Introduction to Automatic Language Identification

Pavel Matějka[1], Jan Černocký[2], Milan Sigmund[1]

[1] Faculty of Electrical Engineering and Communication, BUT
Brno, Purkuňova 118, CZ 612 00, Phone: +420-5-41149156, Fax: +420-5-4114 9244
[2] Faculty of Information Technology, BUT
Brno, Božetěchova 2, CZ 612 66, Phone: +420-5-41141283, Fax: +420-5-41141270
E-mail: matejkap|sigmund@feec.vutbr.cz cernocky@fit.vutbr.cz

This paper provides introduction to automatic language identification. Problems and motivations are described and methods used in this field are discussed. In the end of the paper, we present NIST campaigns run in previous past years and aiming at world-wide quantitative comparison of results of different labs.

1 Introduction

Automatic spoken Language Identification (LID) is the process of classifying an utterance as belonging to one of a number of previously encountered languages. "Automatic", because the decision is performed by machine. It is implied that the process is independent of content, context, task or vocabulary and robust with regard to speaker identity, sex, age as well as to noise and distortion introduced by the communication channel.

2 Problem Specification

As with speech recognition, humans are the most accurate language identification systems in the word today [1]. Within a second of hearing, people are able to determine if it is language they know. If it is language which they don't know or are not familiar with, they often can make subjective judgment about a similarity of the language with which they know, e.g. "sounds like Czech".

There are a variety of cues that humans or machines can use to distinguish between languages. We know that the following characteristics differ from language to language:

- **Phonetics** - Though the human speech system is potentially capable of an unlimited range of sounds, in any language there is a limited number of recurrent, fairly distinctive speech units (phones/phonemes). Even thought many languages share a common subset of phonemes. Phonemes frequencies may also differ, i.e., a phoneme may occur in two languages, but it may be more frequent in one language than the other. The number of phonemes in a language ranges from about 15 to 50, with peak at 30 [2].
- **Phonotactics** - Not only do phoneme inventories differ from language to language, but also does the phoneme combination or sequence of allowable phonemes. Some combinations that occur frequently in one language are illegal in another. Phonotactics help to recapture some of the dynamical nature of speech lost during feature extraction.

- **Prosody** - Prosody is concerned with the ”music” as opposed to the ”lyrics” of speech. Languages have characteristic sound patterns which can be analyzed in terms of duration of phonemes, speech rate, intonation (pitch contour) and stress.
- **Morphology** - 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.
- **Syntax** - The ways in which words can be legally strung together also potentially distinctive information. Even when two languages share a word, e.g. the word ”bin” in English and German, the set of words which can precede and follow the word will be different.

A successful language identification system would use information from all the above mentioned sources to come up with identification decision.

3 Motivation

There are 6000 million people on the world. 64% of them speak 14 languages from around 3000 world known languages [3].

To communicate with each other is necessary to know language with which we are dealing. Language identification is not useful only in personal life, but also in the interest of national security services for monitoring communications and many other fields.

Everybody can imagine the multi-lingual system for example in the airport or rail station for information, checking into hotel, making travel arrangements and so on. All this can be difficult for non-native speakers.

If a system has no input other than speech, then such system has to be capable to determine the language of the speech commands. There are two ways of recognizing commands. First is during the command and second before recognizing commands. Recognition during the command pronouncing would require running many speech recognizers in parallel, one for each language. There can be hundreds nations with different languages in the airport, therefore we might run hundreds recognizers in real-time with a really big computational cost. Alternatively, a language-ID could be run prior to the speech recognizer. This system will output the list of the most probably languages and a few language dependent recognizers can be run to process commands. Language decision can be made only once for set of speech commands. Such system was introduced by Hazen and Zue [4].

Alternatively, LID might be used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language. 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 and requires connections to several human interpreters before the appropriate interpreter is found. The delay in finding appropriate human interpreter can be in the order of minutes as reported by Muthusamy [5]. Such delay can be devastating in emergency situation. An automatic language identification system that can quickly determine the most likely languages of incoming call might cut the delay by one or two orders of magnitude.

4 NIST Evaluations

National Institute of Standards and Technology (NIST)started a Language Identification Evaluations for comparative results all over the word and give support to new efforts in this field. First NIST Language Identification Evaluation was in **1994** and the evaluation data contained speech from OGI multi-language telephone speech corpus [6] which contains monologue speech.

Formal evaluation in **1996** demonstrated that systems that used parallel banks of tokenizer-dependent language models produced the best language identification performance. Evaluation data contained conversational speech merged over several different databases. These were mainly Switchboard and the OGI multi-language telephone speech data. Some results are reported in Table. 1.

SYSTEM EER [%]	**30s**	**10s**	**3s**
MIT 1996	9.9	19.4	29.4
OGI 1996	11.8	20.9	30.7
MIT 2003	6.5	14.2	25.5
OGI 2003	7.7	11.9	22.6

Table 1: Equal Error Rate on NIST 1996 evaluation test set for PPRLM systems from 1996 and 2003

The **2003** NIST Language Recognition Evaluation was very similar to the one in 1996 [7]. It was intended to establish a new baseline of current performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field. The primary evaluation data consisted of excerpts from conversations in twelve languages from the CallFriend Corpus [8]. These test segments had durations of approximately three, ten, or thirty seconds. Six sites from three continents participated in the evaluation. The performance results were significantly improved from those of the previous evaluation. Two of the best systems in 1996 and the systems based on the same strategy about 7 years later is shown in the Table 1. The comparative results from some sites are in Table 2. Six language-dependent phoneme recognizers ran in parallel followed by language modeling are system with label MIT PPRLM [9] and OGI PPRLM [10], one English phoneme recognizer [11], based on novel approach, followed by LM had label 3BT TRAP PRLM, the conventional, but improved, Gaussian mixture model MIT GMM and the system merged acoustic and phonotactic information called MIT Fuse performed the best, but the need of computational power wasn't negligible.

SYSTEM EER [%]	**30s**	**10s**	**3s**
MIT Fuse	2.8	7.8	20.3
MIT GMM	4.8	9.8	19.8
MIT SVM	6.1	16.4	28.2
MIT PPRLM	6.6	14.3	25.5
OGI PPRLM	7.71	11.88	22.60
3BT TRAP PRLM	12.71	22.71	32.19

Table 2: Equal Error Rate on standard NIST 2003 evaluation test set

5 Conclusion

Language recognition together with speech recognition become one of the most interesting field for commercial and noncommercial applications. Language identification system can be employed for example in emergency call routers, information services and to provide information in person's native language. Second large application area are security and defense services. Cues that humans or machines can use to distinguish between languages are described together with the comparison of the latest approaches to LID.

6 Acknowledgments

This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485. Pavel Matějka was supported by doctoral grant of Grant Agency of Czech Republic No. 102/03/H105 and Jan Černocky by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

References

[1] Muthusamy, Y.K., Jain, N., Cole, R.A. Perceptual Benchmarks for Automatic Language Identification. *in Proc. ICASSP 94*, vol. 1, pp. 333-336, Apr. 1994.

[2] Zissman, M.A. Overview of Current Techniques for Automatic Language Identification of Speech. *IEEE Workshop ASRU*, pp. 60-62, December 1995.

[3] Boner. A. *Spracherkennung mit Computer*, AT Verlag, Aarau (Switzerland), 1992, ISBN 3-85502-495-9

[4] Zissman, M.A. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44, January. 1996.

[5] Muthusamy, Y.K., Bernard, E., Cole, R.A. Automatic Language Identification: A Review/Tutorial. *IEEE Signal Processing Magazin*, vol. 11, no. 4, pp. 33-41, October. 1994.

[6] OGI MultiLanguage Telephone Speech. *www.cslu.ogi.edu/corpora/mlts/*, Januar 2004.

[7] Martin, A.F., Przybocki, M.A. NIST 2003 Language Recognition Evaluation. *Proc. Eurospeech'03*, pp. 1341-1344, September 2003.

[8] CallFriend Corpus, telephone speech of 15 different Languages or dialects, *www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone.*

[9] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A. Acoustic,Phonetic,and Discriminative Approaches to Automatic Language Identification. *Proc. Eurospeech'03*, pp. 1345-1348, September 2003.

[10] Yan, Y., Barnard, E. An Approach to Automatic Language Identification Based on Language-dependent Phone Recognition. *Proc. ICASSP'95*, pp. 3511-3514, May 1995.

[11] Schwarz, P., Matějka, P., Černocký, J. Recognition of Phoneme Strings using TRAP Technique. *Proc. Eurospeech'03*, pp. 825-828, September 2003.