that can use, but does not require, labeled speech for each language to be recognized. New PRLM-like systems employing a single phoneme recognizer trained from multi-language speech are being explored.

As automatic speech recognition systems become available for more and more languages, it is reasonable to believe that the availability of standardized, multi-language speech corpora will increase. These large new corpora should allow us to train and test systems that model language dependencies more accurately than is possible with just language-dependent phoneme recognizers employing bigram grammars. Perhaps most obviously, larger corpora would allow training of separate male and female models. LID systems that use language-dependent word spotters and/or continuous speech recognizers may achieve even better performance. We plan to investigate such approaches as larger corpora become available.

6. ACKNOWLEDGEMENTS

The authors are grateful to the Center for Spoken Language Understanding at the Oregon Graduate Institute for making available the OGI-TS speech waveforms and the associated phonetic labels and to BBN for making available their high quality transcriptions of the CREDITCARD excerpts. Tina Chou and Linda Sue Sohn helped prepare the data at Lincoln for PPR processing, and many other members of the Lincoln speech group offered suggestions and advice.

REFERENCES

[1] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *ICSLP '92 Proceedings*, volume 2, pages 1011–1014, October 1992.

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

[3] H. Hermansky et al. RASTA-PLP speech analysis technique. In *ICASSP '92 Proceedings*, volume 1, pages 121–124, March 1992.

[4] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[5] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Proceedings of Eurospeech 93*, volume 2, pages 1303–1306, September 1993.

[6] C. R. Jankowski et al. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In *ICASSP '90 Proceedings*, April 1990.

[7] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *ICASSP '92 Proceedings*, volume 1, pages 517–520, March 1992.

[8] 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.

[9] L. F. Lamel and J.-L. Gauvain. Identifying non-linguistic speech features. In *Proceedings of Eurospeech 93*, volume 1, pages 23–30, September 1993.

[10] Y. Muthusamy et al. A comparison of approaches to automatic language identification using telephone speech. In *Proceedings of Eurospeech 93*, volume 2, pages 1307–1310, September 1993.

[11] 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.