ELSEVIER

Speech Communication 35 (2001) 115–124

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic language identification ☆

Marc A. Zissman, Kay M. Berkling *

MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420-9185, USA

Abstract

Automatic language identification of speech is the process by which the language of a digitized speech utterance is recognized by a computer. In this paper, we will describe the set of available cues for language identification of speech and discuss the different approaches to building working systems. This overview includes a range of historical approaches, contemporary systems that have been evaluated on standard databases, and promising future approaches. Comparative results are also reported. © 2001 Elsevier Science B.V. All rights reserved.

1. Introduction

Automatic language identification of speech is the process by which the language of a digitized speech utterance is recognized by a computer. It is one of several processes in which information is extracted automatically from a speech signal. Language identification can also be performed on other types of data (e.g. text), but this paper focuses narrowly on speech applications.

Language-ID (LID) applications fall into two main categories: pre-processing for machine systems and pre-processing for human listeners. Consider a hotel lobby or international airport of the future that employs a multi-lingual voice-controlled travel-information retrieval system. If no mode of input other than speech is used, then the system must be capable of determining the language of the speech commands either while it is recognizing the commands or before it has recognized the commands. Determining the language during recognition would require many speech recognizers (one for each language) running in parallel. Because tens or even hundreds of input languages would need to be supported, the cost of the required real-time hardware might prove prohibitive. Alternatively, an LID system could be run in advance of the speech recognizer. In this case, the LID system would quickly list 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 LID determination would be made only after speech recognition was complete.

An example of the second category of LID applications is preprocessing for human listeners. In this case, LID is 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 a *Language Line* interpreter service to, among others, police departments handling

☆ This work was sponsored by the Department of Defense under Air Force Contract F19628-95-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the US Air Force.

* Corresponding author. Present address: Buckhauser Strasse 3, 8048 Zurich, Switzerland. Tel.: +41-1400-5650.

E-mail address: maz@ll.mit.edu (K.M. Berkling).

0167-6393/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S 0 1 6 7 - 6 3 9 3 (0 0) 0 0 0 9 9 - 6

emergency calls. When a caller to Language Line does not speak 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 can be used) and can require connections to several human interpreters before the appropriate person is found. As reported by Muthusamy et al. (1994a), when callers to *Language Line* do not speak English, the delay in finding a suitable interpreter can be of the order of minutes, which could prove devastating in an emergency. Thus, an LID system that could quickly determine the most likely languages of the incoming speech might be used to reduce the time required to find an appropriate interpreter by one or two orders of magnitude.

2. Language identification cues

Humans and machines can use a variety of cues to distinguish one language from another. The reader is referred to the linguistics literature (e.g. Comrie, 1990; Crystal, 1987; Fromkin and Rodman, 1993) for in-depth discussions of how specific languages differ from one another and to Muthusamy et al. (1994b), who has measured how well humans can perform LID. In summary, the following characteristics differ from language to language:

- *Phonology*. A "phoneme" is an underlying mental representation of a phonological unit in a language. For example, the eight phonemes that comprise the word "celebrate" are `/s eh l ix b r ey t/`. A "phone" is a realization of an acoustic–phonetic unit or segment. It is the actual sound produced when a speaker is thinking of speaking a phoneme. The phones that comprise the word celebrate might be `[s eh l ax bcl b r ey q]`. Phone and phoneme sets differ from one language to another, even though many languages share a common subset of phones/phonemes. Phone and phoneme frequencies of occurrence 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 and phonemes, can also be different.
- *Morphology*. The word roots and lexicons are usually different from language to language. Each language has its own vocabulary, and its own manner of forming words.
- *Syntax*. The sentence patterns are different among languages. 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 characteristics, pitch contours, and stress patterns are different from one language to another.

3. LID systems

Research in automatic language identification from speech has a history extending back to the 1970s. A few representative LID systems are described below. The reader will find references to other LID systems in reviews by Muthusamy et al. (1994a) and Zissman (1996).

Fig. 1 shows the two phases of LID. During the "training" phase, the typical system is presented with examples of speech from a variety of languages. Each training speech utterance is converted into a stream of feature vectors. These feature vectors are computed from short windows of the speech waveform (e.g. 20 ms) during which the speech signal is assumed to be somewhat stationary. The feature vectors are recomputed regularly (e.g. every 10 ms) and contain spectral or cepstral information about the speech signal (the cepstrum is the inverse Fourier transform of the log magnitude spectrum; it is used in many speech-processing applications). The training algorithm analyzes a sequence of such vectors and produces one or more models for each language. These models represent a set of language-dependent, fundamental characteristics of the training speech to be used during the next phase of the LID process.

During the "recognition" phase of LID, feature vectors computed from a new utterance are compared to each of the language-dependent models. The likelihood that the new utterance was spoken

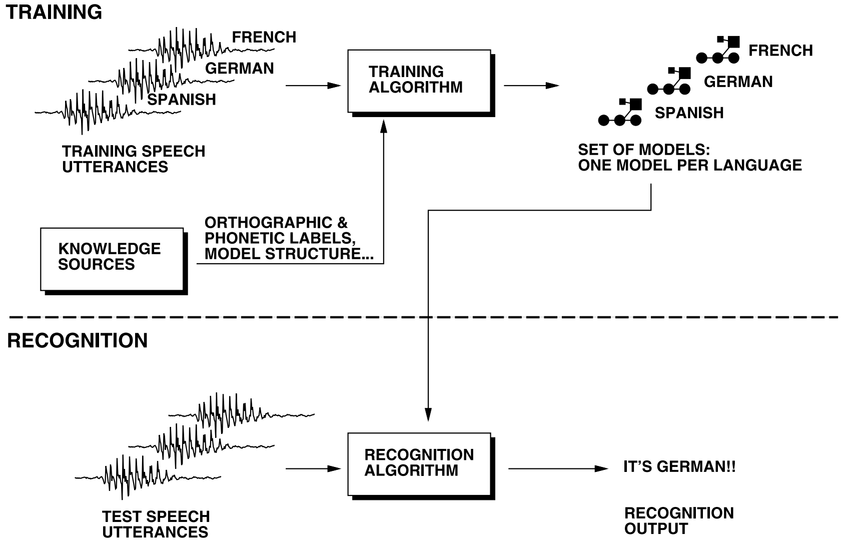

Fig. 1. The two phases of language identification. During training, speech waveforms are analyzed and language-dependent models are produced. During recognition, a new speech utterance is processed and compared to the models produced during training. The language of the speech utterance is hypothesized.

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 model yielding maximum likelihood is hypothesized as the language of the utterance.

Systems vary primarily according to their method for modeling languages. We will discuss a series of different features that have been extracted from speech, yielding increasing amounts of knowledge at the cost of rendering the language identifications system more and more complex. During training, some systems require only the digitized speech utterances and the corresponding true identities of the languages being spoken because the language models are based simply on the signal representation or on self-generated token representation. More complicated LID systems use phonemes to model speech and may require either (1) a phonetic transcription (sequence of symbols representing the spoken sounds), or (2) an orthographic transcription (the text of the words spoken) along with a phonemic transcription dictionary (mapping of words to prototypical pronunciation) for each training utterance. Producing these transcriptions and dictionaries is an expensive, time-consuming process that usually requires a skilled linguist fluent in the language of interest.

3.1. Spectral-similarity approaches

In the earliest automatic LID systems, developers capitalized on the differences in spectral content among languages, exploiting the fact that speech spoken in different languages contains different phonemes and phones. To train these systems, a set of prototypical short-term spectra were computed and extracted from training speech utterances. During recognition, test speech spectra were computed and compared to the training prototypes. The language of the test speech was hypothesized as the language having training spectra that best matched the test spectra.

There were several variations on this spectral-similarity theme. The training and testing spectra could be used directly as feature vectors, or they could be used instead to compute formant-based or cepstral features vectors. The training exemplars could be chosen either directly from the training speech or could be synthesized through the use of K-means clustering. The spectral-similarity could be calculated by the Euclidean, Mahalanobis, or some other distance metric. Examples of spectral-similarity LID systems are those proposed and developed by Cimarusti and Ives (1982), Foil (1986), Goodman et al. (1989) and Sugiyama (1991).

To compute the similarity between a test utterance and a training model, most of the early spectral-similarity systems calculated the distance between each test utterance vector and each training exemplar. The distance between each test vector and its closest exemplar was accumulated as an overall distance, and the language model having lowest overall distance was found. In a generalization of this vector-quantization approach to LID, Riek et al. (1991), Nakagawa et al. (1992) and Zissman (1993) applied Gaussian mixture classifiers to language identification. Here, each feature vector is assumed to be drawn randomly according to a probability density that is a weighted sum of multi-variate Gaussian densities. During training, a Gaussian mixture model for the spectral or cepstral feature vectors is created for each language. During recognition, the likelihood of the test utterance feature vectors is computed given each of the training models. The language of the model having maximum likelihood is hypothesized. The Gaussian mixture approach is "soft" vector-quantization, where more than one exemplar created during training impacts the scoring of each test vector.

Whereas the language identification systems described above perform primarily static classification, hidden Markov models (HMMs) (Rabiner, 1989), which have the ability to model sequential characteristics of speech production, have also been applied to LID. HMM-based language identification was first proposed by House and Neuburg (1977). Savic et al. (1991), Riek et al. (1991), Nakagawa et al. (1992) and Zissman (1993) all applied HMMs to spectral and cepstral feature vectors. 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 each had been testing, though Nakagawa et al. (1994) eventually obtained better performance for his HMM approach than his static approaches.

Li (1994) has proposed the use of novel features for spectral-similarity LID. In his system, the syllable nuclei (i.e. vowels) for each speech utterance are located automatically. Next, feature vectors containing spectral information are computed for regions near the syllable nuclei. Each of these vectors consists of spectral sub-vectors computed on neighboring (but not necessarily adjacent) frames of speech data. Rather than collecting and modeling these vectors over all training speech, Li keeps separate collections of feature vectors for each training speaker. During testing, syllable nuclei of the test utterance are located and feature vector extraction is performed. Each speaker-dependent set of training features vectors is compared to the feature vectors of the test utterance, and the most similar speaker-dependent set of training vectors is found. The language of the speech spoken by the speaker of that set of training vectors is hypothesized as the language of the test utterance.

3.2. Prosody-based approaches

Features that carry prosodic information have also been used as input to automatic language identification systems. This has been motivated, in part, by studies showing that humans can use prosodic features for identifying the language of speech utterances (Muthusamy et al., 1994b, 1999). For example, Itahashi et al. (1994, 1995) have built systems that use features based on pitch estimates alone. He argues that pitch estimation is more robust in noisy environments than spectral parameters.

Hazen (1993), however, showed that features derived from prosodic information provided little language discriminability when compared to a phonetic system. A system that used both prosodic and phonetic parameters performed about the same as a system using phonetic parameters alone.

Finally, Thyme-Gobbel and Hutchins (1996) has also looked at the utility of prosodic cues for language identification. Parameters were designed to capture pitch and amplitude contours on a syllable-by-syllable basis. They were normalized to be insensitive to overall amplitude, pitch and speaking rate. Results show that prosodic parameters can be useful for discriminating one language from another; however, the accuracy of any particular set of features is highly language-pair specific.

3.3. Phone-recognition approaches

Given that different languages have different phone inventories, many researchers have built LID systems that hypothesize exactly which phones are being spoken as a function of time and determine the language based on the statistics of that phone sequence. For example, Lamel built two HMM-based phone recognizers: one in English and another in French (Lamel and Gauvain, 1993). These phone recognizers were then run over test data spoken either in English or French. Lamel found that the likelihood scores emanating from language-dependent phone recognizers can be used to discriminate between English and French speech. Muthusamy ran a similar system on English versus Japanese spontaneous, telephone-speech (Muthusamy et al., 1993).

The novelty of these phone-based systems was the incorporation of more knowledge into the LID system. Both Lamel and Muthusamy trained their systems with multi-language phonetically labeled corpora. Because the systems require phonetically labeled training speech utterances in each language, as compared to the spectral-similarity systems which do not require such labels, it can be more difficult to incorporate new languages into the language recognition process. This problem will be addressed further in Section 3.4.

To make phone-recognition-based LID systems easier to train, one can use a single-language phone recognizer as a front end to a system that uses phonotactic scores to perform LID. Phonotactics are the language-dependent set of constraints specifying which phonemes are allowed to follow other 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 cannot occur in English (except if one word ends in `/sh/` and the next begins with `/p/`, or in a compound word like “flashpoint”). This approach is reminiscent of the work of D’Amore and Mah (1985, 1988), Schmitt (1991) and Damashek (1995), who have used n-gram analysis of text documents to perform language and topic identification and clustering. 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 language identification. Hazen and Zue (1993) and Zissman and Singer (1994) each developed LID systems that use one, single-language front end phone recognizer. An important finding of these researchers was that language ID could be performed successfully even when the front end phone recognizer(s) was not trained on speech spoken in the languages to be recognized. For example, accurate Spanish versus Japanese LID can be performed using only an English phone recognizer. Zissman and Singer (1994) and Yan and Barnard (1995) have extended this work to systems containing multiple, single-language front ends, where there need not be a front end in each language to be identified. Fig. 2 shows an example of these types of systems.

3.4. Using multilingual speech units

Instead of training language-dependent phoneme recognizers, one can build multi-lingual

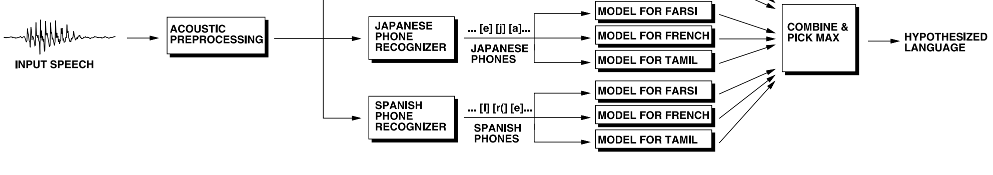

Fig. 2. A LID system that uses several phone recognizers in parallel.

speech units. These are derived by either a mixture of language-dependent and language-independent phones or by deriving tokens automatically from training data. Advantages of this approach include data sharing and discriminant training between phonemes across languages and easy bootstrapping to unseen languages (Wheatly et al., 1994).

Research has also focused on the problem of identifying and processing only those phones that carry the most language discriminating information (Berkling et al., 1994, 1995). These language-dependent phones are called "mono-phonemes" or "key-phones" in the literature. Kwan and Hirose (1997) and Dalsgaard and Andersen (1992) use both language-specific and language-independent phones in their systems. The language-independent phones, sometimes called "poly-phones", can be trained on data from more than one language without loss of LID accuracy. Berkling and Barnard (1995) and Koehler (1997, 1998) have also tested systems that use a single multi-language front end phone recognizer, i.e. a recognizer containing a mixture of poly-phones and mono-phones.

3.5. Word level approaches

Between phone-level systems described in the previous sections and the large-vocabulary speech recognition systems described in a subsequent section are "word-level" approaches to LID. These systems use more sophisticated sequence modeling than the phonotactic models of the phone-level systems, but do not employ full speech-to-text systems.

Kadambe and Hieronymus (1995) proposed the use of lexical modeling for language identification. An incoming utterance is processed by parallel language-dependent phone recognizers. Hypothesized language-specific word occurrences are identified from the resulting phone sequences.

Each language-dependent lexicon contains several thousand entries. This is a bottom-up approach to the language ID problem, where phones are recognized first, followed by words, and eventually language. Thomas et al. (1998) has shown that a language-dependent lexicon need not be available in advance; rather, it can be learned automatically from the training data. Ramesh and Roe (1994), Matrouf et al. (1998), Lund and Gish (1995, 1996) and Braun and Levkowitz (1998) have all proposed similar systems.

3.6. Continuous speech recognition

By adding even more knowledge to the system, researchers hope to obtain even better LID performance. Mendoza et al. (1996), Schultz et al. (1996, 1998) and Hieronymus and Kadambe (1997) have shown that large-vocabulary continuous-speech recognition systems can be used for language ID. 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 hold the promise of 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 LID decision. Furthermore, one obtains a transcription of the utterance as a byproduct of LID. On the other hand, they require many hours of labeled training data in each language to be recognized and are the most computationally complex of the algorithms proposed.

4. Evaluations

From 1993–1996, the National Institute of Standards and Technology (NIST) of the US Department of Commerce has sponsored formal evaluation of language ID systems. At first, these evaluations were conducted using the Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus (Muthusamy et al., 1992). The OGI-TS corpus contains 90 speech messages in each of the following 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Each message is spoken by a unique speaker and comprises responses to 10 prompts. For NIST evaluations, the monologue speech evoked by the prompt "Speak about any topic of your choice" is used for both training and testing. No speaker speaks more than one message or more than one language, and each speaker's message was spoken over a unique long-distance telephone channel. Phonetically transcribed training data is available for six of the OGI languages (English, German, Hindi, Japanese, Mandarin and Spanish).

Performance of the best systems from the 1993, 1994 and 1995 NIST evaluations is shown in Fig. 3. This performance represents each system's first pass over the evaluation data, which means that no system-tuning to the evaluation data was possible. For utterances having duration of either 45 s or 10 s, the best systems can discriminate between two languages with 4% and 2% error, respectively. This error rate is the average computed over all language pairs with English, e.g. English versus Farsi, English versus French, etc. When tested on nine-language forced-choice classification, error rates of 12% and 23% have been obtained on 45-s and 10-s utterances, respectively. The syllabic-feature system and the systems with multiple phone recognizers followed by phonotactic language modeling have exhibited the best performance over the years. Error rate has decreased over time, which indicates that research has improved system performance.

Starting in 1996, the NIST evaluations have employed the Linguistic Data Consortium's CALLFRIEND corpus. CALLFRIEND comprises two-speaker, unprompted, conversational speech messages between friends. Hundred North-American long-distance telephone conversations were recorded in each of twelve languages (the same 11 languages as OGI-TS plus Arabic). No speaker occurs in more than one conversation. In the 1996 evaluation, the multiple phone recognizer followed by language modeling systems again performed best. The error rates on 30 s and 10 s utterances were 5% and 13% for pairwise classification. These same systems obtained 23% and 46% error rates for 12-language classification. The higher error rates on CALLFRIEND are due to the informal conversational style of CALLFRIEND versus the more formal monologue style

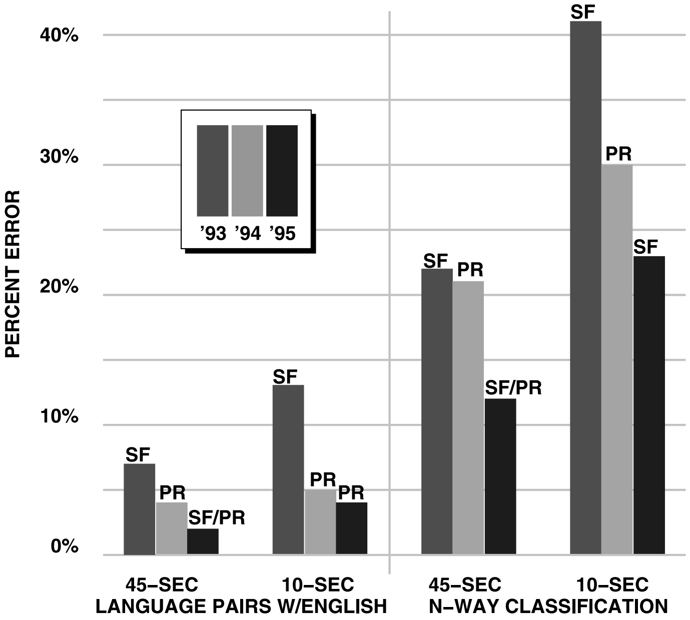

Fig. 3. Error rates of the best LID systems at three NIST evaluations. Performance is shown on the left for average two-alternative, forced-choice classification of the various OGI-TS languages with English. "N-way" classification refers to 10-alternative, forced-choice performance in 1993, 11-alternative, forced-choice performance in 1994, and 9-alternative, forced-choice performance in 1995. "SF" indicates syllabic feature system. "PR" indicates phone recognition followed by language modeling system.

of OGI-TS. The 1996 evaluation also measured each LID system's ability to perform dialect ID on two dialects from each of three languages (English, Mandarin and Spanish). The best systems exhibited 20–30% error rates on two-dialect, forced-choice tests.

The CSR-based LID systems have not been fully evaluated at NIST evaluations, because orthographically and phonetically labeled speech corpora have not been available in each of the requisite languages. As such corpora become available in more languages, implementation and evaluation of CSR-based LID systems will become more feasible. Whether the performance they will afford will be worth their computational complexity remains to be seen.

5. Conclusions

Since the 1970s, language identification systems have become more accurate and more complex. Current systems can perform two-alternative forced-choice identification on extemporaneous monologue almost perfectly, and these same systems can perform 10-way identification with roughly 10% error. Though error rates on conversational speech are somewhat higher, there is every reason to believe that continued research coupled with competitive evaluations will result in improved system performance.

The improved performance of newer LID systems is due to their use of higher levels of linguistic information. Systems which try to model phones, phone frequencies and phonotactics naturally perform better than those that model only lower-level acoustic information. Presumably, systems that model words and grammars will be shown to have even better accuracy.

Improved performance, however, comes at a cost. The higher levels of linguistic information must be programmed or trained into the newer LID systems. Whereas older systems required only digitized speech samples in each language to be recognized, more modern systems tend to require either a phonetic or orthographic transcription of at least some of the training utterances. State-of-the-art large-vocabulary CSR systems are often trained on many hours of transcribed speech. While large corpora of speech are available in many languages, they may not be available in all languages required by a specific application. Thus, the system developer must balance the need for accuracy against the need for speedy deployment and low-cost implementation.

References

Berkling, K.M., Barnard, E., 1995. Theoretical error prediction for a language identification system using optimal phoneme clustering. In: Eurospeech, Vol. 1, pp. 351–354.

Berkling, K.M., Arai, T., Barnard, E., Cole, R.A., 1994. Analysis of phoneme-based features for language identification. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, April 1998, pp. 289–292.

Braun, J., Levkowitz, H., 1998. Automatic language identification with perceptually guided training and recurrent neural networks. In: International Conference on Spoken Language Processing, Vol. 7, pp. 3201–3205.

Cimarusti, D., Ives, R.B., 1982. Development of an automatic identification system of spoken languages: phase I. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 1661–1663.

Comrie, B., 1990. The World's Major Languages. Oxford University Press, New York.

Crystal, D., 1987. The Cambridge Encyclopedia of Language. Cambridge University Press, Cambridge, UK.

Dalsgaard, P., Andersen, O., 1992. Identification of mono- and poly-phonemes using acoustic–phonetic features derived by a self-organizing neural network. In: International Conference on Spoken Language Processing, pp. 547–550.

Damashek, M., 1995. Gauging similarity with n-grams: language-independent categorization of text. Science 267 (5199), 843–848.

D'Amore, R.J., Mah, C.P., 1985. One-time complete indexing of text: theory and practice. In: Proceedings of the Eighth International ACM Conference on Research and Development in Information Retrieval, pp. 155–164.

Foil, J.T., 1986. Language identification using noisy speech. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 861–864.

Fromkin, V., Rodman, R., 1993. An Introduction to Language. Harcourt Brace Jovanovich, Orlando, FL.

Goodman, F.J., Martin, A.F., Wohlford, R.E., 1989. Improved automatic language identification in noisy speech. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 528–531.

Hazen, T., 1993. Automatic language identification using a segment-based approach. Ph.D. thesis, MIT.

Hazen, T.J., Zue, V.W., 1993. Automatic language identification using a segment-based approach. In: Eurospeech, Vol. 2, pp. 1303–1306.

Hieronymus, J.L., Kadambe, S., 1997. Robust spoken language identification using large vocabularly speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 1111–1114.

House, A.S., Neuburg, E.P., 1977. Toward automatic identification of the language of an utterance. I. preliminary methodological considerations.. J. Acoust. Soc. Amer. 62 (3), 708–713.

Itahashi, S., Du, L., 1995. Language identification based on speech fundamental frequency. In: Eurospeech, Vol. 2, pp. 1359–1362.

Itahashi, S., Zhou, J., Tanaka, K., 1994. Spoken language discrimination usin speech fundamental frequency. In: International Conference on Spoken Language Processing, Vol. 4, pp. 1899–1902.

Kadambe, S., Hieronymus, J., 1995. Language identification with phonological and lexical models. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp. 3507–3511.

Kimbrell, R.E., 1988. Searching for text? Send an n-gram!. Byte 13 (5), 297–312.

Koehler, J., 1997. In-service adaptation of multilingual hidden-markov-models. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 1451–1454.

Koehler, J., 1998. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 417–420.

Kwan, H.K., Hirose, K., 1997. Use of recurrent network for unknown language rejection in language identification systems. In: Eurospeech, Vol. 1, pp. 63–67.

Lamel, L.F., Gauvain, J.-L., 1993. Cross-lingual experiments with phone recognition. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 507–510.

Li, K.-P., 1994. Automatic language identification using syllabic spectral features. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 297–300.

Lund, M.A., Gish, H., 1995. Two novel language model estimation techniques for statistical language identification. In: Eurospeech, Vol. 2, pp. 1363–1366.

Lund, M.A., Ma, K., Gish, H., 1996. Statistical language identification based on untranscribed training data. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 793–796.

Matrouf, D., Adda-Decker, M., Lamel, L.F., Gauvain, J.L., 1998. Language identification incorporating lexical information. In: International Conference on Spoken Language Processing, Vol. 2, pp. 181–185.

Mendoza, S., et al. 1996. Automatic language identification using large vocabulary continuous speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 785–788.

Mori, K., Toba, N., Harada, T., Arai, T., Komatsu, M., Aoyagi, M., Murahara, Y., 1999. Human language identification with reduced spectral information. In: Eurospeech.

Muthusamy, Y.K., Cole, R.A., Oshika, B.T., 1992. The OGI multi-language telephone speech corpus. In: International Conference on Spoken Language Processing, Vol. 2, pp. 895–898.

Muthusamy, Y., et al. 1993. A comparison of approaches to automatic language identification using telephone speech. In: Eurospeech, Vol. 2, pp. 1307–1310.

Muthusamy, Y.K., Barnard, E., Cole, R.A., 1994a. Reviewing automatic language identification. IEEE Signal Process. Mag. 11 (4), 33–41.

Muthusamy, Y.K., Jain, N., Cole, R.A., 1994b. Perceptual benchmarks for automatic language identification. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 333–336.

Nakagawa, S., Ueda, Y., Seino, T., 1992. Speaker-independent, text-independent language identification by HMM. In: International Conference on Spoken Language Processing, Vol. 2, pp. 1011–1014.

Nakagawa, S., Seino, T., Ueda, Y., 1994. Spoken language identification by ergodic HMMs and its state sequences. Electron. Commun. Jpn., Part 3 77 (6), 70–79.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Ramesh, P., Roe, E., 1994. Language identification with embedded word models. In: International Conference on Spoken Language Processing, Vol. 4, pp. 1887–1890.

Riek, L., Mistretta, W., Morgan, D., 1991. Experiments in language identification. Technical Report SPCOT-91-002, Lockheed Sanders, Inc., Nashua, NH.

Savic, M., Acosta, E., Gupta, S.K., 1991. An automatic lanuguage identification system. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 817–820.

Schmitt, J.C., 1991. Trigram-based method of language identification. US Patent 5,062,143.

Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. In: International Conference on Spoken Language Processing, Vol. 5, pp. 1819–1823.

Schultz, T., Rogina, I., Waibel, A., 1996. LVCSR-based language identification. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 781–784.

Sugiyama, M., 1991. Automatic language recognition using acoustic features. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 813–816.

Thomas, H.L., Parris, E.S., Wright, J.H., 1998. Recurrent substrings and datafusion for language recognition. In: International Conference on Spoken Language Processing, Vol. 2, pp. 169–173.

Thyme-Gobbel, A.E., Hutchins, S.E., 1996. On using prosodic cues in automatic language identification. In: International Conference on Spoken Language Processing, Vol. 3, pp. 1768–1772.

Wheatly, B., et al. 1994. An evaluation of cross-language adaptation for rapid hmm development in a new language. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 237–240.

Yan, Y., Barnard, E., 1995. An approach to automatic language identification based on language-dependent phone recognition. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp. 3511–3514.

Zissman, M.A., 1993. Automatic language identification using Gaussian mixture and hidden Markov models. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 399–402.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech and Audio Proc. SAP-4 (1), 31–44.

Zissman, M.A., Singer, E., 1994. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 305–308.

Zissman, M.A., Singer, E., 1995. Language identification using phoneme recognition and phonotactic language modeling. In: International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp. 3503–3506.