Comparison of Four Approaches to Automatic Language Identification of Telephone Speech

Marc A. Zissman, *Member, IEEE*

Abstract— **We have compared the performance of four approaches for automatic language identification of speech utterances: Gaussian mixture model (GMM) classification; single-language phone recognition followed by language-dependent, interpolated n-gram language modeling (PRLM); parallel PRLM, which uses multiple single-language phone recognizers, each trained in a different language; and language-dependent parallel phone recognition (PPR). These approaches, which span a wide range of training requirements and levels of recognition complexity, were evaluated with the Oregon Graduate Institute Multi-Language Telephone Speech Corpus. Systems containing phone recognizers performed better than the simpler GMM classifier. The top-performing system was parallel PRLM, which exhibited an error rate of 2% for 45-s utterances and 5% for 10-s utterances in two-language, closed-set, forced-choice classification. The error rate for 11-language, closed-set, forced-choice classification was 11% for 45-s utterances and 21% for 10-s utterances.**

I. Introduction

OVER the past three decades, significant effort has been focused on the automatic extraction of information from speech signals. Many techniques reported previously in this journal and its predecessors have been aimed at obtaining either a transcription of the speech signal or an identification of the speaker's identity and gender. For the most part, the task of determining the language in which the speech was spoken has received far less attention. It is the purpose of this paper to report on the research, development, and evaluation of automatic language-identification systems at MIT Lincoln Laboratory. Where possible, comparisons and contrasts to language-ID systems studied at other sites will be drawn.

Language-ID applications fall into two main categories: pre-processing for machine understanding systems and pre-processing for human listeners. As suggested by Hazen and Zue, consider the hotel lobby or international airport of the future, in which one might find a multi-lingual voice-controlled travel information retrieval system [1]. If the system has no mode of input other than speech, then it must be capable of determining the language of the speech commands either *while* it is recognizing the commands or *before* recognizing the commands. To determine the language during recognition would require running many speech recognizers in parallel, one for each language. As one might wish to support tens or even hundreds of input languages, the cost of the required real-time hardware might prove prohibitive. Alternatively, a language-ID system could be run in advance of the speech recognizer. In this case, the language-ID system would quickly output a list containing the most likely languages of the speech commands, after which the few, most appropriate, language-dependent speech recognition models could be loaded and run on the available hardware. A final language-ID determination would only be made once speech recognition was complete.

Alternatively, language ID might be used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language. Such scenarios are already occurring today: for example, AT&T offers the *Language Line* interpreter service to, among others, police departments handling emergency calls. When a caller to *Language Line* does not speak any English, a human operator must attempt to route the call to an appropriate interpreter. Much of the process is trial and error (for example, recordings of greetings in various languages may be used) and can require connections to several human interpreters before the appropriate person is found. As recently reported by Muthusamy [2], when callers to *Language Line* do not speak any English, the delay in finding a suitable interpreter can be on the order of minutes, which could prove devastating in an emergency situation. Thus, a language-ID system that could quickly determine the most likely languages of the incoming speech might cut the time required to find an appropriate interpreter by one or two orders of magnitude.

Although research and development of automatic language-identification systems has been in progress for the past twenty years, publications have been sparse. Therefore, Section II begins with a brief discussion of previous work. The background discussion does not provide a quantitative report on the performance of each of these systems as, until recently, a standard multi-language evaluation corpus that could allow a fair comparison among the systems did not exist. Section III remarks on some cues that humans and machines use for identifying languages. By reviewing some of the key elements that distinguish one language from another, it serves to motivate the development of specific automatic algorithms. Section IV describes each of the four language-ID approaches that were the main focus of this work: Gaussian mixture modeling (GMM) [3]–[5], single-language phone recognition followed by language-dependent language modeling (PRLM) [6]–[8], parallel PRLM [7], and language-dependent parallel phone recognition (PPR) [9], [10]. Because these approaches

Manuscript received January 23, 1995; revised September 5, 1995. This work was sponsored by the Department of the Air Force. The views expressed are those of the author and do not reflect the official policy or position of the U.S. Government. The associate editor coordinating the review of this paper and approving it for publication was Dr. Douglas D. O'Shaughnessy.

The author is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173 USA.

Publisher Item Identifier S 1063-6676(96)01333-8.

have differing levels of computational complexity and training data requirements, our goal was to study performance while considering the ease with which the systems may be trained and run. Section V reviews the organization of the Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [11], which has become a standard corpus for evaluating language-ID systems. We used the OGI-TS corpus to evaluate our four systems. At the start of our work, the corpus comprised speech from approximately 90 speakers in each of ten languages, though both the numbers of speakers and languages have grown with time. Section VI reports language-ID performance of the four systems we tested on the OGI-TS corpus, and Section VII details results of some additional work that sought to improve the best system. Finally, Section VIII discusses the implications of this work and suggests future research directions.

II. Background

Research in automatic language identification from speech has a history extending back at least twenty years. Until recently, it was difficult to compare the performance of these systems, as few of the algorithms had been evaluated on common corpora. Thus, what follows is a brief description of some representative systems without much indication of quantitative performance. The reader is also referred to Muthusamy's recent review of language-ID systems [2].

Most language-ID systems operate in two phases: training and recognition. During the training phase, the typical system is presented with examples of speech from a variety of languages. Some systems require only the digitized speech utterances and the corresponding true identities of the languages being spoken. More complicated language-ID systems may require either:

- a phonetic transcription (sequence of symbols representing the sounds spoken), or
- an orthographic transcription (the text of the words spoken) along with a pronunciation dictionary (mapping of each word to a prototypical pronunciation)

for each training utterance. Producing these transcriptions and dictionaries is an expensive and time-consuming process that usually requires a skilled linguist fluent in the language of interest. For each language, the training speech is analyzed and one or more models are produced. These models are intended to represent some set of language-dependent, fundamental characteristics of the training speech that can then be used during the second phase of language ID: recognition. During recognition, a new utterance is compared to each of the language-dependent models. In most systems, the likelihood that the new utterance was spoken in the same language as the speech used to train each model is computed, and the maximum-likelihood model is found. The language of the speech that was used to train the maximum-likelihood model is hypothesized as the language of the utterance.

The earliest automatic language-ID systems used the following procedure: examine training speech (either manually or automatically), extract and store a set of prototypical spectra (each computed from about 10 ms of the training speech) for each language, analyze and compare test speech to the sets of prototypical spectra, and classify the test speech based on the results of the comparison. For example, in systems proposed by Leonard and Doddington [12]–[15], spectral feature vectors extracted from training messages were scanned by the researchers for regions of stability and regions of very rapid change. Such regions, thought to be indicative of a specific language, were used as exemplars for template matching on the test data. After this initial work, researchers have tended to focus on automatic spectral feature extraction, unsupervised training, and maximum-likelihood recognition. Cimarusti [16] ran a polynomial classifier on 100-element LPC-derived feature vectors. Foil [17] examined both formant and prosodic feature vectors, finding that formant features were generally superior. His formant-vector-based language-ID system used k-means training and vector quantization classification. Goodman [18] extended Foil's work by refining the formant feature vector and classification distance metric. Ives [19] constructed a rule-based language-ID system. Classification was performed using thresholds on pitch and formant frequency variance, power density centroids, etc. Sugiyama [20] performed vector quantization classification on LPC features. He explored the difference between using one VQ codebook per language versus one common VQ codebook. In the latter case, languages were classified according to their VQ histogram patterns. Riek [3], Nakagawa [4], and Zissman [5] applied Gaussian mixture classifiers to language identification. Gaussian mixture classification is, in some sense, a generalization of exemplar extraction and matching, and is described more fully in Section IV.

In an effort to move beyond low-level spectral analysis, Muthusamy [21] built a neural-net-based, multi-language segmentation system capable of partitioning a speech signal into sequences of seven broad phonetic categories. For each utterance, the class sequences were converted to 194 features used to identify language.

Whereas the language-identification systems described above perform primarily static classification, in that the feature vectors are assumed to be independent of each other and no use of feature vector sequences is made, other systems have used hidden Markov models (HMMs) to model sequential characteristics of speech production. HMM-based language identification was first proposed by House and Neuburg [22]. They created a discrete-observation, ergodic HMM that took sequences of speech symbols as input and produced a source language hypothesis as output. Training and test symbol sequences were derived from published phonetic transcriptions of text. Riek [3], Nakagawa [4], Zissman [5], and Savic [23], all applied HMMs to feature vectors derived automatically from the speech signal. In these systems, HMM training was performed on unlabeled training speech. Riek and Zissman found that HMM systems trained in this unsupervised manner did not perform as well as some of the static classifiers that had been testing. Nakagawa, however, eventually obtained better performance for his HMM approach than his static approaches [24]. In related research, Li and Edwards [25] segmented incoming speech into six broad acoustic-phonetic classes. Finite-state models were used to model transition

probabilities as a function of language. Li has also developed a new language-ID system based on the examination and coding of spectral syllabic features [26].

Recently, language-ID systems that are trained using multi-language, phonetically labeled corpora have been proposed. Lamel and Gauvain have found that likelihood scores emanating from language-dependent phone[1] recognizers are very capable of discriminating between English and French read speech [28], as did Muthusamy on English versus Japanese spontaneous, telephone-speech [10]. This type of system will be covered in Section IV. Andersen [29] and Berkling [30] have explored the possibility of finding and using only those phones that best discriminate between language pairs. While initially these systems were constrained to operate only when phonetically transcribed training speech was available, Tucker [8] and Lamel [31] have utilized single-language phone recognizers to label multi-lingual training speech corpora, which have then been used to train language-dependent phone recognizers for language ID. Kadambe [32] has studied the effect of applying a lexical access module after phone recognition, in some sense spotting words in the phone sequences.

A related approach has been to use a single-language phone recognizer as a front-end to a system that uses phonotactic scores to perform language ID. Phonotactics are the language-dependent set of constraints specifying which phones/phonemes are allowed to follow other phones/phonemes. For example, the German word "spiel" which is pronounced /SH P IY L/ and might be spelled in English as "shpeel" begins with a consonant cluster /SH P/ that is rare in English.[2] This approach is reminiscent of the work of D'Amore [33], [34], Schmitt [35], and Damashek [36], who have used n-gram analysis of text documents to perform language and topic identification and clustering. Albina [37] extended the same technique to clustering speech utterances by topic. By "tokenizing" the speech message, i.e., converting the input waveform to a sequence of phone symbols, the statistics of the resulting symbol sequences can be used to perform either language or topic identification. Hazen [6], Zissman [7], and Tucker [8] have each developed such language-ID systems by using single-language front-end phone recognizers. Zissman [7] and Yan [38] have extended this work to system using multiple, single-language front-ends, for which there need not be a front-end in each language to be identified. Meanwhile, Hazen [39] has pursued a single multi-language front-end phone recognizer. Examples of some of these types of systems will be explored more fully below.

Prosodic features, such as duration, pitch, and stress have also been used to distinguish automatically one language from another. For example, Hutchins [40] has been successful in applying prosodic features to two-language LID (e.g., English versus Spanish, English versus Japanese, etc.), and Itahashi [41] has applied such features to six-way language ID.

Finally, within the past year, efforts at a number of sites have focused on the use of continuous speech recognition systems for language ID (e.g., [42]). During training, one speech recognizer per language is created. During testing, each of these recognizers is run in parallel, and the one yielding output with highest likelihood is selected as the winning recognizer—the language used to train that recognizer is the hypothesized language of the utterance. Such systems promise high-quality language identification because they use higher-level knowledge (words and word sequences) rather than lower-level knowledge (phones and phone sequences) to make the language-ID decision. Furthermore, one obtains a transcription of the utterance as a byproduct of language ID. On the other hand, continuous speech recognition systems require many hours of labeled training data in each language and also are the most computationally complex of the algorithms proposed. In a somewhat similar vein, Ramesh [43] has proposed text-dependent language ID via word spotting for situations in which the speaker's vocabulary is likely to be constrained.

III. Language-ID Cues

There are a variety of cues that humans and machines can use to distinguish one language from another. The reader is referred to the linguistics literature (e.g., [27], [44], [45]) for in depth discussions of how specific languages differ from one another, and to Muthusamy [46], who has measured how well humans can perform language ID. We know that the following characteristics differ from language to language:

- *Phonology.* Phone/phoneme sets are different from one language to another, even though many languages share a common subset of phones/phonemes. Phone/phoneme frequencies may also differ, i.e., a phone may occur in two languages, but it may be more frequent in one language than the other. Phonotactics, i.e., the rules governing the sequences of allowable phones/phonemes, can be different, as can be the prosodics.
- *Morphology.* The word roots and lexicons are usually different. Each language has its own vocabulary, and its own manner of forming words.
- *Syntax.* The sentence patterns are different. Even when two languages share a word, e.g., the word "bin" in English and German, the sets of words that may precede and follow the word will be different.
- *Prosody.* Duration, pitch, and stress differ from one language to another.

At present, all automatic language-ID systems of which the author is aware take advantage of one or more of these sets of language traits in discriminating one language from another.

IV. Algorithms

The algorithms described above have varying levels of computational complexity and different requirements for the training data. Our primary goal in this work was to evaluate

[1] The term "phone" is used to identify the realization of acoustic-phonetic units or segments, whereas a "phoneme" is an underlying mental representation of a phonological unit in a language [27]. Because most of the recognizers described in this paper are trained on phonetically labeled speech, i.e., the labels describe what was actually said, rather than phonemically labeled speech, in which the labels are found by dictionary lookup, the term "phone recognizer" will be used instead of "phoneme recognizer." Admittedly, this choice is somewhat arbitrary.

[2] This cluster can occur in English only if one word ends in /SH/ and the next begins with /P/, or in a compound word like "flashpoint."

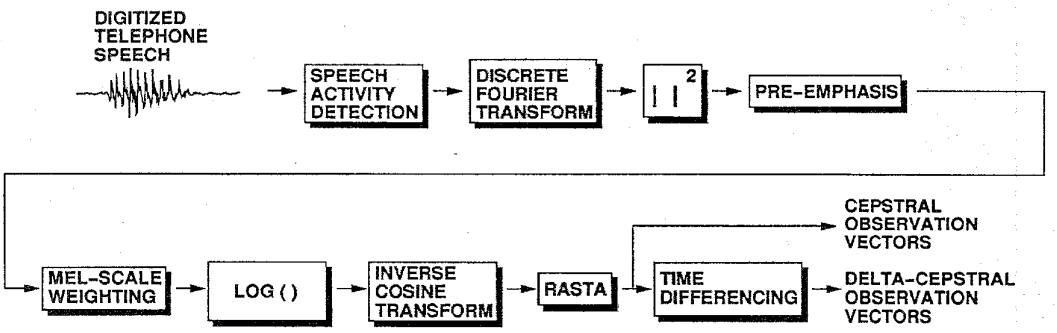

Fig. 1. Acoustic preprocessing used to convert telephone speech into feature vectors. Digitized speech is passed through a mel-scale filter bank from which cepstral and delta-cepstral feature vectors are created. Silence is removed automatically. RASTA is applied to help remove telephone channel effects.

a few of these techniques in a consistent manner to compare their language-ID capabilities. We tested four language-ID algorithms: Gaussian mixture modeling, single-language phone recognition followed by language modeling, parallel phone recognition followed by language modeling, and parallel phone recognition. Each of these systems is described in this section. The descriptions are preceded by a discussion of the conversion of speech to feature vectors, which is a process common to all four algorithms.

A. Converting Telephone Speech into Feature Vectors

In the four systems we examined, training and recognition are preceded by feature extraction, i.e., the speech waveforms are converted from their digital waveform representations (usually 16-bit linear or 8-bit μ-law encodings) to one or more streams of feature vectors. Fig. 1 shows a block diagram of the pseudo filter-bank. The acoustic preprocessor produces one mel-cepstral observation vector every 10 ms using a 20 ms window. This type of front-end was studied by Davis and Mermelstein [47], and the version used at Lincoln Laboratory for speech recognition, speaker ID, and language ID was implemented by Paul [48]. For language ID, only the lowest 13 coefficients of the mel-cepstrum are calculated (c_0 through c_{12}), thereby retaining information relating to the speaker's vocal tract shape while largely ignoring the excitation signal. The lowest cepstral coefficient (c_0) is ignored, because it contains only overall energy level information. The next twelve coefficients (c_1 through c_{12}) form the cepstral feature vector. Because the mel-cepstrum is a relatively orthogonal feature set, in that its coefficients tend not to be linearly related, it has been used widely for many types of digital speech processing.

In an effort to model cepstral transition information, difference cepstra are also computed and modeled. This vector of cepstral differences, or "delta" cepstral vector, (Δc_0 through Δc_{12}) is computed every frame as

$$\Delta c_i(t) = c_i(t+1) - c_i(t-1). \tag{1}$$

Note that Δc_0 is included as part of the delta-cepstral vector, thus making 13 coefficients altogether. For historical reasons relating to our use of tied mixture GMM systems, we process this vector as a separate, independent stream of observations, though it could be appended to the cepstral vector to obtain a 25-D composite vector.

When training or test speech messages comprise active speech segments separated by long regions of silence, we have found it desirable to train or test only on the active speech regions because the nonspeech regions typically contain no language-specific information. The speech activity detector we use was developed by Reynolds for pre-processing speech in his speaker-ID system [49]. To separate speech from silence, it relies on a time-varying estimate of the instantaneous signal-to-noise ratio (SNR).

Recognizing that the cepstral feature vectors can be influenced by the frequency response of the communications channel and in light of the possibility that each individual message is collected over a channel that is different from all other channels, we apply RASTA to remove slowly varying, linear channel effects from the raw feature vectors [50]. In this process, each feature vector's individual elements, considered to be separate streams of data, are passed through identical filters that remove near-DC components along with some higher frequency components. For each vector index i, the RASTA filtered coefficient, c_i', is related to the original coefficient, c_i as follows:

$$c_i'(t) = h(t) * c_i(t) \tag{2}$$

where $*$ denotes the convolution operation, and t is the time index measured in frames. We use the standard RASTA IIR filter

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.98z^{-1})}. \tag{3}$$

In some initial LID experiments using the GMM system, RASTA's impact on language-ID performance was found to be almost identical to that of long-term cepstral mean subtraction, but with the computational advantage of requiring only a single pass over the input data. Additionally, RASTA is capable of tracking changes in channel characteristics with time. RASTA may be performed equivalently in either the log mel-filter domain or the cepstral domain; we applied it to the log mel-filter coefficients.

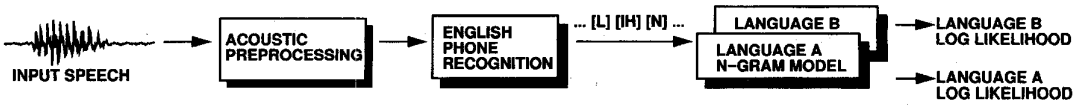

Fig. 2. PRLM block diagram. A single-language phone recognition front end is used to tokenize the input speech. The phone sequences output by the front end are analyzed, and a language is hypothesized.

B. Algorithm 1: Gaussian Mixture Model (GMM) Classification

A GMM language-ID system served as the simplest algorithm for this study. As will be shown below, GMM language ID is motivated by the observation that different languages have different sounds and sound frequencies. It has been applied to language ID at several sites [3]–[5].

Under the GMM assumption, each feature vector $\vec{v}_t$ at frame time t is assumed to be drawn randomly according to a probability density that is a weighted sum of multi-variate Gaussian densities:

$$p(\vec{v}_t|\lambda) = \sum_{k=1}^{N} p_k b_k(\vec{v}_t) \tag{4}$$

where λ is the set of model parameters

$$\lambda = \{p_k, \vec{\mu}_k, \Sigma_k\} \tag{5}$$

k is the mixture index ($1 \leq k \leq N$), the p_k's are the mixture weights constrained such that $\Sigma_{k=1}^{N} p_k = 1$, and the b_k's are the multi-variate Gaussian densities defined by the means $\vec{\mu}_k$ and variances Σ_k.

For each language l, two GMMs are created: one for the cepstral feature vectors, $\{\vec{x}_t\}$, and one for the delta-cepstral feature vectors, $\{\vec{y}_t\}$, as follows:

- From training speech spoken in language l, two independent feature vector streams are extracted: centisecond mel-scale cepstra (c_1 through c_{12}) and delta-cepstra (Δc_0 through Δc_{12}), as described in Section IV-A.
- A modified version of the Linde, Buzo, and Gray algorithm [51] is used to cluster each stream of feature vectors, producing 40 cluster centers for each stream (i.e., $N = 40$).
- By using the cluster centers as initial estimates for $\vec{\mu}_k$, multiple iterations of the estimate-maximize (E-M) algorithm are run, producing, for each stream, a more likely set of p_k, $\vec{\mu}_k$, Σ_k [52], [53].

During recognition, an unknown speech utterance is classified by first converting the digitized waveform to feature vectors and then by calculating the log likelihood that the language l model produced the unknown speech utterance. The log likelihood, $\mathcal{L}$, is defined as

$$\mathcal{L}(\{\vec{x}_t, \vec{y}_t\}|\lambda_l^C, \lambda_l^{DC}) = \sum_{t=1}^{T} [\log p(\vec{x}_t|\lambda_l^C) + \log p(\vec{y}_t|\lambda_l^{DC})] \tag{6}$$

where λ_l^C and λ_l^{DC} are the cepstral and delta-cepstral GMM, respectively, for language l, and T is the duration of the utterance. 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

$$\hat{l} = \arg\max_{l} \mathcal{L}(\{\vec{x}_t, \vec{y}_t\}|\lambda_l^C, \lambda_l^{DC}) \tag{7}$$

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

C. Algorithm 2: Phone Recognition Followed by Language Modeling (PRLM)

The second language-ID approach we tested comprises a single-language phone recognizer followed by an n-gram analyzer, as shown in Fig. 2 [6]–[8]. In this system, training messages in each language l are tokenized by a single-language phone recognizer, the resulting symbol sequence associated with each of the training messages is analyzed, and an n-gram probability distribution language model is estimated for each language l. Note that the n-gram probability distributions are trained from the output of the single-language phone recognizer, not from human-supplied orthographic or phonetic labels. During recognition, a test message is tokenized and the likelihood that its symbol sequence was produced in each of the languages is calculated. The n-gram model that results in the highest likelihood is identified, and the language of that model is selected as the language of the message.

PRLM is motivated by a desire to use speech sequence information in the language-ID process, thereby exploiting a larger range of phonology differences between languages than is possible with GMM. We view it as a compromise between:

- modeling the sequence information using hidden Markov models (HMM's) trained from unlabeled speech (such systems have performed no better than static classification [4], [5], though Nakagawa has had some success more recently [24]), and
- employing language-dependent phone recognizers trained from orthographically or phonetically labeled speech (such systems, which are the subject of Section IV-E, can be difficult to implement, as labeled speech in every language of interest is often not available).

1) The Front-End: Single-Language Phone Recognition: Though PRLM systems can employ a single-language phone recognizer trained from speech in any language, we focused initially on English front-ends, because labeled English speech

corpora were the most readily available.[3] The phone recognizer, implemented using the hidden Markov model toolkit (HTK) [54], is a network of context-independent phones ("monophones"), in which each phone model contains three emitting states. The output vector probability densities are modeled as GMMs with six underlying Gaussian densities per state per stream. The observation streams are the same cepstral and delta-cepstral vectors used in the GMM system. Phone recognition is performed via a Viterbi search using a fully connected null-grammar network of monophones. Phone recognition, which dominates PRLM processing time, takes about 1.5× real-time on a Sun SPARCstation-10 (i.e., a 10 s utterance takes about 15 s to process).

2) The Back-End: N-gram Language Modeling: Using the English phone recognizer as a front-end, a language model can be trained for each language l by running training speech for language l into the phone recognizer and computing a model for the statistics of the phones and phone sequences that are output by the recognizer. We count the occurrences of n-grams: subsequences of n symbols (phones, in this case). Training is performed by accumulating a set of n-gram histograms, one per language, under the assumption that different languages will have different n-gram histograms. We then use interpolated n-gram language models [55] to approximate the n-gram distribution as the weighted sum of the probabilities of the n-gram, the (n −1)-gram, etc. An example for a bigram model (i.e., $n = 2$) is

$$\tilde{P}(w_t|w_{t-1}) = \alpha_2 P(w_t|w_{t-1}) + \alpha_1 P(w_t) + \alpha_0 P_0. \quad (8)$$

where w_{t-1} and w_t are consecutive symbols observed in the phone stream. The P's are ratios of counts observed in the training data, e.g.:

$$P(w_t|w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_{t-1})} \quad (9)$$

where $C(w_{t-1}, w_t)$ is the number of times symbol w_{t-1} is followed by w_t, and $C(w_{t-1})$ is the number of occurrences of symbol w_{t-1}. P_0 is the reciprocal of the number of symbol types. 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 front-end phone recognizer, producing a phone sequence, $W = \{w_0, w_1, w_2, \cdots\}$. The log likelihood, $\mathcal{L}$, that the interpolated bigram language model for language l, λ_l^{BG}, produced the phone sequence W, is

$$\mathcal{L}(W|\lambda_l^{BG}) = \sum_{t=1}^{T} \log \tilde{P}(w_t|w_{t-1}, \lambda_l^{BG}) \quad (10)$$

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

$$\hat{l} = \arg\max_l \mathcal{L}(W|\lambda_l^{BG}). \quad (11)$$

Based on early experiments, we set $n = 2$, $\alpha_2 = 0.399$, $\alpha_1 = 0.6$, and $\alpha_0 = 0.001$ for PRLM experiments, as we found that peak performance was obtained in the region of $0.3 < \alpha_1, \alpha_2 < 0.7$. We have found little advantage to using $n > 2$ for PRLM, and this observation is consistent with other sites [1].[4] Our settings for α and n are surely related to the amount of training speech available; for example, one might weight the higher order α's more heavily as the amount of training data increases.

[3] Ultimately, we tested single-language front-ends in six different languages.

D. Algorithm 3: Parallel PRLM

Although PRLM is an effective means of identifying the language of speech messages (as will be shown in Section VI), we know that the sounds in the languages to be identified do not always occur in the one language used to train the front-end phone recognizer. Thus, it seems natural to look for a way to incorporate phones from more than one language into a PRLM-like system. For example, Hazen has proposed to train a front-end recognizer on speech from more than one language [39]. Alternatively, our approach is simply to run multiple PRLM systems in parallel with the single-language front-end recognizers each trained in a different language [7], [38]. This approach requires that labeled training speech be available in more than one language, although the labeled training speech does not need to be available for all, or even any, of the languages to be recognized. An example of such a parallel PRLM system is shown in Fig. 3. In the example, we have access to labeled speech corpora in English, Japanese, and Spanish, but the task at hand is to perform language classification of messages in Farsi, French, and Tamil. To perform the classification, we first train three separate PRLM systems: one with an English front-end, another with a Japanese front-end, and the last with a Spanish front-end. This parallel PRLM system would have a total of nine n-gram language models—one for each language to be identified (Farsi, French, and Tamil) per each front-end (English, Japanese, Spanish). During recognition, a test message is processed by all three PRLM systems, and their outputs are averaged in the log domain (multiplied in the linear domain, as if each PRLM system were operating independently) to calculate overall language log likelihood scores. Note that this approach extends easily to any number of parallel PRLM systems. The only limitation is the number of languages for which labeled training speech is available. The phone recognizer parameters (e.g., number of states, number of Gaussians, etc.) used in parallel PRLM are identical to those used in PRLM. Parallel PRLM processing time is about 1.5× real-time on a Sun SPARCstation-10 per front-end phone recognizer; therefore, a system with phone recognizers in three languages (e.g., English, Japanese, and Spanish) would take about 4.5× real-time.

E. Algorithm 4: Parallel Phone Recognition (PPR)

The PRLM and parallel PRLM systems perform phonetic tokenization followed by phonotactic analysis. Though this approach is reasonable when labeled training speech is not available in each language to be identified, the availability

[4] Though not yet effective for PRLM-based language ID, trigrams have been used successfully in other types of language-ID systems (e.g., [24], [32]).

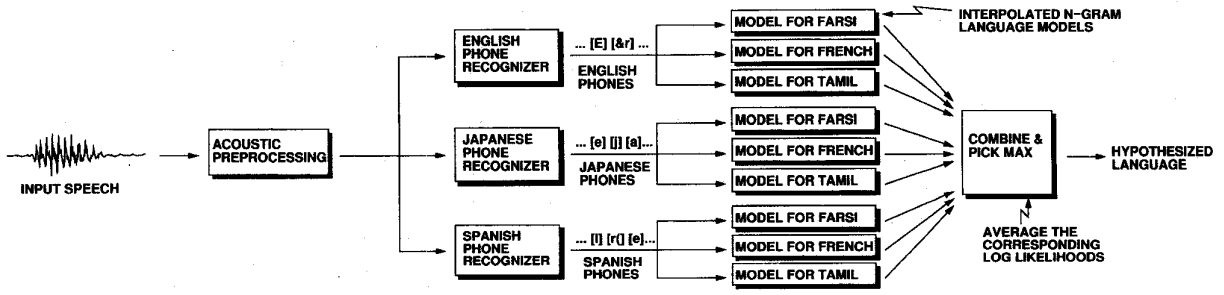

Fig. 3. Parallel PRLM block diagram. Several single-language phone recognition front ends are used in parallel to tokenize the input speech. The phone sequences output by the front ends are analyzed, and a language is hypothesized.

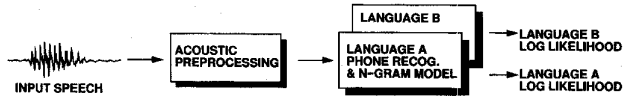

Fig. 4. PPR block diagram. Several single-language phone recognition front ends are used in parallel. The likelihoods of the Viterbi paths through each system are compared from which a language is hypothesized.

of such labeled training speech broadens the scope of possible language-ID strategies; for example, it becomes easy to train and use integrated acoustic/phonotactic models. By allowing the phone recognizer to use the language-specific phonotactic constraints *during* the Viterbi decoding process rather than applying those constraints *after* phone recognition is complete (as is done in PRLM and parallel PRLM), the most likely phone sequence identified during recognition is optimal with respect to some combination of both the acoustics and phonotactics. The joint acoustic-phonotactic likelihood of that phone sequence would seem to be well-suited for language ID. Thus, we tested such a parallel phone recognition (PPR) system, as shown in Fig. 4. Like PRLM, PPR makes use of the phonological differences between languages. Such systems had been proposed previously by Lamel [9] and Muthusamy [10].

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

$$a_{ij} = s \log \tilde{P}(j|i) \tag{12}$$

where s is the grammar scale factor, and the $\tilde{P}$'s are bigram probabilities derived from the training labels. Based on preliminary testing, $s = 3$ was used in these experiments, as performance was seen to have a broad peak near this value. Another difference between PRLM and PPR phone recognizers is that while both can use context-dependent phone models, our PRLM phone recognizers use only monophones while our PPR phone recognizers use the monophones of each language plus the 100 most commonly occurring right (i.e., succeeding) context-dependent phones. This strategy was motivated by initial experiments showing that context-dependent phones improved PPR language-ID performance but had no effect on PRLM language-ID performance.

PPR language-ID is performed by Viterbi decoding the test utterance once for each language-dependent phone recognizer. Each phone 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. During some initial experiments, we found that the log likelihood scores were biased, i.e., the scores out of a recognizer for language l were higher, on average, than the scores from the language m recognizer. We speculate that this effect might be due to the use of the Viterbi (best-path) log likelihood rather than the full log likelihood across all possible paths. Alternatively, the bias might have been caused by a language-specific mismatch between speakers or text used for training and testing. Finally, it might be that these biases represent different degrees of mismatch between the HMM assumptions and various natural languages. In any case, to hypothesize the most likely language in our PPR system, we use a modified maximum-likelihood criterion in which a recognizer-dependent bias is subtracted from each log likelihood score prior to applying the maximum-likelihood decision rule. Instead of finding $\hat{l}$,

$$\hat{l} = \arg\max_{l} \mathcal{L}(\hat{p}_l|\lambda_l) \tag{13}$$

we find $\hat{l}'$,

$$\hat{l}' = \arg\max_{l} [\mathcal{L}(\hat{p}_l|\lambda_l) - K_l] \tag{14}$$

where $\mathcal{L}(\hat{p}|\lambda_l)$ is the log likelihood of the Viterbi path $\hat{p}_l$ through the language l phone recognizer and K_l is the recognizer-dependent bias. The recognizer-dependent bias is set to the average of the normalized log likelihoods for all messages processed by the recognizer. In preliminary tests, this heuristic bias-removal technique was shown to reduce the error rate by a factor of two. The PPR recognizer for each language runs at about 2× real-time on a Sun SPARCstation-10.

Note that PPR systems require labeled speech for *every* language to be recognized; therefore, it may be more difficult to implement a PPR system than any of the other systems already discussed, although Tucker [8] and Lamel [31] have bootstrapped PPR systems by using labeled training speech in only one language, and Lund [56] has developed a technique for using acoustic models in one language to train language models (and run phone recognizers) in many languages (a bridge between PRLM and PPR).

TABLE I
OGI MULTI-LANGUAGE TELEPHONE SPEECH CORPUS

Language	Initial Training		Development Test		Extended Training		Final Test	
	male	female	male	female	male	female	male	female
English	33	17	14	6	72	30	16	4
Farsi	39	10	15	4	8	1	18	2
French	40	10	15	5	11	2	12	8
German	25	25	11	9	10	5	15	5
Hindi	47	3	13	4	25	11	14	6
Korean	32	17	18	2	3	2	15	5
Japanese	30	20	15	5	1	0	11	8
Mandarin	34	15	14	6	8	8	10	10
Spanish	34	16	16	4	14	5	11	8
Tamil	43	7	17	3	20	2	19	1
Vietnamese	31	19	16	4	11	6	13	7

V. SPEECH CORPUS

The Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [11] was used to evaluate the performance of each of the four language-ID approaches outlined above.[5] Each message in the corpus was spoken by a unique speaker over a telephone channel and comprises responses to ten prompts, four of which elicit fixed text (e.g., "Please recite the seven days of the week," "Please say the numbers zero through ten") and six of which elicit free text (e.g., "Describe the room from which you are calling," "Speak about any topic of your choice"). The ten responses contained in each message together comprise about two minutes of speech.

Table I contains a listing of the number of messages per language in each of the four segments of the corpus: initial training, development test, extended training, and final test. Our GMM, PRLM, parallel PRLM, and PPR comparisons were run with the initial training segment for training and the development test set for testing. Because the Hindi messages were not yet available when we performed our preliminary test, only ten languages were used. Test utterances were extracted from the development test set according to the April 1993 National Institute of Standards and Technology (NIST) specification [57]:

"45 s" Utterance Testing: Language ID is performed on a set of 45 s utterances spoken by the development test speakers. These utterances are the first 45 s of the responses to the prompt "speak about any topic of your choice." OGI refers to these utterances as "stories before the tone," and they are denoted `story-bt`.[6]

"10 s" Utterance Testing: Language ID is performed on a set of 10 s cuts from the same stories utterances used in "45 s" testing.

Phonetic labels for six of the languages were provided by OGI during the course of this work. English, Japanese, and Spanish labels were provided first, followed by German, Hindi, and Mandarin. For all six languages, labels were provided only for the `story-bt` utterances. We compared GMM, PRLM, parallel PRLM, and PPR using only the English, Japanese, and Spanish messages. Additional experiments that compared only the GMM, PRLM, and parallel PRLM systems used messages in all ten languages.

Though the same OGI-TS messages were used to train each of the four systems, the systems used the training data in different ways. The Gaussian mixture models were trained on the responses to the six free-text prompts. The PRLM back-end language models and the phone recognizers for the parallel PRLM and PPR systems were trained on the `story-bt` utterances. For the PRLM system, three different English front-ends were trained:

- A phone recognizer was trained on the phonetically labeled messages of the English initial training segment of the OGI-TS corpus.[7] Models for 48 monophones were trained.
- A second phone recognizer was trained on the entire[8] training set of the NTIMIT telephone-speech corpus [58]. The data comprised 3.1 hr of read, labeled, telephone-speech recorded over many telephone channels using a single handset. Models for 48 monophones were trained.
- A third phone recognizer was trained on CREDITCARD excerpts from the SWITCHBOARD corpus [59]. The data comprised 3.8 hr of spontaneous, labeled, telephone-speech recorded using many handsets. Models for 42 monophones were trained.

Note that the number of monophone models trained is dependent on the corpus labeling scheme.

We conducted further testing of the parallel PRLM system after OGI released the extended-training segment and the Hindi messages. Single-language front-ends were eventually trained in six languages (English, German, Hindi, Japanese, Mandarin, Spanish). Language model training was performed on the union of the initial training, development test, and extended training segments. Test utterances were selected according to the March 1994 NIST specification, with both "45 s" and "10 s" utterances extracted from the final test set [57].

VI. EXPERIMENTS AND RESULTS

The four algorithms were compared by performing two-alternative and three-alternative, forced-choice classification experiments using the English, Japanese, and Spanish OGI-TS messages. As defined in Table I, this first set of experiments used the "Initial Training" data for training, and the "Development Test" data for testing. For the two-alternative testing, one model was trained on English speech and another on Japanese speech. Test messages spoken in English and Japanese were then presented to the system for classification. Similar experiments were run for English versus Spanish and Japanese versus Spanish. For the three-alternative testing, models were trained in all three languages, and test messages in all three languages were presented to the system for forced-choice classification. Results of all of these experiments are

[5] The OGI-TS corpus is available from the Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[6] A tone signaled the speaker when 45 s of speech had been collected, indicating 15 s remaining.

[7] Because the forward-backward algorithm would have had trouble aligning phone models against 45 s utterances, shorter, hand endpointed segments of the story utterances were used. This also resulted in less heavy reliance on the OGI supplied phone start and end times.

[8] Except for the shibboleth sentences.

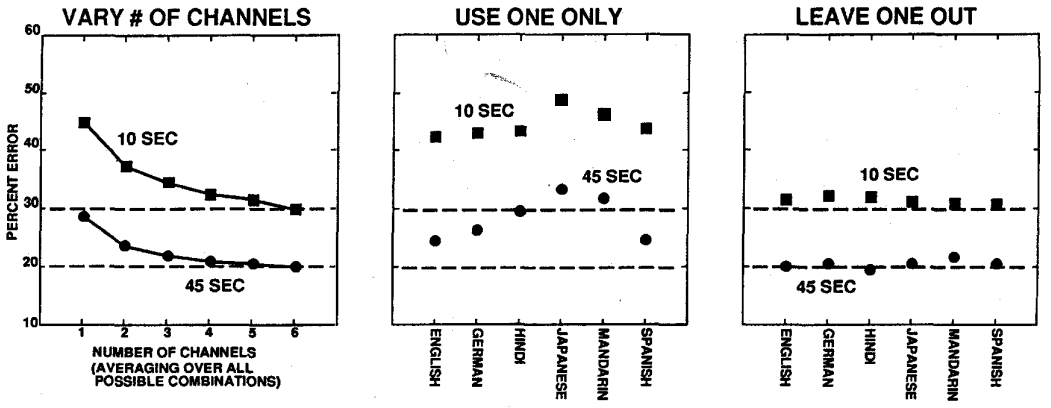

Fig. 5. Using fewer than six front-ends. Left panel shows the average effect of reducing the number of channels. Middle panel shows the effect of using only one channel. Right panel shows the effect of omitting one of the six channels.

TABLE II
RESULTS COMPARING ALL FOUR SYSTEMS (% ERROR)

System	Eng./Jap.		Eng./Spa.		Jap./Spa.		2L Average		3L	
	45-s	10-s	45-s	10-s	45-s	10-s	45-s	10-s	45-s	10-s
GMM	17	16	17	16	35	36	23	23	35	36
PRLM (SWITCHBOARD)	6	12	3	15	12	22	7	16	10	27
Parallel PRLM	9	10	3	12	6	10	6	11	8	15
PPR	6	8	3	8	15	13	8	10	14	15
σ (standard deviation)							4	2	6	3

shown in Table II. In the table, the averages were computed with equal-weighting per language pair. Standard deviations for the last four columns were computed with the assumption of a binomial distribution. Generally, the results show that parallel PRLM and PPR perform about equally. This result is not surprising because the major difference between the two systems for these three languages is the manner in which the language model is applied. For the 45 s utterances, SWITCHBOARD-based PRLM performs about as well as parallel PRLM and PPR, though it performs worse than parallel PRLM and PPR for the shorter, 10 s utterances.

Some additional experiments were run comparing PRLM, parallel PRLM, and GMM using all ten languages of the OGI-TS corpus. PPR could not be run in this mode, as phonetic labels did not exist for all of the languages. The first two columns of Table III show ten language, forced-choice results. Next, two language, forced-choice average results for English versus each of the other nine languages are presented. The final two columns show two-language, forced-choice results averaged over all of the 45 language pairs. Approximate standard deviations are shown in the bottom row. Table III shows that parallel PRLM generally performs best. Also note that PRLM with a SWITCHBOARD front-end performs about equally to PRLM with an OGI-TS English front-end. PRLM with an NTIMIT front-end performs rather poorly, perhaps because there are significant differences between the recording conditions of the OGI-TS and NTIMIT corpora: NTIMIT speech is read, and the entire corpus was collected using a single handset, while the OGI-TS corpus is extemporaneous, and it was recorded using hundreds of handsets. We suspect that the lack of handset variability in NTIMIT caused the poor performance.

TABLE III
FULL TEN-LANGUAGE RESULTS (% ERROR)

System	10L		Eng. vs. L		L vs. L'	
	45-s	10-s	45-s	10-s	45-s	10-s
GMM	47	50	19	16	20	21
PRLM (NTIMIT)	33	53	12	18	10	16
PRLM (SWITCHBOARD)	28	46	5	12	8	14
PRLM (OGI-ENGLISH)	28	46	7	13	8	14
Parallel PRLM	21	37	8	12	6	10
σ (standard deviation)	3	2	2	1	1	1

TABLE IV
PARALLEL PRLM RESULTS (% ERROR) USING MARCH 1994 NIST GUIDELINES

11L		Eng. vs. L		L vs. L'	
45s	10s	45s	10s	45s	10s
20	30	4	6	5	8

Table IV shows the results of evaluating the parallel PRLM system according to the March 1994 NIST guidelines. With the addition of Hindi, the first two columns refer to eleven-alternative, forced-choice classification, the next two columns refer to an average of the ten two-alternative, forced-choice experiments with English and one other language, and the last two columns refer to an average of the 55 two-alternative, forced-choice experiments using each pair of languages. Six front-end phone recognizers (English, German, Hindi, Japanese, Mandarin, and Spanish) were used for this experiment. As defined in Table I, this second set (and all subsequent sets) of experiments used the initial training, development test, and extended training data for training, and the final test data for testing. Table IV shows our first pass through the final test evaluation data, so for these results there was no possibility of tuning the system to specific speakers or messages.

Further analysis of our NIST March 1994 results was performed to determine the effect of reducing the number of front-end phone recognizers. The results on the eleven-language classification task are shown in Fig. 5. The left panel shows that reducing the number of channels generally increases the error rate more quickly for the 10 s utterances than the 45 s utterances. The middle panel shows that using

TABLE V
PPR PHONE RECOGNITION RESULTS

	Error Rate %	N	I	S	D	H	Number of monophones	# phone classes
English	58.1	8269	966	2715	1120	4434	52	39
Japanese	44.5	7949	864	945	1730	5274	27	25
Spanish	45.1	7509	733	1631	1021	4857	38	34

only one channel, no matter which one it is, greatly increases the error rate. The right panel shows that omitting any one of the six channels has only a small impact.

We also measured the within-language accuracy of a few of the front-end recognizers; i.e., we tested the English recognizer with English, the Spanish recognizer with Spanish, and so on. Table V shows the within-language phone recognition performance of the PPR recognizers. The results are presented in terms of error rate, i.e., the sum of the substitution, deletion, and insertion errors, divided by the true number of phones. N is number of actual phone tokens in test set, I is number of insertions, S is number of substitutions, D is number of deletions, H is number of phones correctly identified. Note that for each language, the number of equivalence classes (i.e., those classes of similar phones that are, for the intent of scoring, considered equivalent) is less than the number of monophones. Equivalence classes for English were motivated by Lee and Hon [60]. For Japanese and Spanish, similar rules were applied. The 10 s utterances from the development test set were used to evaluate phone recognition performance. For these evaluations, the phone networks included all context-independent and right context-dependent phones observed in the training data. The results of Tables II and V indicate that individual PPR recognizers can exhibit a high phone-recognition error rate while still allowing the overall PPR system to achieve good language-ID performance.

Although not measured, we believe that our PRLM phone recognizers, which do not employ any context-dependent phones, have even higher error rates than our PPR phone recognizers. Therefore, it is interesting that their output can be used to perform language ID effectively. Some preliminary studies indicate that mutual information, as opposed to phone accuracy, might be a better measure of front-end utility. As suggested by Gish [61], mutual information of the front-end measures jointly the resolution of the phone recognizer and its consistency. Consistency, rather than accuracy, is what is required by the language models; after all, if phone a is always recognized by a two-phone front-end as phone b, and phone b is always recognized as phone a, the accuracy of the front-end might be zero, but the ability of the language model to perform language ID will be as high as if the front-end made no mistakes. That bigram performance is better than unigram performance, even though we rarely recognize a bigram "accurately," might be due to the fact that we can recognize bigrams "consistently."

VII. ADDITIONAL EXPERIMENTS

Given the high performance of the parallel PRLM approach, our attention has focused on ways of boosting its language-ID capabilities even further. In this section, we report on efforts to use gender-dependent phonotactic weighting and duration tagging to improve parallel PRLM language-ID performance.

A. Gender-Dependent Channels

Using gender-dependent acoustic models is a popular technique for improving speech recognition performance (e.g., [62]–[64]). We were motivated to use gender-dependent front-ends and back-ends for two reasons:

- Gender-dependent phone recognizers should produce a more reliable tokenization of the input speech relative to their gender-independent counterparts; therefore, n-gram analysis should prove more effective.
- The acoustic likelihoods output by gender-dependent phone recognizers could be used to weight the phonotactic scores output by the interpolated language models. This weighting procedure would represent our first use of acoustic likelihoods in a PRLM-type system.

The general idea of employing gender-dependent channels for language ID is to make a preliminary 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 Fig. 6. During training, three phone 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 l 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 phone recognizer are trained only on male messages, the female language models only on 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[9]

$$\Pr(\text{male}|x) = \frac{p(x|\Lambda_m)}{p(x|\Lambda_m) + p(x|\Lambda_f)} \tag{15}$$

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:

$$W_m = \begin{cases} \dfrac{\Pr(\text{male}|x) - K}{1 - K} & \text{if } \Pr(\text{male}|x) \geq K \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

$$W_f = \begin{cases} \dfrac{K - \Pr(\text{male}|x)}{K} & \text{if } \Pr(\text{male}|x) < K \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

$$W_{gi} = \begin{cases} 1 - W_m & \text{if } \Pr(\text{male}|x) \geq K \\ 1 - W_f & \text{if } \Pr(\text{male}|x) < K \end{cases} \tag{18}$$

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

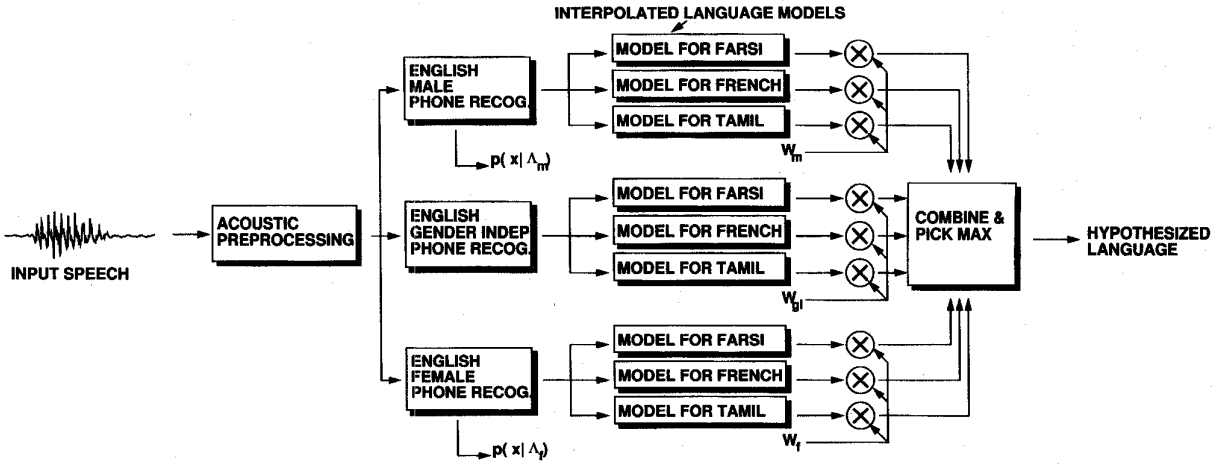

Fig. 6. Example of gender-dependent processing for a channel with an English front-end. The acoustic likelihoods $p(x|\Lambda_m)$ and $p(x|\Lambda_f)$ are used to compute the weights W_m, W_f, and W_{gi}.

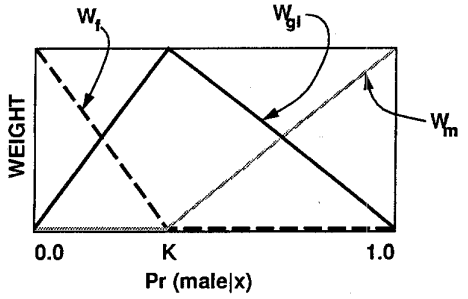

Fig. 7. Three weight functions. The value of each weight is a function of $\Pr(_{\text{male}}|x)$.

where W_m, W_f, and W_{gi} are the weights for the male, female, and gender-independent channels, respectively, and K is a constant set empirically during training (typically ranging from 0.30–0.70). The weight functions are shown graphically in Fig. 7. The W's are used to weight the phonotactic language model scores as follows.

$$p(x|l) = W_m p(x|\lambda_l^m) + W_f p(x|\lambda_l^f) + W_{gi} p(x|\lambda_l^{gi}) \quad (19)$$

where λ_l^m is the interpolated n-gram language model trained by passing male language l speech through the male phone recognizer, λ_l^f is the interpolated n-gram language model trained by passing female language l speech through the female phone recognizer, λ_l^{gi} is the interpolated n-gram language model trained by passing both male and female language l speech through the gender-independent phone recognizer.

B. Duration Tagging

On advice from Mistretta at Lockheed-Martin Sanders [65], we have begun to use phone duration to improve the performance of our parallel PRLM system. Duration tagging makes explicit use of phone-duration information that is output from the front-end phone recognizers. Our version of the Lockheed-Martin Sanders approach for using duration information is shown in Fig. 8. The training data for all languages are passed through each of the front-end phone recognizers. A histogram of durations for each phone emitted from each recognizer is compiled and the average duration determined. A $-$L suffix is appended to all phones having duration longer than the average duration for that phone, and a $-$S suffix is appended to all phones having duration shorter than the average duration for that phone. This modified sequence of phone symbols is then used in place of the original sequence for training the interpolated language models. During recognition, we use the duration thresholds determined during training to apply the same procedure to the output symbols from the phone recognizer.

TABLE VI
PARALLEL PRLM PERFORMANCE WITH SEVERAL ENHANCEMENTS (% ERROR)

	11L		Eng. vs. L		L vs. L'	
System	45s	10s	45s	10s	45s	10s
Baseline	20	30	4	6	5	8
New Baseline	14	26	5	3	4	7
Gender	13	23	2	4	3	6
Duration	14	23	2	5	3	6
Gender + Duration	11	21	2	4	2	5
σ (standard deviation)	3	2	2	1	<1	<1

C. Results and Analysis

Use of gender-dependent front-ends together with gender-independent front-ends has resulted in a modest improvement in LID performance. Table VI compares the performance of six systems:

Baseline: Our baseline six-channel parallel PRLM system from the March 1994 evaluation (first row of results),

New Baseline: A newer version of the baseline system that has better silence detection and a better set of language model interpolation weights ($\alpha_2 = 0.599$, $\alpha_1 = 0.4$, and $\alpha_0 = 0.001$),

Gender: A 16 channel system having three front-ends (one male, one female, and one gender-independent) for English, German, Japanese, Mandarin, and Spanish, and one front-end for Hindi (as there was insufficient female speech to train gender-dependent front-ends for Hindi). These results

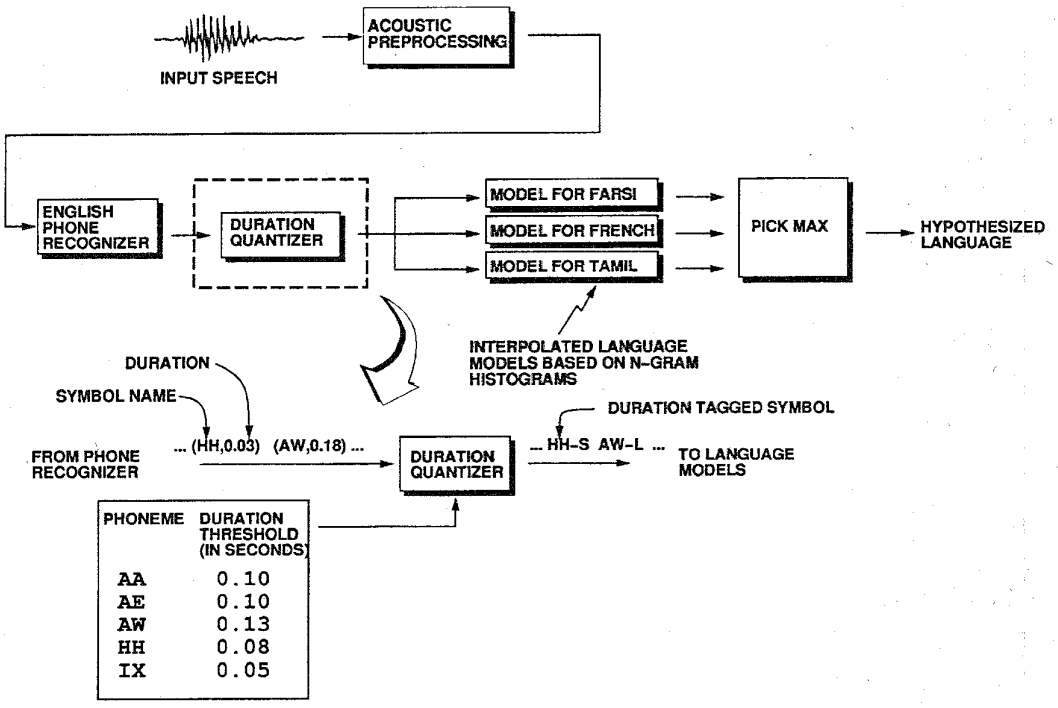

Fig. 8. Approach to duration tagging.

TABLE VII
CONFUSION MATRIX FOR 45-s UTTERANCES

	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
English	18	0	0	0	0	0	0	0	1	0	0
Farsi	0	19	0	0	0	0	0	0	0	0	0
French	2	0	13	1	0	1	0	0	0	0	0
German	0	1	1	17	0	0	0	0	0	0	0
Hindi	0	0	0	0	18	0	0	0	0	1	0
Japanese	0	0	1	0	1	16	0	0	1	0	0
Korean	0	0	0	0	0	0	11	0	0	0	1
Mandarin	0	0	1	0	0	0	2	14	0	0	0
Spanish	0	0	0	0	3	0	0	0	13	0	1
Tamil	0	0	0	0	0	0	0	0	0	14	0
Vietnamese	0	0	0	0	1	0	0	0	0	1	13

TABLE VIII
CONFUSION MATRIX FOR 10-s UTTERANCES

	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
English	61	0	2	0	0	0	2	0	2	2	0
Farsi	2	47	3	2	2	0	0	0	2	0	0
French	5	0	41	9	2	3	0	0	1	0	1
German	3	3	2	53	1	1	0	1	0	0	1
Hindi	3	2	0	2	51	0	0	1	1	5	0
Japanese	0	0	2	0	2	49	2	0	5	0	1
Korean	0	1	2	1	0	0	34	1	0	0	6
Mandarin	0	4	0	1	3	1	2	40	0	0	1
Spanish	2	1	1	1	6	2	0	0	40	1	4
Tamil	0	0	0	0	0	0	0	0	0	43	0
Vietnamese	2	2	1	0	3	0	1	0	3	1	34

represent an attempt to use the acoustic likelihoods output by the front-end phone recognizer to improve the phonotactic scores output by the n-gram language models.

Duration: A system that uses the simple technique for modeling duration using the $-S$ and $-L$ tags.

Gender + Duration: A system that combines the gender and duration enhancements.

Tables VII and VIII show the confusion matrices for the 45 s and 10 s utterances, respectively, using a parallel PRLM system with gender-dependent and duration processing. Each row shows the number of utterances truly spoken in some language, and each column shows the number of utterances classified by the system as spoken in some language. Therefore, entries along the main diagonal indicate utterances correctly identified, while off-diagonal entries are errors. From studying the confusion matrices, it becomes clear that confusions are not necessarily correlated with the linguistic "closeness" of the language pair. For example, there are many more Spanish/Hindi confusions than Spanish/French confusions. This may be due to the small size of the test corpus, which limits our confidence in these statistics.

We also have some evidence that the parallel PRLM system has trouble with nonnative speakers of a language. For Spanish, an expert Spanish dialectologist listened to each message, classifying the dialect of the speaker [66]. Though the Spanish speakers are generally native of Spain or Latin America, some were born and/or raised in other countries (e.g., the United States and France). Of the 13 Spanish speakers correctly identified as shown in Table VII, 10 are native speakers, and three are not. Of the four Spanish speakers incorrectly identified, all are nonnative, and one was nonfluent.

Although the phone recognizer acoustic likelihoods used in the gender weighting are already being calculated as part of the phone recognition process, and hence, we get them for free in our system, we have begun to use a simpler GMM-based algorithm for making the gender-ID decision. The GMM-

based approach to gender ID yields language-ID performance comparable with our original approach but allows for a more reliable separation between male and female speakers and obviates the computation of the K factor.

The use of even more fine-grain duration tags has been studied both by us and by Lockheed–Martin Sanders. In both cases, quantizing duration into more than two values has not improved language-ID performance.

VIII. Discussion

This paper has reviewed the research and development of language-identification systems at MIT Lincoln Laboratory. We began by comparing the performance of four approaches to automatic language identification of telephone-speech messages: Gaussian mixture modeling (GMM), single-language phone recognition followed by language modeling (PRLM), parallel PRLM, and parallel phone recognition (PPR). The GMM system, which requires no phonetically labeled training speech and runs faster than real-time on a conventional UNIX workstation, performed poorest. PRLM, which requires phonetically labeled training speech in only one language and runs a bit slower than real-time on a conventional workstation, performs respectably as long as the front-end phone recognizer is trained on speech collected over a variety of handsets and channels. Even better results were obtained when multiple front-end phone recognizers were used with either the parallel PRLM or PPR systems, but these systems run more slowly (e.g., $4\times$ to $24\times$ real-time). Because phonetically or orthographically labeling foreign language speech is expensive, the high performance obtained with the parallel PRLM system—which can use, but does not require, labeled speech for each language to be recognized—is encouraging.

With respect to a parallel PRLM system, we have shown that using gender-dependent front-ends in parallel with gender-independent front-ends can improve performance. This result is consistent with the experience of using gender-dependent models for continuous speech recognition. We have also used phone duration tagging to improve performance. On 45-s telephone speech messages, our very best system yields a 11% error rate in performing 11-language closed-set classification and a 2% error rate in performing two-language closed-set tasks.

NIST-sponsored language-ID evaluations occurred again in March 1995 and are planned for 1996. Details regarding the corpora used, the test guidelines, and the results may be obtained from NIST [57]. There is no doubt that the existence of the OGI-TS corpus together with the specification of evaluation scenarios by NIST has greatly enhanced our ability to do language-ID research. It has also allowed us to compare one language-ID system to another under carefully controlled conditions.

As automatic speech recognition systems become available for more and more languages, it is reasonable to believe that the availability of standardized, multilanguage 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 phone recognizers employing bigram grammars. Language-ID systems that use language-dependent word spotters [32], [43] and continuous speech recognizers [42] are evolving. These systems are moving beyond the use of phonology for language ID, incorporating both morphologic and syntactic information. It will be interesting to compare the performance and computational complexity of these newer systems to the systems we studied.

Acknowledgment

The author is grateful to R. Cole and his team at the Center for Spoken Language Understanding at the Oregon Graduate Institute for making available the OGI Multi-Language Telephone Speech Corpus. W. Mistretta and D. Morgan of Lockheed–Martin Sanders suggested the approach of duration tagging described in Section VII-B. K. Ng of BBN kindly provided high-quality transcriptions of the SWITCHBOARD CREDITCARD corpus. T. Chou and L. S. Sohn helped prepare the OGI-TS data at Lincoln for PPR processing. D. Reynolds, D. Paul, R. Lippmann, B. Carlson, T. Gleason, J. Lynch, J. O'Leary, C. Rader, E. Singer, and C. Weinstein of the Lincoln speech group offered significant technical advice. Lippmann, O'Leary, Reynolds, Weinstein, and five anonymous IEEE-appointed reviewers made numerous suggestions for improving this manuscript. A. Hayashi of the Lincoln Publications Group also contributed numerous editorial improvements.

References

[1] T. Hazen, Private communication.
[2] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Mag.*, vol. 11, no. 4, pp. 33–41, Oct. 1994.
[3] L. Riek, W. Mistreta, and D. Morgan, "Experiments in language identification," Lockheed Sanders, Inc., Nashua, NH, Tech. Rep. SPCOT-91-002, Dec. 1991.
[4] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text-independent language identification by HMM," in *Proc. ICSLP '92*, vol. 2, Oct. 1992, pp. 1011–1014.
[5] M. A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proc. ICASSP '93*, vol. 2, Apr. 1993, pp. 399–402.
[6] T. J. Hazen and V. W. Zue, "Automatic language identification using a segment-based approach," in *Proc. Eurospeech '93*, vol. 2, Sept. 1993, pp. 1303–1306.
[7] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *Proc. ICASSP '94*, vol. 1, Apr. 1994, pp. 305–308.
[8] R. C. F. Tucker, M. J. Carey, and E. S. Paris, "Automatic language identification using sub-words models," in *Proc. ICASSP '94*, vol. 1, Apr. 1994, pp. 301–304.
[9] L. F. Lamel and J.-L. Gauvain, "Identifying non-linguistic speech features," in *Proc. Eurospeech '93*, vol. 1, Sept. 1993, pp. 23–30.
[10] Y. Muthusamy *et al.*, "A comparison of approaches to automatic language identification using telephone speech," in *Proc. Eurospeech '93*, vol. 2, Sept. 1993, pp. 1307–1310.
[11] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. ICSLP '92*, vol. 2, Oct. 1992, pp. 895–898.
[12] R. G. Leonard, "Language recognition test and evaluation," RADC/Texas Instruments, Inc., Dallas, TX, Tech. Rep. RADC-TR-80-83, Mar. 1980.
[13] R. G. Leonard and G. R. Doddington, "Automatic language identification," RADC/Texas Instruments, Inc., Dallas, TX, Tech. Rep. RADC-TR-74-200/TI-347650, Aug. 1974.
[14] ——, "Automatic classification of languages," RADC/Texas Instruments, Inc., Dallas, TX, Tech. Rep. RADC-TR-75-264, Oct. 1975.

[15] ——, "Automatic language discrimination," RADC/Texas Instruments, Inc., Dallas, TX, Tech. Rep. RADC-TR-78-5, Jan. 1978.
[16] D. Cimarusti and R. B. Ives, "Development of an automatic identification system of spoken languages: Phase I," in *Proc. ICASSP '82,* May 1982, pp. 1661–1663.
[17] J. T. Foil, "Language identification using noisy speech," in *Proc. ICASSP '86,* vol. 2, Apr. 1986, pp. 861–864.
[18] F. J. Goodman, A. F. Martin, and R. E. Wohlford, "Improved automatic language identification in noisy speech," in *Proc. ICASSP '89,* vol. 1, May 1989, pp. 528–531.
[19] R. B. Ives, "A minimal rule AI expert system for real-time classification of natural spoken languages," in *Proc. Second Ann. Artificial Intell. Adv. Comput. Technol. Conf.,* Long Beach, CA, May 1986, pp. 337–340.
[20] M. Sugiyama, "Automatic language recognition using acoustic features," in *Proc. ICASSP '91,* vol. 2, May 1991, pp. 813–816.
[21] Y. K. Muthusamy and R. A. Cole, "Automatic segmentation and identification of ten languages using telephone speech," in *Proc. ICSLP '92,* vol. 2, Oct. 1992, pp. 1007–1010.
[22] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *J. Acoust. Soc. Amer.,* vol. 62, no. 3, pp. 708–713, Sept. 1977.
[23] M. Savic, E. Acosta, and S. K. Gupta, "An automatic language identification system," in *Proc. ICASSP '91,* vol. 2, May 1991, pp. 817–820.
[24] S. Nakagawa, T. Seino, and Y. Ueda, "Spoken language identification by ergodic HMMs and its state sequences," *Electron. Commun. Japan, Pt. 3,* vol. 77, no. 6, pp. 70–79, Feb. 1994.
[25] K. P. Li and T. J. Edwards, "Statistical models for automatic language identification," in *Proc. ICASSP '80,* vol. 3, Apr. 1980, pp. 884–887.
[26] K.-P. Li, "Automatic language identification using syllabic spectral features," in *Proc. ICASSP '94,* vol. 1, Apr. 1994, pp. 297–300.
[27] V. Fromkin and R. Rodman, *An Introduction to Language.* Orlando, FL: Harcourt Brace Jovanovich, 1993.
[28] L. F. Lamel and J.-L. Gauvain, "Cross-lingual experiments with phone recognition," in *Proc. ICASSP '93,* vol. 2, Apr. 1993, pp. 507–510.
[29] 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 *Proc. ICASSP '94,* vol. 1, Apr. 1994, pp. 121–124.
[30] K. M. Berkling, T. Arai, and E. Barnard, "Analysis of phoneme-based features for language identification," in *Proc. ICASSP '94,* vol. 1, Apr. 1994, pp. 289–292.
[31] L. F. Lamel and J. L. Gauvain, "Language identification using phone-based acoustic likelihoods," in *Proc. ICASSP '94,* vol. 1, Apr. 1994, pp. 293–296.
[32] S. Kadambe and J. L. Hieronymus, "Language identification with phonological and lexical models," in *Proc. ICASSP '95,* vol. 5, May 1995, pp. 3507–3510.
[33] R. J. D'Amore and C. P. Mah, "One-time complete indexing of text: Theory and practice," in *Proc. Eighth Int. ACM Conf. Res. Dev. Inform. Retrieval,* 1985, pp. 155–164.
[34] R. E. Kimbrell, "Searching for text? Send an N-gram!," *Byte,* vol. 13, no. 5, pp. 297–312, May 1988.
[35] J. C. Schmitt, "Trigram-based method of language identification," US Patent 5 062 143, Oct. 1991.
[36] M. Damashek, "Gauging similarity via N-grams: Language-independent text sorting, categorization, and retrieval of text," submitted for publication in *Sci.*
[37] T. A. Albina *et al.,* "A system for clustering spoken documents," in *Proc. Eurospeech '93,* vol. 2, Sept. 1993, pp. 1371–1374.
[38] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," in *Proc. ICASSP '95,* vol. 5, May 1995, pp. 3511–3514.
[39] T. J. Hazen and V. W. Zue, "Recent improvements in an approach to segment-based automatic language identification," in *Proc. ICASSP '94,* vol. 4, Sept. 1994, pp. 1883–1886.
[40] S. Hutchins, Private communication.
[41] S. Itahashi and L. Du, "Language identification based on speech fundamental frequency," in *Proc. ICASSP '95,* vol. 2, Sept. 1995, pp. 1359–1362.
[42] S. Mendoza, Private communication.
[43] P. Ramesh and D. B. Roe, "Language identification with embedded word models," in *Proc. ICASSP '94,* vol. 4, Sept. 1994, pp. 1887–1890.
[44] B. Comrie, *The World's Major Languages.* New York: Oxford University Press, 1990.
[45] D. Crystal, *The Cambridge Encyclopedia of Language.* Cambridge UK: Cambridge University Press, 1987.
[46] Y. K. Muthusamy, N. Jain, and R. A. Cole, "Perceptual benchmarks for automatic language identification," in *Proc. ICASSP '94,* vol. 1, Apr. 1994, pp. 333–336.
[47] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[48] D. B. Paul, "Speech recognition using hidden Markov models," *Lincoln Lab. J.,* vol. 3, no. 1, pp. 41–62, Spring 1990.
[49] D. A. Reynolds, R. C. Rose, and M. J. T. Smith, "PC-Based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system," in *Proc. ICSPAT '92,* vol. 2, Nov. 1992, pp. 967–973.
[50] H. Hermansky *et al.,* "RASTA-PLP speech analysis technique," in *Proc. ICASSP '92,* vol. 1, Mar. 1992, pp. 121–124.
[51] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.,* vol. COM-28, no. 1, pp. 84–95, Jan. 1980.
[52] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.,* vol. 39, pp. 1–38, 1977.
[53] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities,* vol. 3, pp. 1–8, 1972.
[54] P. C. Woodland and S. J. Young, "The HTK tied-state continuous speech recognizer," in *Proc. Eurospeech '93,* vol. 3, Sept. 1993, pp. 2207–2210.
[55] F. Jelinek, "Self-organized language modeling for speech recognition," in *Readings in Speech Recognition,* A. Waibel and K.-F. Lee, Eds. Palo Alto, CA: Morgan Kaufmann, 1990, pp. 450–506.
[56] M. A. Lund and H. Gish, "Two novel language model estimation techniques for statistical language identification," in *Proc. Eurospeech '95,* vol. 2, Sept. 1995, pp. 1363–1366.
[57] A. F. Martin, *Language ID Guidelines and Results.* Gaithersburg, MD: Nat. Inst. Std. Technol. (NIST), Spoken Natural Language Processing Group.
[58] C. R. Jankowski *et al.,* "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. ICASSP '90,* Apr. 1990, pp. 109–112.
[59] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP '92,* vol. 1, Mar. 1992, pp. 517–520.
[60] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 37, no. 11, pp. 1641–1648, Nov. 1989.
[61] H. Gish, Private communication.
[62] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.
[63] L. F. Lamel and J.-L. Gauvain, "High performance speaker-independent phone recognition using CDHMM," in *Proc. Eurospeech '93,* vol. 1, Sept. 1993, pp. 121–124.
[64] D. B. Paul, and B. F. Necioğlu, "The Lincoln large-vocabulary stack-decoder HMM CSR," in *Proc. ICASSP '93,* vol. 2, Apr. 1993, pp. 660–663.
[65] W. Mistretta, Private communication.
[66] D. M. Rekart and M. A. Zissman, "Dialect labels for the Spanish segment of the OGI multi-language telephone speech corpus," Sept. 1994, Mass. Inst. Technol., Lincoln Lab., Project Rep. DVPR-2, Lexington, MA, USA.

Marc A. Zissman (M'86) was born in Chicago, IL, USA, in 1963. He received the S.B. degree in computer science in 1985 and the S.B., S.M, and Ph.D. degrees in electrical engineering in 1986, 1986, and 1990, respectively, all from the Massachusetts Institute of Technology (MIT), Cambridge, USA.

From 1983 to the present, he has been a member of the Speech Systems Technology Group at MIT Lincoln Laboratory, where his research has focused on digital speech processing, including parallel computing for speech coding and recognition, co-channel talker interference suppression, language and dialect identification, and cochlear-implant processing for the profoundly deaf. From 1992 to the present, he has also been a research affiliate at the MIT Research Laboratory of Electronics.

Dr. Zissman is a member of Tau Beta Pi and Eta Kappa Nu.