Automatic Language Identification: A Review/Tutorial

Yeshwant K. Muthusamy†, Etienne Barnard‡ and Ronald A. Cole‡

† Systems and Information Sciences Laboratory
Texas Instruments, Inc.

‡ Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology

1 Introduction

1.1 The Problem

Automatic language identification (language ID for short) is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. As with speech recognition, humans are the most accurate language identification systems in the world today. Within seconds of hearing speech, people are able to determine whether it is a language they know. If it is a language with which they are not familiar, they often can make subjective judgments as to its similarity to a language they know, e.g., "sounds like German".

Languages have characteristic sound patterns; they are described subjectively as "singsong", "rhythmic", "guttural", "nasal" etc. Languages differ in the inventory of phonological units (speech sound categories) used to produce words, the frequency of occurrence of these units, and the order in which they occur in words. The presence of individual sounds, such as the "clicks" found in some sub-Saharan African languages, or the velar fricatives found in Arabic, are readily apparent to speakers of languages that do not contain these phonemes. Less obvious acoustic patterns are also observed. Mandarin Chinese has a higher frequency of occurrence of nasals than English. Hawaiian is known for its very limited consonant inventory. Prosodic patterns also differ significantly between languages. For example, it has been shown that fundamental frequency (F_0) patterns of continuous speech display different characteristics in Mandarin Chinese (a tone language) and American English (a stress language) [3]. The key to solving the problem of automatic language identification then, is the detection and exploitation of such differences between languages. Of course, if we had a system or set of systems that could "understand" each language, it would also be identifying the correct one in the process. However, speech recognition or understanding in multiple languages is still very much an unsolved problem.

1.2 Importance of language ID

There are several important applications for automatic language identification. As the global economic community expands, there is an increasing need for automatic spoken language identification services. For example, checking into a hotel, arranging a meeting or making travel arrangements can be difficult for non-native speakers. Telephone companies will be better equipped to handle foreign language calls if an automatic language identification system can be used to route the call to an operator fluent in that language.

Rapid language identification and translation can even save lives. There are many reported cases of 911 operators being unable to understand the language of a distressed caller. In response to these needs, AT&T recently introduced its *Language Line* Interpreter Service to serve business, the general public and police departments handling 911 emergencies. The service uses trained human interpreters, handles 140 languages and satisfies an important need in our increasingly cosmopolitan communities. However, tremendous responsibility is placed on the human operator who must route the call to the appropriate interpreter. A call to the *Language Line* Service by the first author, who spoke only in Tamil, resulted in a 3 minute delay before the language was identified and a Tamil interpreter was brought on-line. The delay was caused by the operator unsuccessfully trying out three South-East Asian interpreters and playing recordings of greetings in other languages. The delay would have been longer if the author had not relented and spoken the name 'Tamil' in English rather than in Tamil! This anecdote emphasizes the point that if automatic language identification could be made sufficiently fast and accurate, it could aid human operators.

An automatic language identification system could also serve as a front-end for a multi-language translation system in which the input speech can be in one of several languages. The input language needs to be quickly identified before translation to the target language(s) can begin.

2 Sources of Information Useful for Language ID

What makes this problem so challenging and interesting? In mono-lingual spoken language systems, the objective is to determine the content of the speech, typically implemented by phoneme recognition coupled with word recognition and sentence recognition. This requires that researchers cue in on small portions of the speech—frames, phonemes, syllables, sub-word units, and so on, to determine what the speaker said. In contrast, in text-independent language identification, phonemes and other sub-word units alone are not sufficient cues, since several phonemes and syllables and even words are common across different languages. One also needs to examine the sentence as a whole to determine the "acoustic signature" of the language, the unique characteristics that make one language sound distinct from another.

Decoding this "acoustic signature" requires information from several sources:

- **Acoustic Phonetics.** Phonetic inventories differ from language to language. Even when languages have identical phones, the frequencies of occurrence of phones differ across languages.

- **Prosodics.** Languages vary in terms of the duration of phones, speech rate and the intonation (pitch contour). Tonal languages (i.e. languages in which the intonation of a word determines its meaning) such as Mandarin and Vietnamese have very different intonation characteristics than stress languages such as English.

- **Phonotactics.** Phonotactics refers to the rules that govern the combinations of the different phones in a language. There is a wide variance in phonotactic rules across languages. For example, the phone cluster /sr/ is very common in the Dravidian language Tamil, whereas it is not a legal cluster in English.

- **Vocabulary.** Conceptually the most important difference between languages is that they use different sets of words – that is, their vocabularies differ. Thus, a non-native speaker of English is likely to use the phonemic inventory, prosodic patterns and even (approximately) the phonotactics of her/his native language, but will be judged to speak English if the vocabulary used is that of English.

A successful language ID algorithm would exploit information from all of the above sources to arrive at its identification decision.

3 Previous Approaches to Language ID

We were able to locate only fourteen studies in language ID published in English in the two decades preceding the recent developments in language ID described in Section 6, The speech data have spanned the range from phonetic transcriptions of text, laboratory-quality speech, to telephone and radio speech. The number of languages has varied from three to twenty. The approaches to language identification have used "reference sounds" in each language [16, 17, 18, 19], segment- and syllable-based Markov models [21], pitch contours [5, 31], formant vectors [5, 6], acoustic features [2], broad phonetic and prosodic features [25], and just raw waveform features [12]. A variety of classification methods have been tried, including HMMs [9, 21], expert systems [11], clustering algorithms [5, 17, 32], quadratic classifiers [5] and artificial neural networks [12]. A detailed review of these studies can be found in [22].

The literature does not present a coherent picture. While the performance figures of some of the studies might look impressive in isolation, meaningful comparisons across studies is not possible, for the following reasons:

- Many of the studies represented classified or sensitive research, so experimental details (e.g., languages used) were often not described[5, 6, 16, 17, 18, 19, 21].
- There was no common, public-domain database (like the TIMIT corpus [4, 13] for continuous speech recognition) with which to evaluate different approaches to language ID.

Thus, despite initiating some interesting ideas (the use of VQ, pitch contours, and formant vectors), these studies have had little impact on current work.

4 OGI Multi-language Telephone Speech Corpus

4.1 Motivation

Research in automatic language identification requires a large corpus of multi-lingual speech data to capture the many sources of variability within and across languages. These include variability due to speaker differences (e.g., age, gender, dialect), microphones, telephone handsets, communication lines, background noise and the language being spoken. It is also important that the corpus contain a wide variety of speech from each speaker, ranging from fixed-vocabulary utterances to natural, continuous speech. This makes it useful for both content-dependent and content-independent language identification. Further, the availability of such a corpus in the public-domain would enable researchers to study languages and to develop, evaluate and compare multi-language recognition algorithms.

From one perspective, it would be ideal for such a corpus to consist of high-quality laboratory-recorded speech, since this would enable researchers to concentrate on the core issues of language identification to the exclusion of extraneous complicating factors. Several practical concerns have, however, lead to the adoption of corpora recorded over the telephone channel as the common standard, despite the introduction of additional complications. In the real world, an automatic language identification system is more likely to be used over some form of communication channel. If the system is to perform accurately under these conditions, it needs to be trained on speech recorded under these conditions. Moreover, there are definite advantages of telephone speech data collection over that of high-quality (laboratory) speech.

- The collection process can be easily automated. Once the recording protocol and equipment are set up, speech data can be collected very rapidly with minimum human supervision.
- Long-distance telephone networks provide access to speakers of different languages spread over a wide geographical area.

4.2 OGI_TS

The OGI Multi-language Telephone Speech Corpus (OGI_TS for short) was designed specifically for language ID research. It currently consists of spontaneous and fixed-vocabulary utterances in 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. These utterances were produced by 90 native speakers in each language over real-world telephone lines. The utterances ranged in duration from 1 second to 50 seconds, with an average duration of 13.4 seconds. Hindi is a recent addition to the corpus. The design, collection and development of the original ten-language corpus is described in detail in [22, 27]. OGI_TS has been placed in the public domain[1].

4.3 Impact of OGI_TS: NIST Evaluations

The advent of OGI_TS has sparked renewed interest in language ID. In March 1993, it was designated as the standard for evaluating language ID algorithms by the National Institute of Standards and Technology (NIST). NIST has since been coordinating the evaluation process. At last count, eight research sites in the U.S. (AT&T, ITT, Lockheed-Sanders, MIT, MIT Lincoln Laboratories, Natural Speech Technologies, OGI, and RPI) are participating in this ongoing evaluation. In addition, the last two years have seen a substantial increase in papers on language ID in major speech conferences and symposia such as ICASSP, Eurospeech and SRS (Speech Research Symposium) [7, 15, 20, 23, 24, 28, 33, 34], with complete sessions devoted to language ID in each of them. This proliferation of different approaches to the problem using the same corpus has led to an open exchange of ideas—a process so essential for research progress.

5 Perceptual Studies on Language ID

5.1 Perceptual Benchmarks are Essential

Humans are able to identify languages using very short excerpts of speech, drawing upon several different sources of information. The fact that humans are so adept at this task illustrates the considerable gap between our perceptual capabilities and our attempts at automating them. If we are to shorten this gap, it is imperative that we study human performance on language identification tasks. Perceptual studies with listeners from different language backgrounds provide benchmarks for evaluating machine performance. In addition, patterns of confusions between languages provide insights about the salient acoustic and other characteristics that can be useful for language ID.

[1] It can be obtained from the Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

5.2 Perceptual Experiments

Before the advent of OGI_TS, there were only two attempts to study human performance on language identification [1, 32]. These studies, performed on different corpora, were very limited in their scope and in the number of speakers used.

A series of perceptual experiments were conducted using 1-, 2-, 4- and 6-second excerpts of speech excised from the spontaneous speech utterances in the original ten-language OGI corpus [28]. These experiments used an interactive graphical interface that played out excerpts of speech at random from the 10 languages, and maintained a log of listener responses. The listeners were given feedback on every trial so that they were constantly being trained as the experiment progressed. They could also listen to an utterance *after* making a choice—a feature that was included to aid in the learning process.

The first set of experiments examined listening performance of monolingual English speakers. The second set of experiments used twice as many excerpts of speech and listeners who were native speakers of the ten languages. The listeners in the second series of experiments were debriefed after the experiment to determine the cues that they used or developed during the course of the experiment to distinguish between the languages.

Towards the end of the experiment, listeners were able to identify the ten languages with accuracies ranging from 39.2% to 100.0% (average performance: 69.4%) using just 6-second excerpts of speech. Figure 1 displays the average performance, over all listeners, on the 6-second excerpts in the first and last quarters of the second experiment.

The responses in the post-experiment interviews provided valuable information on the way human listeners learn to discriminate between languages. They appear to have used a combination of phoneme- and word-spotting strategies (e.g. the velar fricative /kh/ in German as in *ich*, the phoneme pair *eh-s* in Spanish; the words *imnida* in Korean and *mashita* in Japanese) and prosodic cues (e.g. sing-song intonation of Vietnamese). Non-native speakers of Korean had a lot of trouble identifying Korean. Many of them confessed to choosing Korean when they were unsure of the language of the excerpt!

The second experiment showed that increased exposure to each language and longer training sessions contribute to improved classification performance. Listeners who knew more languages tended to perform better, on the average, than subjects who knew just one language. While *a priori* knowledge of the language definitely helped, the listeners seem to have learned to develop their own cues as the experiment progressed.

While these experiments have provided interesting results, the mix of subjects and multitude of languages makes it difficult to determine the cues that human listeners would use to distinguish between two *completely unfamiliar* languages. Perceptual experiments using just pairs of languages and appropriately selected subjects might provide some answers to that intriguing question.

Another important issue not addressed by these experiments is the performance of listeners who have access to linguistic information beyond the acoustic samples provided during training. How, for instance, would subjects trained in linguistic phonetics fare, or subjects who are provided with prior information on the salient differences between the languages? Many such interesting questions remain to be addressed.

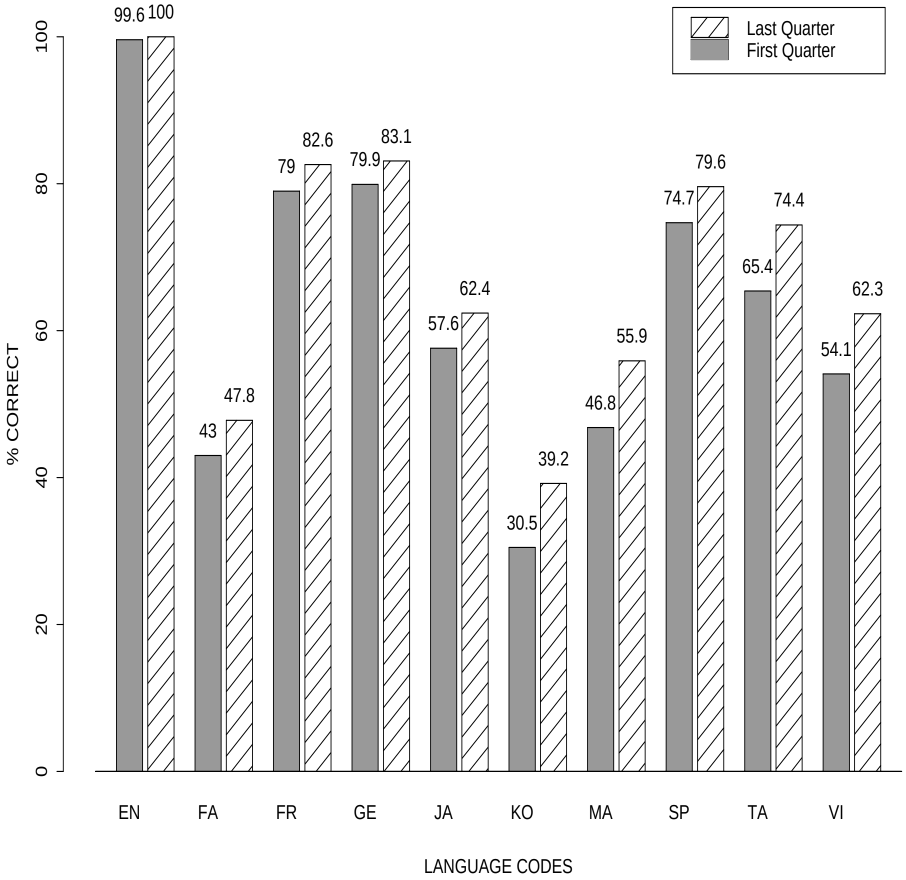


Figure 1: Average Subject Performance on 6-second Excerpts in the First and Last Quarters: All Subjects

6 Recent Approaches to Language ID

Language identification is related to speaker-independent speech recognition and speaker identification in several interesting ways. It is therefore not surprising that many of the recent developments in language identification can be related to developments in those two fields. In this section we review some of the more important recent approaches to language identification against the background of successes in speaker and speech recognition.

In particular, we demonstrate how approaches to language identification based on acoustic modeling and language modeling, respectively, are similar to algorithms used in speaker-independent continuous speech recognition. Thereafter, prosodic and duration-based information sources are studied. We then review an approach to language identification that draws heavily on speaker identification. Finally, the performance of some representative algorithms is reported.

6.1 Stochastic Models of Language Acoustics

In general, languages differ significantly from each other with respect to their typical short-term acoustics. This is not only caused by differences in phonemic inventories employed in the different languages (as noted in Section 1), but also by subtle differences in the realization of similar phonemes in those languages. Thus, the fricative in the German "ich" has no English counterpart – all English fricatives are formed at the lips, teeth, or glottis. In contrast, the "r" in American English differs somewhat from both its Spanish counterparts (one of which is realized as a flap, and the other as a trill).

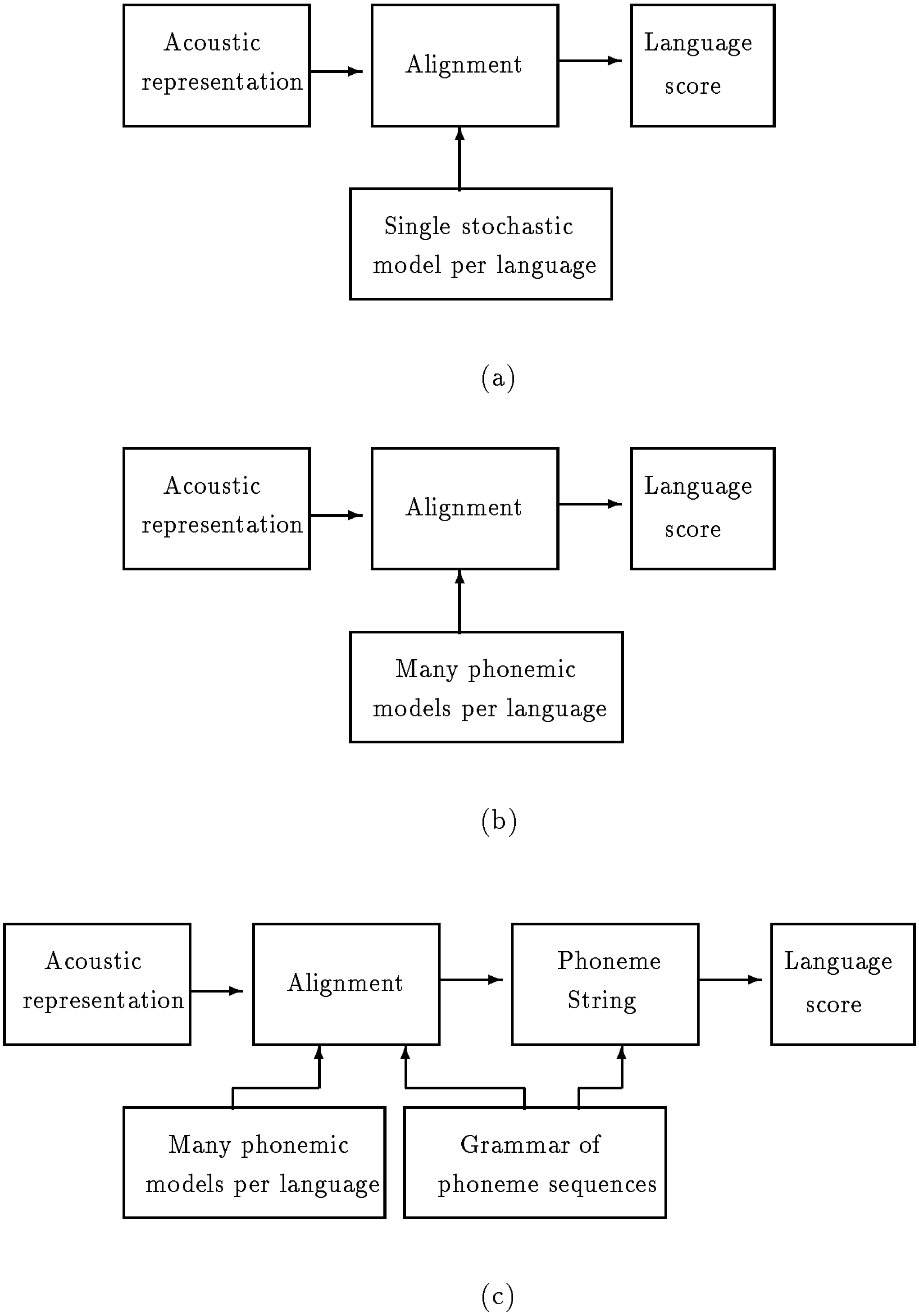
During the past decade, much progress has been made in speaker-independent speech recognition by using sophisticated methods such as Hidden Markov Models (HMMs) and artificial neural networks (NNs) to model short-term acoustics. The most successful systems have used HMMs or NNs to model the acoustics of speech units such as (context-dependent) phonemes. These models have proven to be sufficiently robust with respect to factors like speaker differences and contextual variations for successful speech recognition to be possible.

These advances have carried over into language ID in a variety of forms. One approach is to model an entire language by a single stochastic model, as shown schematically in Figure 2(a). For instance, an ergodic HMM (i.e. an HMM with all states connected to all other states —see Figure 3(a)) is trained for every language to be recognized, using mel-scaled cepstral coefficients as input [33, 29]. To identify the language of an unknown utterance, it is decoded with each of these models in turn. The language of the model which thus predicts the utterance with the highest likelihood is taken as the language of the utterance.

This approach has not been very successful; it is obvious that it is ambitious to hope that a single HMM can be trained to capture all of the complexities of a language. Researchers have therefore moved to systems which are even more akin to those used for speech recognition. A separate stochastic model is trained for each phoneme in each of the target languages (see Figure 2(b)). Since we now try to model the temporal structure of these phonemes explicitly, a left-to-right topology for the Markov models as in Figure 3(b) is appropriate.

An unknown utterance is then decoded using the set of phoneme models from each of the languages in turn; the language with the largest likelihood is again selected. Note that the decoding process produces an optimal string of phonemes in each of the languages as a side effect. This fact is used in language modeling, as is described below.

It is more realistic to learn representative phoneme models than models of a whole language, and experience has shown that these approaches do indeed outperform those relying on a single stochastic model per language. The main disadvantage of phonemic approaches is that they



(a)

(b)

(c)

Figure 2: Three System Structures for Language ID

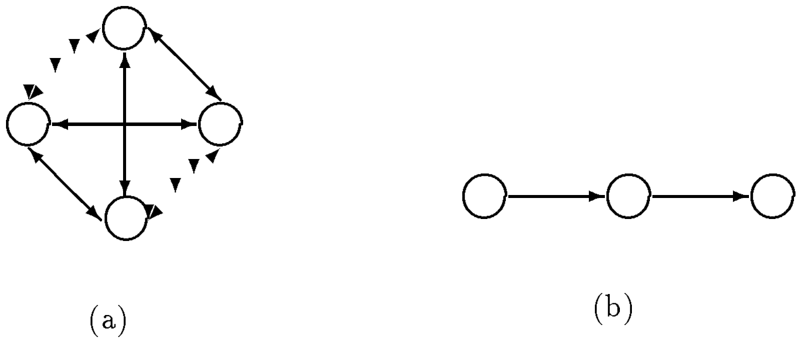


Figure 3: HMM topologies: (a) Ergodic and (b) Left-to-right

require phonemically labeled data in each of the target languages for use during training of the models (or at least to seed such training).

6.2 Language Modeling for Language ID

In speech-recognition systems for continuous speech, the incorporation of stochastic grammars has been very important to the success of the most powerful systems. These grammars take into account the likelihoods that certain words will appear together, and thereby correct many of the errors that the unconstrained word recognizer would otherwise have made.

For text-independent language recognition, it is generally not feasible to construct word models in each of the target languages (it would, for instance, be very difficult to obtain dictionaries with sufficient coverage in each of the 11 languages in OGI_TS). It is possible, though, to create models which model the sequential statistics of more basic units in each of the languages, e.g., the phonemes or broad categories of phonemes. (In language ID, speech is often segmented by broad category, e.g. vowel, fricative, nasal, etc., rather than phonetic category. The main reason for this is that these broad categories are more language-independent, and classifiers can thus be used even for languages for which labeled training data are not available.) If a stochastic grammar is now used to compute likelihoods of co-occurrence of these units, we will capture some of the so-called *phonotactic* regularities in the target languages. In English, for instance, the phoneme /w/ is quite likely to follow after /s/, but will rarely follow after /r/.

Training now consists of two stages. First, a stochastic model is trained for each of the fundamental units, exactly as in Section 6.1. Various combinations of models and units have been used: HMMs with broad-phonetic classes [33] or phone classes [14, 15, 33, 34], and neural networks with phone classes [23, 24]. These trained models are then used to estimate the stochastic grammar appropriate for each language. Grammars equivalent to a bigram grammar have generally been employed; that is, these models capture the likelihood that each phoneme is followed by any other phoneme. The language of an utterance is then determined by successively decoding it with the unit models and grammar of each of the target languages. The decoding with the highest likelihood is taken to indicate the language in which the utterance was spoken. Since the likelihood computed during the decoding process is a product of both acoustic and grammatic terms, this score actually incorporates both acoustic and phonotactic information. This procedure is depicted abstractly in Fig. 2(c).

An important simplification is made possible by noting that one can build a stochastic grammar for one language based on the acoustic models of a *different* language. Thus, one can build acoustic models for just one language (English, say, for which data are readily available), and

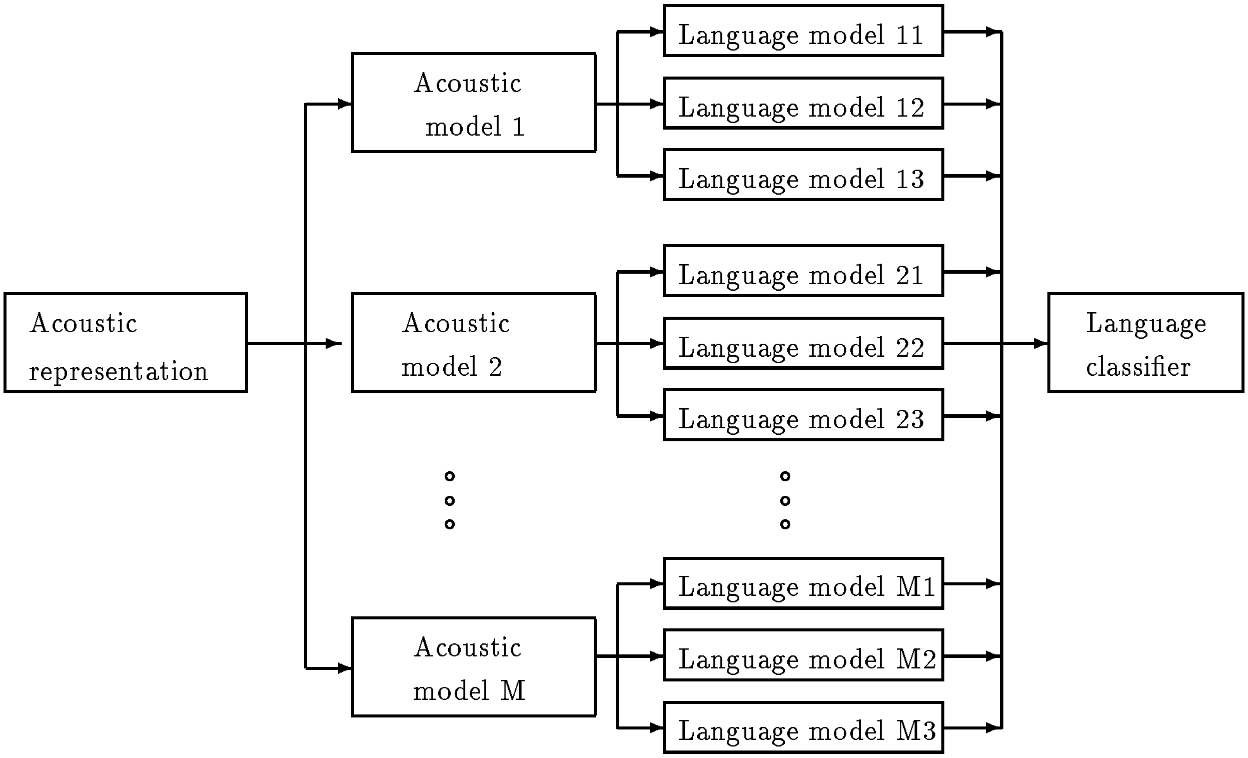


Figure 4: Language ID using M phonetic front-ends; a separate language model for each target language is based on the output of each front-end

calculate language models for any other target language based on the strings of phonemes produced by the English recognizer when utterances from the target language are decoded by it. This eliminates the need for training corpora in each of the target languages. One foregoes, however, the acoustic scores that were included in the discrimination before.

The most successful language ID system on the OGI_TS corpus to date [34] generalizes this idea somewhat: rather than using a single acoustic recognizer and N language models for N target languages, it uses as many acoustic recognizers as possible ($M = 6$ languages in OGI_TS have been labeled phonemically to date), and computes N language models for each of these. The likelihood for a particular language is taken as the sum of the likelihoods produced by each of the M language models constructed for it. This structure is depicted in Fig. 4.

6.3 Incorporating Duration and Prosodic Information

It has long been recognized that prosodic information (that is, information derived from speech characteristics such as pitch, amplitude, and rate, which span several phonemes) should contribute much to speech recognition. This insight has, however, not contributed much to the success of current systems. Similarly, the incorporation of explicit prosodic information was not as useful in early language-identification systems of the current generation[7, 22] as the designers may have hoped. There is, nonetheless, much reason to think that prosodic differences will contribute significantly to language identification in the future, and recent research has begun to fulfill this promise[10].

Hazen and Zue [7] incorporated pitch information by multiplying their acoustic and phonotactic probabilities by a third factor which captured the probability densities of pitch distributions in the various languages. (The first derivative of the pitch is treated similarly.) They have reported systems with both broad-category and phonetic acoustic models. Hazen and Zue find that the prosodic factor is useful, but not by much, in both cases.

Muthusamy [22] considered more complex prosodic models, which take into account the pitch variation within and across the different segments marked by a broad-category classifier. He also extracted features indicative of speech rate and syllabic timing. Again, these prosodic features were found to be marginally useful—much less so than the durational and phonotactic features that he also employed.

Segmental duration has been much more useful in characterizing language differences. In [7] and [26], the distributions of the durations in each broad category were modeled as additional factors in the computation of language likelihoods, and this was seen to be quite useful in both cases.

Perhaps the most successful work on incorporating prosodic information [10] proceeds by first segmenting an utterance into syllables based on amplitude and pitch information. Various statistics based on the rhythmic and tonal characteristics are then computed. Information related to rhythm is encapsulated in terms of syllable timing and duration, and descriptors of amplitude patterns. Tone information is described in terms of phrase characteristics of the pitch (such as the range of pitch levels and the variation of pitch over a whole utterance) and syllable characteristics thereof (including descriptors of the shape of the pitch trajectory in each syllable, and measures of how strongly the pitch falls off at the end of each syllable). Finally, the correlation between pitch and amplitude is also described by several measures. Using just these prosodic features, Hutchins and Thymé-Gobbel have obtained results comparable to those reported by other groups using many additional sources of information.

6.4 Language-based Speaker Similarity

Li [20] has achieved much success by importing ideas from speaker identification into language identification. His basic concept is to classify an incoming utterance by measuring the similarity between the speaker thereof and the most similar speakers of each of the target languages. (Although many of the systems described above bear some similarity to algorithms used in speaker identification, Li's work is unique to our knowledge in explicitly computing speaker similarities as a precursor to language identification.)

During training, an artificial neural network is used to extract all syllabic nuclei for all utterances in the training corpus. Spectral coefficients are extracted at several locations within each nucleus, and stored. During recognition the syllabic nuclei are similarly extracted, and the spectral coefficients are compared to all those stored for each speaker. The smallest difference between each nucleus in the utterance to be classified and the stored nuclei of each speaker is computed. The sum of these differences is taken as the difference between the speaker of the utterance and each of the reference speakers. The average difference with the S most similar speakers in each language is taken as the difference between the new utterance and the various target languages; the language with the smallest difference is selected. S is a parameter optimized for performance; it varies from 1 to 25 in the experiments reported in [20]. This process is schematized in Fig. 5.

6.5 Performance Results

Although all of these systems have been tested on the OGI_TS corpus, not all groups have performed exactly the same tests. It is therefore not possible to provide direct comparisons between

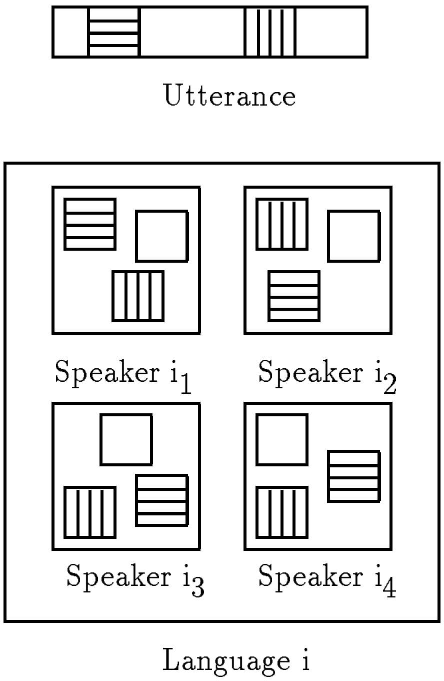


Figure 5: Language ID by speaker identification – the syllabic nuclei in an utterance (indicated by the shaded regions) are matched with the most similar syllabic nucleus of each training speaker. The "distance" between the utterance and each of the speakers is determined as the sum of the errors in these matches. The "distance" of a language is the average of the S smallest speaker distances

Method	Whole	10-sec	Desc
Speaker-identification based	78	59	[20]
Acoustic model per language	53	50	[33]
Phonotactics per broad category; pitch; timing; duration	66	48	[26]
Acoustics per broad category; phonotactics; pitch; duration	57	46	[7]

Table 1: Percentage of test utterances correctly classified in 10-language-recognition task with various approaches developed before 1994. "Whole" and "10-sec" refer to results obtained with whole stories and 10-second segments, respectively. "Desc" cites the papers in which the systems are described

Method	Whole	10-sec	Desc
Speaker-identification based	78	63	[20]
Acoustic model per phone; phonotactics	79	70	[34]
Acoustics per phone; phonotactics; pitch; duration	69	64	[8]

Table 2: Results for 11-language recognition with various more recent approaches

all systems. Also, it has been found that factors not explicitly mentioned above can have a substantial impact on performance – factors such as noise compensation, channel equalization, etc. Since the various systems differ in this regard, a comparison is further complicated. It is nevertheless possible to obtain at least some impression of the utility of different approaches by looking at comparable results that have been published.

The first set of results refer to the original OGI corpus, which consisted of data in 10 languages [27, 30]. Results are given for the "stories" part of the corpus, consisting mostly of extemporaneous speech, which is widely used as a test set. In Table 1, we list the results obtained with various methods when classifying the whole stories, and also for the classification of 10-second segments excised from these stories. Table 2 lists results obtained on the 11-language OGI_TS.

The following conclusions can be drawn from these results:

- The results reported for systems which have been employed on both pairs of tests show that the 10-language and 11-language tasks are of comparable difficulty: the complexity of an additional language introduced into the newer corpus is approximately compensated for by the additional training material available.

- In comparing systems that use phonetic rather than broad-category information, it seems that the former perform better. This has been confirmed in experiments more directly aimed at comparing phonetic and broad-category based approaches [24, 35]. Thus, although front-end classifiers are less successful in extracting detailed phonetic information than they are with broad-category information, the additional information contained in the phonetic categories more than compensates for the performance differential.

- A single acoustic model per language does not seem to be sufficient for capturing language differences; much better results are obtained by dividing the acoustic-modeling task into phonetic subtasks.

- There is, as of yet, no clear "preferred approach" to language ID. Systems designed from quite dissimilar principles (e.g. based on phonotactics [34] versus speaker similarity[20]) perform comparably on the task described here.

7 Conclusions

After languishing for almost two decades, the field of language ID has been rejuvenated with the advent of a public-domain multilingual corpus of speech. It has sparked renewed interest in the field and a proliferation of different approaches to the problem. Meaningful comparisons between systems can now be made and ideas on improving performance exchanged. This free-flow of information within the research community is an essential prerequisite for progress towards solving this problem.

Although there has been encouraging progress in language ID in the past two years, current systems are still not very reliable in distinguishing between languages in a set of 10 or 11 languages. It is informative to note that all the reported systems still perform much better when classifying 50-second utterances than 10-second segments. This is to be contrasted with perceptual experiments in which human identification performance asymptotes for much shorter durations of speech. Perhaps a more detailed examination of the patterns of confusion between the languages, obtained from such perceptual experiments, is required. In addition, analysis of listener feedback from post-experiment interviews could provide new insights into the sources of information that are not being captured well enough in current systems.

References

[1] K. Atkinson. Language identification from nonsegmental cues. *Journal of the Acoustical Society of America*, 44:378(A), 1968.

[2] D. Cimarusti and R. B. Ives. Development of an automatic identification system of spoken languages: Phase 1. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 82*, Paris, France, May 1982.

[3] S. J. Eady. Differences in the F_0 patterns of speech: Tone language versus stress language. *Language and Speech*, 25(1):29–42, 1982.

[4] W. Fisher, G. R. Doddington, and K. Goudie-Marshall. The DARPA speech recognition research database: Specification and status. In *Proceedings DARPA Speech Recognition Workshop*, pages 93–100, February 1986.

[5] J. T. Foil. Language identification using noisy speech. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 86*, Tokyo, Japan, 1986.

[6] F.J. Goodman, A.F. Martin, and R.E. Wohlford. Improved automatic language identification in noisy speech. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 89*, Glasgow, Scotland, May 1989.

[7] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech 93)*, September 1993.

[8] T. J. Hazen and V. W. Zue. Recent improvements in an approach to segment-based automatic language identification. In *Proceedings International Conference on Spoken Language Processing 94*, Yokohama, Japan, September 1994.

[9] A. S. House and E. P. Neuberg. Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3):708–713, 1977.

[10] S. E. Hutchins and A. Thymé-Gobbel. Experiments using prosody for language identification. In *Proceedings Speech Research Symposium XIV*, Baltimore, Maryland, June 1994.

[11] R. B. Ives. A minimal rule AI expert system for real-time classification of natural spoken languages. In *Proceedings 2nd Annual Artificial Intelligence and Advanced Computer Technology Conference*, Long Beach, CA, April-May 1986.

[12] S. C. Kwasny, B. L. Kalman, W. Wu, and A. M. Engebretson. Identifying language from speech: An example of high-level, statistically-based feature extraction. In *Proceedings 14th Annual Conference of the Cognitive Science Society*, 1992.

[13] L. Lamel, R. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings DARPA Speech Recognition Workshop*, pages 100–110, February 1986.

[14] L. F. Lamel and J-L. S. Gauvain. Language identification using phone-based acoustic likelihoods. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 94*, pages I–293–I–296, Adelaide, Australia, April 1994.

[15] L. F. Lamel and J-L. S. Gauvain. Identifying non-linguistic speech features. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech 93)*, pages 23–30, Berlin, Germany, September 1993.

[16] R. G. Leonard. Language recognition test and evaluation. Technical Report RADC-TR-80-83, Air Force Rome Air Development Center, March 1980.

[17] R. G. Leonard and G. R. Doddington. Automatic language identification. Technical Report RADC-TR-74-200, Air Force Rome Air Development Center, August 1974.

[18] R. G. Leonard and G. R. Doddington. Automatic language discrimination. Technical Report RADC-TR-78-5, Air Force Rome Air Development Center, January 1978.

[19] R. G. Leonard and G. R. Doddington. Automatic language identification. Technical Report RADC-TR-75-264, Air Force Rome Air Development Center, October 1975.

[20] K. P. Li. Automatic language identification using syllabic features. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 94*, pages I–297–I–300, Adelaide, Australia, April 1994.

[21] K.P. Li and T. J. Edwards. Statistical models for automatic language identification. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 80*, Denver, CO, April 1980.

[22] Y. K. Muthusamy. *A Segmental Approach to Automatic Language Identification.* PhD thesis, Oregon Graduate Institute of Science & Technology, 1993.

[23] Y. K. Muthusamy, T. Arai, K. M. Berkling, R. A. Cole, and E. Barnard. Two approaches to automatic language identification with telephone speech. In *Speech Research Symposium XIII*, pages 443–449, Baltimore, Maryland, June 1993.

[24] Y. K. Muthusamy, K. M. Berkling, T. Arai, R. A. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech 93)*, Berlin, Germany, September 1993.

[25] Y. K. Muthusamy and R. A. Cole. A segment-based automatic language identification system. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[26] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[27] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[28] Y. K. Muthusamy, Neena Jain, and Ronald A. Cole. Perceptual benchmarks for automatic language identification. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 94*, Adelaide, Australia, April 1994.

[29] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[30] D. S. Pallett and A. F. Martin. Language identification: Testing protocols and evaluation procedures. In *Speech Research Symposium XIII*, pages 428–442, Baltimore, Maryland, June 1993.

[31] M. Savic, E. Acosta, and S. K. Gupta. An automatic language identification system. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 91*, Toronto, Canada, May 1991.

[32] M. Sugiyama. Automatic language recognition using acoustic features. Technical Report TR-I-0167, ATR Interpreting Telephony Research Laboratories, 1991.

[33] M. A. Zissman. Automatic language identification using gaussian mixture and hidden markov models. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing 93*, Minneapolis, MN, April 1993.

[34] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 94*, pages I–305–I–308, Adelaide, Australia, April 1994.

[35] M. A. Zissman and E. Singer. Language identification using phonetic class recognition and N-gram analysis. In *Speech Research Symposium XIII*, pages 400–409, Baltimore, Maryland, June 1993.