Automatic Language Identification

Martine Adda-Decker

July 20, 2008

8.1. Introduction

When listening to our native language we, speech and hearing enabled humans, immediately identify the language being spoken. Furthermore for familiar languages, we are able to do almost as well. This is a capacity we would like to share with automatic devices. Human capacities are certainly impressive on short speech fragments, however the number of native and familiar languages generally remains very limited as compared to the number of distinct spoken languages in the world. Furthermore the risk of error, even if low, is not completely negligible.

Technically speaking, language identification (LId) consists in identifying the language being spoken by some unknown speaker using a given speech sample. Automatic LId appears as one of the expected capacities of smart speech-to-text or speech-to-speech systems in a multilingual context. Automatic LId raises many challenging research issues which contribute to progress towards multilingual speech technologies [SCH 06, NAV 06, GEO 04, ZIS 99]. Major questions in the domain of automatic language identification include: which levels of information contribute most effectively to high accuracy language identification? What are the best acoustic parameters to catch language-specific information? What algorithms, methods and approaches are the most promising for LId? Which language-specific resources are required for spoken language modeling? What is the correlation between signal duration and the confidence of language identification? Related questions of interest concern human performance: how do humans proceed in identifying languages? Can insight into human cognition be helpful for automatic devices or vice-versa? Another important question concerns the definition of what is to be considered a language. Beyond the LId issue in itself, automatic processing of speech from multiple languages produces baseline resources which enable or enhance large scale comparative studies in phonetics, phonemics and prosody. Although this is not the main goal of LId research, it is an important interdisciplinary byproduct, giving rise to fruitful interactions between researchers in linguistics and speech scientists (e.g. [GEN 07]). For spoken minority languages, which are not (or only scarcely) described and documented, automatic processing tools may help linguists to elaborate phonemic, morphological and text processing systems.

Automatic LId has been an active research domain for about 30 years, with the pioneering works of Leonard & Doddington (1974), which relied on an acoustic filterbank approach, and of House & Neuburg (1977) [LEO 74, HOU 77], who made the first contribution to LId using language-specific phonotactic constraints. Different sources of information are known to contribute to human language identification, among which the most important ones are the acoustic, phonetic, phonemic and phonotactic levels, but also prosody as well as lexical and morphological knowledge. All these levels are not of the same importance to LId, nor are they equivalently easy to capture in computationally tractable models. Acoustic-phonemic and phonotactic approaches benefit from decades of research, first by linguists to describe languages using compact phonetic and phonemic systems [IPA 99, LAD 96, SCH 97, VAL 99], and more recently by computer speech scientists elaborating phone models for automatic speech recognition together with appropriate language resources. Hence acoustic-phonemic and phonotactic modeling became the most popular approaches to LId [LAM 94, LAM 95, DAL 96, ZIS 94, ZIS 96a] in the early years of LId research. Other sources of information, such as prosody or morphology, can be used as a complement rather than in a stand-alone approach.

Until recently, research in speech processing has primarily addressed the world's major languages for which standardized writing systems exist and for which a prolific production of various types of language resources allows for language-specific modeling beyond a mere acoustic level. An important issue is the seamless extension of automatic LId devices to additional languages. The relative sparsity of language-specific resources for most spoken languages has motivated research work in acoustic-based approaches, as the lack of resources restricts the range of LId modeling options. Positive outcomes of these acoustic-based approaches consist both in easily bootstrapping LId devices to more languages and in reducing the related costs in terms of human effort. However it is worthwhile to note that there are important ongoing efforts to build language resources in an increasingly high number of spoken languages, so as to enable a variety of multilingual speech technologies, including for example automatic transcription and indexing of audio documents, speech translation and synthesis. As a matter of fact, research in multilingual speech processing has been supported by the European Commission for almost thirty years now, as linguistic diversity (with more than 20 official languages now) is one of Europe's challenging specificities [MAR 98, MAR 05]. Resources are collected and distributed by the European Language Resource Distribution Agency (ELRA/ELDA) [ELD]. In the United States the Defense Advanced Research Project Agency (DARPA) has been massively fostering multilingual research for more than fifteen years now and the Linguistic Data Consortium (LDC) has an impressive catalogue of multilingual resources (speech, text and lexica) in at least 80 languages [LDC 07].

Given the progress achieved in automatic LId, there is a growing interest in more subtle LId problems, such as dialect and accent identification [ZIS 96b, KUM 96], as witnessed by recent studies on dialectical forms of Mandarin, Spanish, Arabic, different South-African languages, regional accents of English and French[TOR 04, BAR 99, NIE 06]. However, in multilingual contexts, speakers do not necessarily communicate in their native languages. A speech signal may include code switching, code mixing and possibly non-native speech. The definition of language identification, as stated at the beginning of this chapter, may then become more complex: the speech signal conveys information about the language being spoken and the speaker's native language. Should LId systems only identify the language being spoken or should they also be able to give some information about the speaker's accent (related to his/her native language)? Until recently LId has mainly addressed the problem of identifying a spoken language, where speakers use their native language for communication. Research in automatic language identification may also include the challenges of non-native speech to increase the usability of speech systems in multilingual environments [TEI 97, WAN 99, WIT 99, COM 01, BAR 07].

A range of applications can be envisioned for automatic language identification: telephone companies need to quickly identify the language of foreign callers to route their calls to operators who speak that language. Automatic language identification can be useful in governmental intelligence and monitoring applications. For multilingual translation services, systems need to identify the language being spoken prior to translation. Multilingual audio indexing systems may be enhanced, by integrating dialect and accent identification capacities beyond LId in itself.

8.2. Language characteristics

The number of distinct spoken languages is very high, even though the exact number is difficult to establish. Depending on the sources, the number of distinct languages varies, between 4000 and 8000 [CRY 97, COM 90, KIR 06, ETH 05, LAV]. The reasons for this lack in precision is mainly due to variable definitions of what is to be counted as a language. On a diachronic axis, there is the question of living vs extinct languages [HOM 06]. Latin or ancient Greek, although extinct in the sense that there are no longer native speakers, remain alive through education, as well as their literary and historical legacy. An important number of still living minority languages however are in real danger of extinction today. From the synchrony viewpoint, the distinction between language and dialect is not clear-cut. A living language can be considered as a shared and collectively agreed upon code or competence [CHO 77] of a community, a population or a social group at a given time and place. A spoken language then corresponds to the oral performance of a set of speakers issued from such a population, community or group. With this view of spoken language in mind, the number of existing languages blatantly depends on the applied granularity to achieve homogeneous populations, as well as time and space divisions. Dialects of a given language are supposed to imply, beyond potential pronunciation shifts, some changes on lexical and grammatical levels. However they are supposed to keep a mutual intelligibility. Different language/dialect classifications helps to explain diverging language counts.

Whatever the exact number of different spoken languages, the vast majority of the world's population speaks a very limited set of languages. Zipf's law [ZIP 49], which applies to word occurrences in large corpora, also seems to apply to language populations: about 95% of all the speakers make use of only 5% of the world's languages. Another important fact concerning spoken language technologies in general and automatic language identification in particular is that only 5-10% of the world's spoken languages admit a corresponding writing system [DAN 96]. The vast majority of languages are only spoken. This point is also worth remembering, when discussing different approaches to LId, which may more or less rely on high-level language-specific resources. Table 8.1 shows the most widely spoken languages in the world ranked by the number of native speakers [CRY 93]. The numbers of official language speakers (L1+L2) are also given. It is interesting to note that Chinese (including most importantly Mandarin but also Xiang, Hakka, Gan and Minbei) ranks first due to the huge number of native speakers, whereas English becomes number one, when official language populations are added. French, while holding rank 11 for native speakers, progresses to rank 6, when including secondary speakers. For Bengali, the opposite tendency is observed. The field of historical linguistics aims at organizing the set of inventored languages in family trees according to their genetic relatedness. Figure 8.1 represents the language family tree for the 11 languages of the first LId corpus collected by the Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute (OGI) [MUT 93]. The Indo-European branch is the family with the largest number of speakers. Spoken languages, although sharing common features which are exploited to achieve genetic classifications, may vary significantly in their surface forms. This qualitative observation raises the question of how to quantify these perceived differences. The sound structure of a language can be described at different levels in terms of phonetics [LAD 96, IPA 99], phonemics [IPA 99] and prosody [HIR 99]. Whereas *phonetics* aim at an objective

Language	*# L1*		*# L1+L2*		*Family*
	Rank	*Speakers (M)*	*Rank*	*Speakers (M)*	
Chinese	1	1,000	2	1,000	Sino-Tibetan
English	2	350	1	1,400	Indo-European
Spanish	3	250	4	280	Indo-European
Hindi	4	200	3	700	Indo-European
Arabic	5	150	7	170	Afro-Asiatic
Bengali	6	150	10	150	Indo-European
Russian	7	150	5	270	Indo-European
Portuguese	8	135	8	160	Indo-European
Japanese	9	120	10	120	Japanese
German	10	100	11	100	Indo-European
French	11	70	6	220	Indo-European

Table 8.1. *Most frequent languages ranked according to the number of native speakers (L1), including secondary speakers (L1+L2), expressed in millions (M), after [CRY 93].*

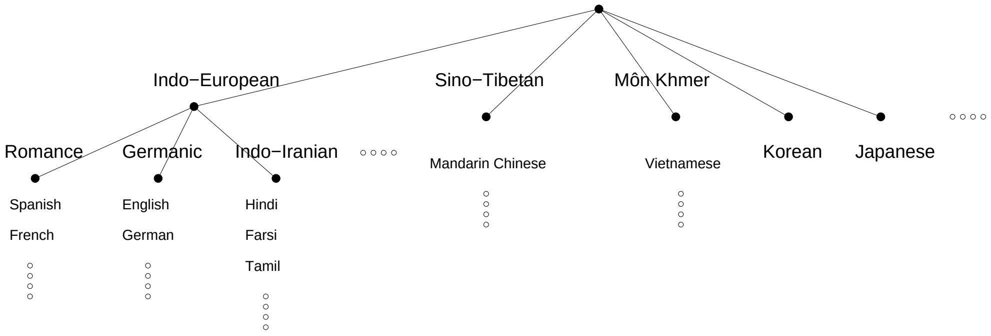

Figure 8.1. *Language family tree for the 11 languages of the first LId telephone speech corpus collected by CSLU-OGI in the early 90ies. Japanese and Korean are isolated in specific families.*

description of sound acoustics with a focus on linguistically relevant aspects, phonemics focus on the functional and linguistically distinctive roles of abstract sound representations in a language system. *Prosody* deals with fundamental frequency, energy and duration, tones, rhythmic patterns, intonation and prominence. Sounds, when studied as mere physical events are referred to as phones, whereas phonemic sounds are termed phonemes. The IPA [IPA 99] (International Phonetic Alphabet), which aims at providing a universally agreed upon system of notation for the spoken sounds of the world's languages, lists more than hundred distinct elementary phonetic symbol codes (without considering diacritics and diacritical combinations). IPA symbols (or derived alphabets such as SAMPA [WEL 97], Arpabet [SHO 80], Worldbet [HIE 94]) are very popular for acoustic-phonetic modelling of speech. A very comprehensive description of phonetic alphabet symbols used by linguists and speech scientists to record the sounds of the world's languages can be found in [PUL 96]. Phonemic inventories as well as corresponding phonetic realizations certainly reflect language-specific cues. Even hesitation segments carry some language-specific information [VAS 05]. An interesting question concerns the efficient use of rare, but salient segment information [HOM 99], e.g. clicks in African languages. Language-specific inventories generally range from 20 to 60 symbols. For example German has twice as many phonemic symbols than Spanish. The inventory size may significantly increase when tonal or gemmination information is explicited for the concerned languages (e.g. Mandarin for tones, Italian for gemminates). Consonant and vowel distributions as well as cooccurences of consonants and vowels are highly language-specific, and phonotactic constraints are known to be very important to identify languages and language styles [HOU 77, DEL 65, MAR 13]. Even simple acoustic measures, such as long term spectra exhibit differences among languages which might be related to supralaryngeal settings. Syllabic skeletons vary among languages and language-specific prosodic contours are among the earliest acquisitions in infant's babbling [LEV 91, HAL 91]. Research in language typology aims at extracting language universals and at organizing language differences [GRE 78, RUH 05, VAL 99].

Depending on the languages under consideration, differences in acoustic surface forms spread in variable proportions over the different levels of information. In general, largest discrepancies can be observed on the lexical level. Less abstract levels can feature more or less easily detectable differences depending on the examined language characteristics. For LId, the list of languages to classify and discriminate certainly influences identification and detection capacities of humans as well as of automatic devices. To illustrate typical differences between the acoustic signal of spoken languages, Figure 8.2 compares short speech excerpts of a Romance and a Germanic language (French vs Luxembourgish). Language-specific differences may be observed on different levels of the information representation. While the respective phonemic inventories share many symbols, there are important differences, especially in the vowel sets: French has nasal vowels, which are not part of the Luxembourgish native inventory, whereas the latter (similar to English) makes extensive use of diphthongues, which do not belong to the French inventory. Diphthongues entail a change of color in the corresponding vocalic segments, whereas formant frequencies of French vowels tend to remain very stable. Differences in distributions of phones and phone sequences are also illustrated by Figure 8.2. The example shows a large proportion of CV syllables for French, whereas complex syllable structures are rather common in Luxembourgish, as they are in other Germanic languages, likewise English or German. Even if not familiar with both languages, human listeners may easily detect differences arising from their sound structures. However, distinguishing between two close languages of the same family, such as Italian and Spanish, might be much more complex, even for listeners acquainted with both languages [BOU 04].

8.3. Language identification by humans

In this section, a short overview of human language identification competence is presented. The underlying questions are the following: how do humans proceed to identify a language? How important are the different levels of information (acoustic, phonetic, phonemic, phonotactic, prosodic, lexical...)? Different identification strategies are probably used depending on the degree of knowledge and of exposure to the languages to be identified by the human listener. The “distance” between languages (differences in terms of phonetic realizations, phonemic inventories, of syllabic structures, of the use or not of tonal information. . .) is certainly another important parameter.

Whereas new-born babies are considered as universal phoneticians, they rapidly specialize towards their native language throughout the first months of their lives, and by the end of the first year, native language characteristics (vowels, consonants,

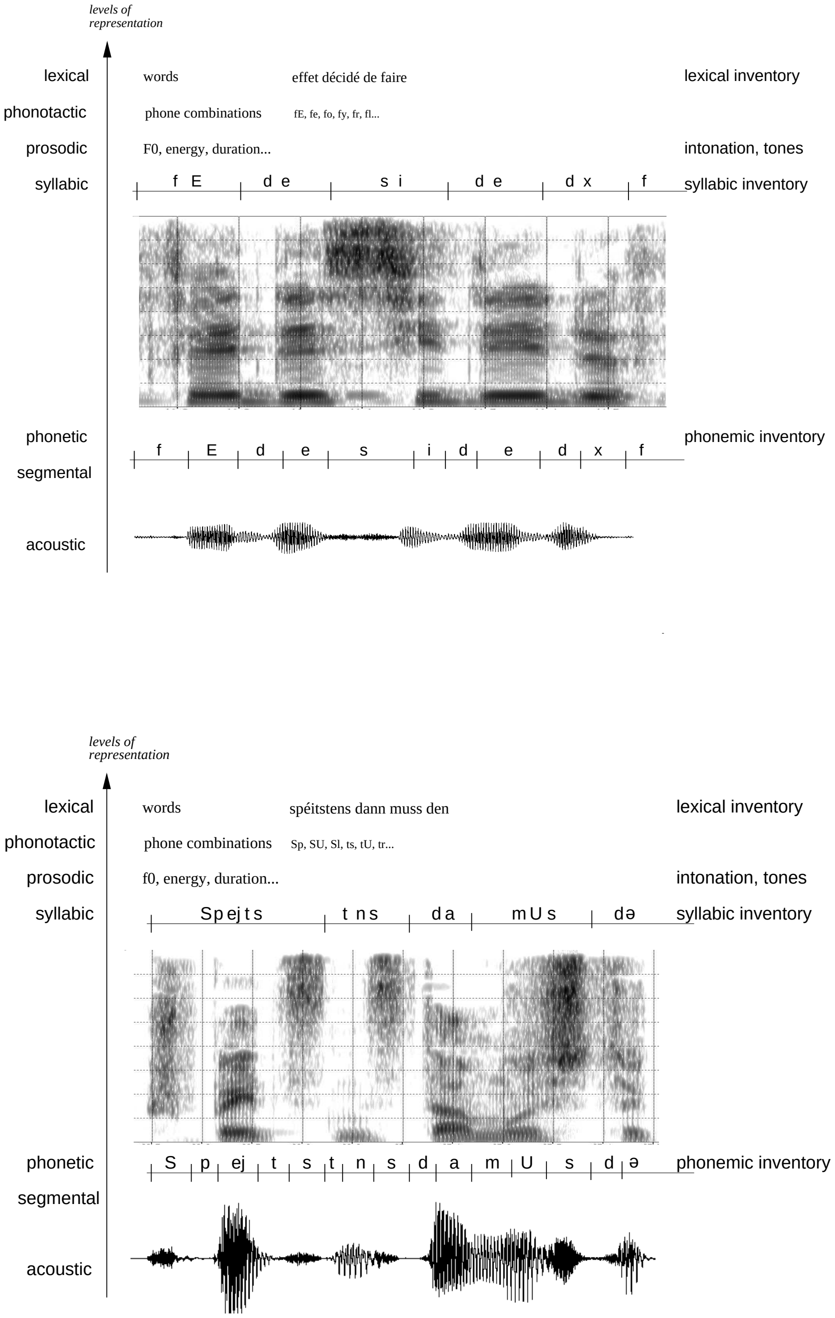

Figure 8.2. *Speech encodes language-specific information at different levels of representation. The first example shows an excerpt of French* (Romance language) *read speech* (`...en effet décidé de faire...` *(eng.* `...actually decided to do...`*)). The second example displays a Luxembourgish* (Germanic language) *speech sample* (`spêitstens da(nn) muss den...`*(eng.* `by then at the latest it needs...`*)) extracted from radio news.*

syllables, prosody) are clearly established [BOY 91] in both perception and production. Concerning adults' capacities, perceptual experiments have been conducted by many researchers to evaluate human performances in identifying languages [RAM 99, MAD 02]. Different experimental setups can be envisioned: language discrimination corresponding to ABX classifications or language identification corresponding to a necessary choice among a closed set of K languages. Original or modified speech stimuli can be used depending on whether the question of interest is either to assess the intrinsic difficulty of discrimination between languages (respectively classifying languages) or, in the case of modified speech, to measure the relative importance of some type of information encoded in the acoustic signal (e.g. prosodic information by filtering out the phonetic content) [RAM 99, MOR 99]. To address the question of the relative importance of different information levels, interesting results have been produced by J. Navrátil [NAV 01] on five languages (English, German, French, Mandarin and Japanese). Three series of perceptual tests have been carried out: 1) original stimuli, 2) randomly concatenated syllables extracted from the original stimuli and 3) filtered speech with a flattened spectral shape preserving only the $f0$ contour (see Table 8.2). In this experiment with a relatively small number of five languages mainly stemming from different families, identification results are very high (the average of 93.6% on 3 second stimuli increases up to 96% for 6 second excerpts). Next, results are highest on languages for which listeners had the largest background knowledge. The second condition keeps segmental information unmodified and preserves most of the phonotactic information. However the meaning and most of the prosodic information is lost. This loss of information entails a significant decrease in identification accuracy (22.1% absolute). Whereas the loss is almost negligible for English (the language which is most familiar to all listeners), the performance decrease is particularly high for the languages from the Asian area, which are the least familiar ones to most of the listeners. This suggests that the information of language identity may be very redundant for well known languages, as a partial loss has no major impact. This redundancy also makes it possible to perform reasonably well on less known languages.

Test	*English*	*German*	*French*	*Mandarin*	*Japanese*	*Average*
original (3s)	100.0	98.7	98.7	88.7	81.7	93.6
shuffled syll (6s)	98.7	79.7	79.1	57.7	54.6	73.9
$f0$ contour (6s)	34.3	34.3	69.4	65.9	45.3	49.4

Table 8.2. *Language classification accuracy (%) using different speech stimuli from telephone conversations: original (duration 3s) , shuffled syll (duration 6s) corresponds to a random concatenation of unmodified syllables, $f0$ contour (6s) corresponds to filtered speech keeping $f0$ information unmodified (after Navrátil 2001).*

However, in degraded conditions, the subtraction of part of the language identity related information may entail dramatic losses of performance. This is particularly true in the last condition, where all information, except the $f0$ contour, is filtered out. Here, English is not better identified than German, and both are significantly worse than French and Mandarin. The prosodic contours of both of these languages prove to be very informative to LId given this particular perceptual test configuration.

Similar perceptual experiments with French native listeners using unmodified short stimuli (2s) of broadcast news speech from eight languages (English, German, French, Italian, Spanish, Portuguese, Arabic, Mandarin) gave identification results of about 85% [ADD 03], which must be compared to the 93% obtained in the previous experiment with five languages. The stimuli here are significantly shorter and the number of languages higher. Whereas results on Mandarin are almost perfect, the results for the three non-French Romance languages provide a low average accuracy level of only 78%, even though French listeners are more familiar with the Romance languages (without necessarily practicing them) than with Mandarin. These results highlight that perceptual identification results must be examined with respect to the chosen language set: this is an important parameter when interpreting perceptual (and possibly also automatic) identification performances.

The experiments described in this section tend to show that the information of language identity is not exclusively encoded in one single information level, but across different levels, each one providing partly redundant information. This hypothesis then allows to envision different approaches, models and architectures to automatic LId, which may rely more or less on these different levels.

8.4. Language identification by machines

Automatic language identification is the process of using a computer system to identify the language of a spoken speech sample. Despite this very simple formulation, the process may become complex, as the identity of a spoken language may be blurred by code mixing and code switching, by dialects [BAR 99], regional or foreign [BOU 06, FLE 03] accents, or other variations of different languages in contact. Language differences are only part of the observed differences arising from speakers, uttered messages and environmental conditions. In the following, we briefly describe the major LId tasks, the corresponding performance measures and give some pointers to LId evaluations.

8.4.1. *LId tasks*

LId can be addressed via different task setups of varying academic and applicative interest, corresponding to different classification tasks [NIS 03]. Traditionnally the

LId problem has been addressed as a **closed set identification** task, which consists in identifying a speech input as corresponding to one language among a collection of K a priori given languages (Figure 8.3, left). A collection of K language-dependent models is thus required. Whereas this condition is certainly of scientific interest, the closed set assumption is too restrictive for most real-life applications. Another language recognition setup is given by the **open set detection/verification** task (Figure 8.3, right). The system's task is to decide whether or not the signal corresponds to the target language L. Speech input may come from an open set of languages, i.e. it doesn't necessarily belong to one of the modeled target languages. This task corresponds to selective language filtering. A language detection system can take as input any language and ideally outputs NO for all languages except for the target language L. Besides a model for the target L, there is generally a complementary model $\bar{L}$, also termed as the universal background model (UBM). The detection task better matches some of the real application needs (e.g. language filtering in multilingual audio streams). The most general task, which combines the positive features of the two preceding ones, corresponds to an **open-set identification** or multi-target detection [SIN 04]. It consists in either rejecting a speech input, if it stems from an unknown language, or otherwise, in identifying it as one of the multiple known languages (multi-target). Open-set identification can be implemented as a closed set identification (producing L^*) followed by a language detection system (producing YES or NO for L^*) or as K detection systems in parallel [SIN 04] [1].

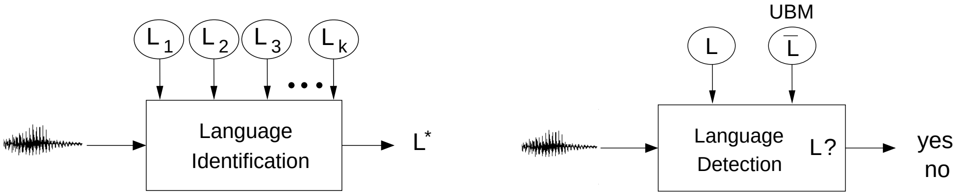

Figure 8.3. *Schematic representation of a language recognition system, which may be implemented either as language identification (left) or as language detection (right).*

8.4.2. *Performance measures*

Depending on the language recognition task, different evaluation measures may apply. For the closed set identification task, the errors correspond to substitutions, and LId error rates can simply be measured as average substitution rates. For detection tasks, the situation is slightly more complicated: a speech signal from a target

[1] Throughout this chapter, if not specified otherwise, LId stands for the general problem of language recognition, either identification or detection.

language can be rejected, and vice-versa a stimulus of an unknown language can be falsely identified as the target. Hence, performance measures for language detection are generally given as equal error rates (EER), equalizing the contribution of the two types of errors: false alarm and miss rates. This measure corresponds to a specific operating point in the Receiver Operating Characteristics (ROC) curve or equivalently, the Detection Error Trade-off (DET) curve [MAR 97]. ROC/DET curves give the full set of operating points, i.e. the rejection rate as a function of the acceptance rate. Examples of LId DET curves are given in subsequent sections (Figures 8.9, 8.13)[2]. Comparing the problems of automatic language vs automatic speaker recognition, it is worthwhile noting that distinct identities can be uniquely associated to each speaker (at least in ideal conditions), whereas spoken languages may correspond to more or less clearly defined classes, given first the difficulty of defining exact language contours and next the potential bilingual and more generally polyglot capacities of speakers. Independently from the recognition task and from the adopted approach, the decision accuracy largely depends on the stimulus length.

8.4.3. *Evaluation*

Unlike the numerous ARPA/NIST evaluation campaigns in automatic speech transcription and speaker recognition, a small number of language recognition evaluations have been organized by NIST (National Institute of Standards and Technology). The first evaluations took place in 1996, followed more recently by other campaigns in 2003 and 2005 [MAR 03, NIS 03, NIS 05]. NIST is presently scheduling LId evaluations every other year. The evaluation task corresponds to the open set detection/verification task. Figure 8.4 summarizes LRE evaluation objectives together with multilingual speech collection efforts in the US. The primary speech data for the evaluations was the multilanguage CALLFRIEND corpus, which is composed of telephone conversations by native speakers from 12 languages collected in North America. The open test corpus also included conversations in Russian, which is not among the 12 target languages. Significant progress of detection performance results could be observed between 1996 and 2003. Since this progress was achieved without significant evaluation focus over the seven years, G. Doddington jested that the *right action to encourage progress is to ignore the problem.* Naturally, speaker recognition evaluations took place in the mean time, and the progress achieved proved to be very beneficial for LId.

Most successful approaches to automatic language identification rely on a statistical modeling framework, requiring appropriate observation sets for model estimation. Multilingual language resource collections are thus vital to LId research.

[2] Many methodological and technical aspects of automatic language recognition are very similar to those of automatic speaker recognition. For details the interested reader is referred to chapter 9 in this book.

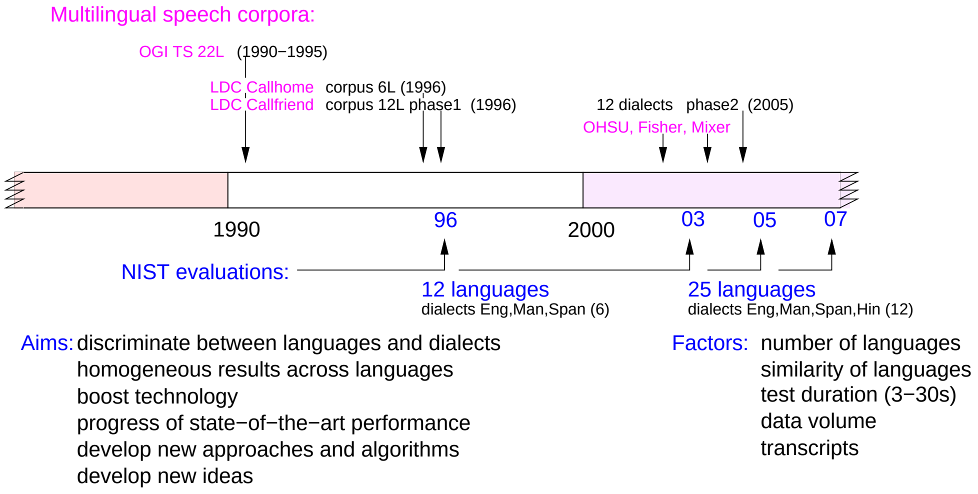
Multilingual speech corpora:
OGI TS 22L (1990−1995)
LDC Callhome corpus 6L (1996)
LDC Callfriend corpus 12L phase1 (1996)
12 dialects phase2 (2005)
OHSU, Fisher, Mixer
1990
96
2000
03
05
07
NIST evaluations:
12 languages
dialects Eng,Man,Span (6)
25 languages
dialects Eng,Man,Span,Hin (12)
Aims: discriminate between languages and dialects
homogeneous results across languages
boost technology
progress of state−of−the−art performance
develop new approaches and algorithms
develop new ideas
Factors: number of languages
similarity of languages
test duration (3−30s)
data volume
transcripts

Figure 8.4. *Chronological representation of multilingual data collection efforts in the US and of language recognition evaluations (NIST LRE). The aims of the language recognition evaluation campaigns are recalled and major factors impacting LId performance are listed.*

8.5. LId resources

Until the early nineties, research in LId had suffered from the lack of common, public-domain multilingual speech corpora to evaluate and compare systems, as well as their underlying models and approaches. In the following section, we will describe major efforts engaged for collecting multilingual speech corpora to support research in the domain of automatic LId. Although Europe and member states are actively involved in multilingual language resource collection efforts [MAR 05], coordinated collections dedicated to LId research are mainly driven by the United States.

OGI_TS 11L and 22L : Efforts to collect large-scale multilingual speech corpora initiated in the early nineties. The first major multilingual speech corpus for LId was designed and collected by Y. Muthusamy and colleagues [MUT 92] of the former CSLU-OGI institute, which is now known as the OHSU-OGI School of Science and Engineering. The public availability of the OGI Multilanguage Telephone Speech Corpus (OGI_TS has significantly contributed to stimulate interest in the field, boosting the development of LId specific approaches. The first OGI_TS corpus contained two hours of speech per language from eleven languages (see Figure 8.1). The speech was collected via domestic telephone calls from native speakers established for years in the United States. For each speaker, read and spontaneous speech was recorded, spontaneous speech consisting in monologue answers to a selection of predefined every-day-life questions. Phonetic transcriptions were manually produced for subsets of the recorded speech in 6 languages. Language-dependent acoustic-phonetic models can hence be trained for these languages and many researchers developed LId systems based on acoustic-phonetic decoding, as well as on phonotactic approaches [ZIS 96a, KWA 95]. In the following years OGI extended the multilingual database to twenty-two languages [LAN 95], adding Italian, Portuguese, Swedish, Polish, Czech, Hungarian, Russian, Arabic, Swahili, Cantonese and Malaysian.

The data distributed for the first LId evaluations of 1996 organized by the National Institute of Standards and Technology (NIST) included the OGI_TS corpus. It was also distributed by the Linguistic Data Consortium (LDC) [LDC 07]. The impact of the OGI corpus was most significant for LId research during the nineties. Later corpora tried to overcome several shortcomings. In particular later multilingual speech collections aimed at gathering larger amounts of data with native speakers, either through international calls connecting US residents with relatives living in their home countries, or by recording calls locally. However, the OGI corpus may become of renewed interest for research on accented native speech, where the focus goes to the impact of long-time L2 practice on L1.

CallHome 6L : Released in the later nineties by the LDC, the CALLHOME corpus was originally collected to support conversational speech recognition in multiple languages: American English, Arabic (Egyptian), Mandarin, Latin American Spanish, German and Japanese. It has also been used for different NIST language recognition evaluations. CALLHOME consists of two hundred international telephone calls per language, each call lasting about half an hour. Along with these hundred hours of speech per language, transcripts have been produced for subsets of the data together with corresponding pronunciation dictionaries, part-of-speech information as well as frequency counts. As CALLHOME conversations connect familiar speakers with their home country, English L2 effects and code switching should be limited. However for each language, differences in telephone networks might leave a signature in the acoustic signal, biasing LId by partial telephone network identification.

CallFriend 12L : The CALLFRIEND collections have specifically been designed and collected for LId research and NIST language recognition evaluations (LRE). They consist of two parts corresponding to different recording phases with different objectives. While the first phase collection [CAL 96], carried out around 1995, was designed for standard LId tasks, the more recent collection phase initiated in 2004 aims at collecting data for more subtle problems, such as dialect and regional variety identification. CALLFRIEND (phase 1) includes twelve languages with a hundred half-hour continental telephone conversations per language. Beyond the six CALLHOME languages (American English, Egyptian-Arabic, Mandarin, Latin American Spanish, German and Japanese), CALLFRIEND includes Canadian French, Farsi, Hindi, Korean, Tamil and Vietnamese. The second phase collection consists of shorter telephone calls (10 minutes per call) recorded as before in the United States and includes more languages (Georgian, Italian, Pundjabi, Russian, Aceh, Amharic, Bengali, Burmese, Chechen, Guarani, Khmer, Lao, Tagalog, Thai, Tigrigna, Urdu, Uzbek) and so-called "dialects" for American English (separated by continental origins: US, Britain, India, Australia), for Arabic, Hindustani and Chinese dialects of Mandarin, Shanghai Wu and Min. This second phase collection named LVDID (Language, Variety and Dialect Identification) [CAL2] data are to be released by LDC after their use in the corresponding NIST LRE evaluations.

SpeechDat : SPEECHDAT corresponds to a series of multilingual speech data collection projects funded by the European Commission for more than ten years now. The aim of SPEECHDAT is to enable the development of multilingual teleservices and speech interfaces. Countries involved include Belgium, Denmark, Finland, France, Germany, Greece, Italy, The Netherlands, Norway, Poland, Portugal, Spain, Sweden, UK... Depending on the collections, focus is put either on acoustic conditions: fixed or mobile telephones, in-vehicle recording SPEECHDAT-M, SPEECHDAT-CAR [SPDM], or on geographical and/or linguistic criteria (SPEECHDAT-E for Eastern European languages, ORIENTEL for Mediterranean and Middle East, LILA for languages of the Asian Pacific area, SALA for Latin America). For example, the ORIENTEL project focused on the development of language resources for speech-based telephony applications across the area between Morocco and the Gulf States. Within the SPEECHDAT project series, the projects are launched by industrial consortia including European and International groups such as Siemens, Nokia, IBM, Sony, Nuance, Scansoft, Microsoft... The speech databases are specified concerning content (application words and phrases, phonetic rich words and sentences), coverage concerning dialectal regions, speaker age and gender and recording devices. Information on SpeechDat projects and corpora can be found on the following WEB site: `http://www.speechdat.org/` and in the proceedings of the LREC (Linguistic Ressources and Evaluation Conference) conferences, which are held every second year since the first conference in Granada, Spain in 1998. They have occasionally been used for LId experiments [CAS 98, MAT 05].

GlobalPhone : The GLOBALPHONE project was launched in the mid nineties at the Karlsruhe University by A. Waibel and T. Schultz [SCH 02]. The goal was to produce similar resources in a large number of languages to enable multilingual speech technologies: speech-to-text, speaker and language recognition. Within this ambitious project, text and speech from the world's major languages were collected, with a focus on European and Eastern-European languages. GLOBALPHONE languages comprise Croatian, Czech, Polish, Russian, Spanish, French, German, Swedish, Turkish, Arabic, Mandarin, Wu (Shanghai), Tamil, Thai, Japanese, Korean, and Brazilian Portuguese. Many different writing systems were experienced and had to be transliterated into a common machine readable format. For each language, news texts were collected via the WEB and recruited speakers read a text composed of about 100 sentences. A hundred adult speakers were recorded in each language using close-talking microphones which resulted in a total volume of 300 hours. Although the GLOBALPHONE corpus seems less appropriate for the development of telephone-speech based LId applications, it may contribute to studying language specificities using controlled read speech, which is generally more carefully articulated than spontaneous conversations.

Fisher : The FISHER corpus collection started in 2002 for the DARPA EARS program [EAR 05], aiming at rich speech transcription and metadata extraction (MDE) in English, Mandarin, Spanish and Arabic. FISHER was designed to collect a large number of short calls from different regions, with speakers dealing with several of a large set of predefined subjects in order to broaden the vocabulary covered by the corpus. The collection was coordinated by a robot operator using a dozen of telephone lines in parallel. For the English language, the recorded Fisher subjects indicated their region of origin using, if appropriate, the major dialectal regions [LAB 04] (North, Midland, South, West) or indicating their accent or country of origin. More than 15,000 calls totaling close to 3,000 hours of speech were thus collected from English speakers and organized according to regional origins. Similarly for Mandarin Chinese, Spanish and Levantine colloquial Arabic, about 300 hours of speech were collected for different dialects and thousands of speakers. Beyond their use for text-to-speech and MDE research, these corpora are available for LId research.

Mixer : The MIXER collection was originally designed for speaker recognition research, with a focus on forensic needs. A new feature here is to collect bilingual and multilingual speakers to measure speaker recognition performance independently of the language being spoken. Channel-independence, which is classically seen as independence towards telephone and recording conditions, includes here the language channel. The MIXER collection started in 2003 and totaled about 5000 speakers in English, but also Arabic, Mandarin, Russian and Spanish by 2005. Each speaker was encouraged to accomplish a large number of calls, either in English or in one of his/her other performing languages.

Table 8.3 summarizes examples of multilingual speech corpora which have been collected since the nineties for language and multilingual speech recognition . These are or will become publicly available via the LDC and ELDA, except the LIMSI-IDEAL corpus which is owned by the French *France-Télécom.* The LIMSI-IDEAL corpus [LAM 98] has been collected according to similar criteria than the OGI corpus. The latter encompasses a relatively high number of languages with all the callers settled in the US, whereas IDEAL comprises only four European languages (French, English, Spanish and German) with a great number of native speakers both calling either from France or from their native countries. Major dialectal regions have been distinguished in the four countries and the speakers have been balanced according to their gender and regions.

Corpus	*Date*	*#L*	*Vol. (h)*	*Speech type*	*Record*
human talks to automated telephone service					
OGI-11*	1992	11	k*10	read & free monologues	domestic tel.
OGI-22*	1995	22	k*10	read & free monologues	domestic tel.
LIMSI-Ideal	1995	4	k*100	read & free monologues	dom./internat.
human reads prompted texts					
GlobalPhone$^\diamond$	1997	30	k*100	read	close-talk
human-human telephone conversations					
CallHome*	1996	6	k*100	spont. dialogues	internat. tel.
Callfriend*	1996	12	k*100	spont. dialogues	domestic tel.
Fisher*	2002	4	k*1000	spont. dialogues (regional accents)	domestic tel.
Mixer*	2004	3	k*100	spont. dialogues (bilingual speakers)	domestic tel.

Table 8.3. *Examples of (*publically available via LDC, $\diamond$publically available via ELDA) multilingual speech corpora. For each corpus, the number of languages is given in the #L. column. Dates approximately correspond to the start of data collections and total volumes are indicated by orders of magnitude.*

Ever-growing efforts to collect huge amounts of speech corpora in multiple languages highlight the increasing importance of multilingual speech technologies, among which language, dialect and accent identification represent key issues. An ISCA special interest group on speaker and language characterization was launched [BON 01] in 1998. The Odyssey workshop focuses on research in speaker and language recognition and takes place every other year. Furthermore, papers on spoken language, dialect and accent recognition can be regularly found in major speech processing conferences (, ,).

8.6. LId formulation

The problem of language identification can be approached with the help of mathematical formulation, which makes it possible to decompose the problem into simpler or more focused sub-problems. Let X denote the acoustic evidence (speech) on the basis of which the LId decision is to be taken. Without loss of generality, we can make the assumption that X be a sequence of symbols from a finite alphabet $\mathcal{X}$. With a statistical approach [JEL 98] the LId problem can be stated as follows:

$$L^* = \underset{L \in \mathcal{L}}{argmax}\ P(L|X) \tag{8.1}$$

L^* being the identified language, X the symbol sequence of the speech sample, $\mathcal{L}$ the set of potential languages, and $P(L|X)$ the probability of language L given X. Applying the well known Bayes' formula of probability theory to equation (8.1), the LId problem can be reformulated as follows:

$$L^* = \underset{L}{argmax}\ P(X|L)\, P(L) \tag{8.2}$$

where $P(X|L)$ stands for the probability that X is observed, when language L is spoken and $P(L)$ is the a priori probability of language L. With an equiprobability assumption for the different languages, the formula can be simplified as:

$$L^* = \underset{L}{argmax}\ P(X|L) \tag{8.3}$$

Written as such, the formula is not yet of great help, as language L is considered as a whole. It merely evokes an acoustic approach to LId, relying on some language-specific acoustic models. However, depending on multilingual resources available for language-specific model training (see Figure 8.2), various decompositions can be proposed. Different LId approaches then correspond to different ways of decomposing $P(X|L)$ and all through the LId litterature, proposed approaches and available language-specific resources are tightly coupled.

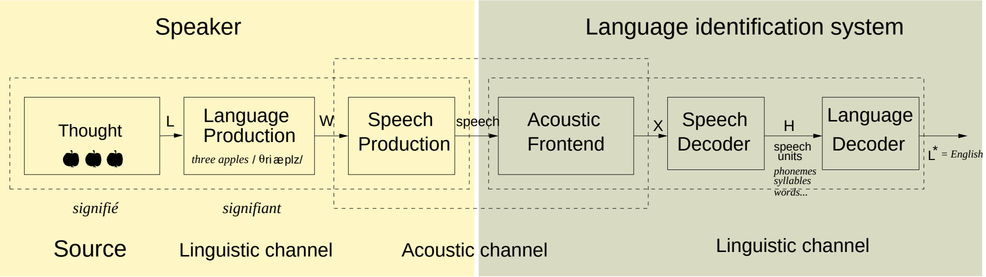

Figure 8.5. *Multi-lingual source-channel model of language recognition and multi-lingual speech recognition.*

Figure 8.5 gives a tentative source-channel model of the LId problem, adapted after the corresponding model of speech recognition proposed by F. Jelinek [JEL 98]. Even though the involved cognitive processes are certainly more intricate and complex than depicted in Figure 8.5, it can be instructive to examine the LId problem within this extended source-channel model. A "thought" is to be encoded in one of various different linguistic surface forms, with variable word choices for a given language and across languages. The left (generation and encoding) side illustrates the different language-dependent formulation choices for a given language-independent message. Using Saussure's terminology [SAU 15], the language-independent *signifié* is associated with a language-dependent *signifiant*. For example, given a *signifié* of three apples, if the message is to be implemented by a French *signifiant*, the generated word string would be most probably W_{fr}=*trois pommes* with the corresponding phonemic sequence /tʁwa#pɔm/. However in English the *signifiant* would most likely become W_{en}=*three apples* with an underlying phonemic representation /θri#æpəlz/. The right side of Figure 8.5 represents the decoding process with an acoustic front-end converting the acoustic waveform into an acoustic feature stream X which is then processed by a "speech decoder" (acoustic decoder) to produce a hypothesis stream H. The H stream enters the "language decoder", which outputs the most likely language identity and, if desired, the most likely corresponding hypothesis: some unit sequence, which may be of linguistic type (such as phoneme, syllable or word sequences) or of acoustic type (such as acoustic segment or Gaussian label sequences). Various LId implementations have been proposed for the decoding block corresponding to variants of the mathematical formulation of the LId process developed hereafter.

Following the source-channel model and related work on automatic speech and language recognition [GAU 04], $P(X/L)$ can be more generally rewritten as:

$$L^* = \underset{L}{argmax} \sum_{H} P(X|H, L, \Lambda) P(H|L) \tag{8.4}$$

where $P(X|H, L, \Lambda)$ corresponds to the speech decoder and $P(H|L)$ to the language decoder (in Figure 8.5). Λ is the acoustic model and H a stream of symbols (e.g. phoneme, syllable or word sequences).

The LId problem has thus been formally divided into two blocks. The acoustic decoder can be seen as a speech tokenizer, generating flows of weighted symbols H (either language-dependent or language-independent). The language decoder associates additional weights to the H symbol flow for each language L. The language decision may rely on the language decoder only, the acoustic decoder then plays merely the role of a speech tokenizer considering acoustic and symbolic information simultaneously.

In general, the first term reflecting the acoustic decoder is implemented by approximating the summation over the whole hypothesis space by the best hypothesis, resulting in the one-best H_L^* sequence. This approximation corresponds to:

$$\sum_{H} P(X|H, L, \Lambda) \approx \underset{H}{max}\ P(X|H, L, \Lambda)$$

which then results in:

$$L^* \approx \underset{L}{argmax}\, \underset{H}{max}\ P(X|H, L, \Lambda) P(H|L) \tag{8.5}$$

The acoustic decoder may nonetheless contribute to the language-dependent scores, either by weighting via the achieved acoustic likelihoods [GAU 04] or by decoding with language-specific acoustic phone models [LAM 94]. Numerous research contributions aim at minimizing the dependence of the acoustic model (Λ) on the language: it is not required that for each language a specific acoustic model decode the acoustic input stream X. In the following equation, the speech decoder becomes language-independent, the language-dependent information is captured by the language decoder:

$$L^* \approx \underset{L}{argmax} \max_{H} P(X|H, \Lambda)P(H|L) \tag{8.6}$$

Let $H^* = argmax_H P(X|H, \Lambda)$ be the one-best hypothesis of a language-independent speech decoder, then the LId problem may be rewritten as:

$$L^* \approx \underset{L}{argmax} \; P(H^*|L) \tag{8.7}$$

This formulation of the LId problem globally corresponds to the well known phonotactic approach (see below). Language-independent acoustic models may entail lower phone (or other relevant units) recognition accuracy. It is known that the decoded H stream accuracy correlates with LId accuracy [LAM 95], and the trade-off between language-independent vs language-specific acoustic models is a challenging research issue. In order to take into account the hidden nature of the speech symbols and to increase the amount of information contained in H, limited by the approximation of the sum by the max operator in equation 8.5, Gauvain et al. [GAU 04] have recently proposed to take into account hypothesis lattices instead of single one-best hypotheses.

$$L^* = \underset{L}{argmax} \sum_{H \in lattice} P(X|H, L, \Lambda)P(H|L) \tag{8.8}$$

The arcs of the lattice are labeled by the decoded speech symbols and may be weighted by the corresponding acoustic likelihoods. A phonotactic lattice-based approach can then be written as

$$L^* = \underset{L}{argmax} \sum_{H \in lattice} P(H|L) \tag{8.9}$$

The benefits of this approach [GAU 04, SHE 06, ZHU 06] have been demonstrated in terms of LId accuracy and of computational efficiency.

8.7. LId modeling

Beyond language-dependent models, successful LId systems also include effective acoustic feature extraction and language decision modules. In state-of-the art systems [CAM 06, SHE 06, GAU 04, MAT 06], the $argmax_L$ operator is implemented, not just as a simple max picking, but with the help of specialized linear or non-linear classifiers. Figure 8.6 shows a block diagram of the major components of an LId system, where the “LId Scoring” block includes both speech and language decoding blocks of Figure 8.5. The acoustic front-end aims at extracting appropriate feature vectors. The decision module (back-end classifier), if appropriately optimized, contributes to significant LId accuracy gains [ZIS 97].

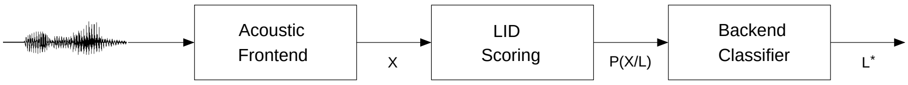

Figure 8.6. *Block diagram of the main modules of an LId system: acoustic front-end, language scoring and a back-end classifier.*

If only multilingual audio data are available, language recognition has to rely on mere acoustic properties. Such approaches have successfully been explored using Gaussian Mixture Models (GMM) [ZIS 93, TOR 02a, TOR 04, WON 00] and support vector machines (SVM) [CAM 04, CAM 02]. Additional resources may be required for higher level modeling: phonetic or phonemic labeling and segmentation, transcriptions, pronunciation dictionaries, morphological information, prosodic annotations... Different types of language-specific information can then be modeled ranging from mere acoustics to more sophisticated linguistic levels, including phonemics, information on vowel systems [PAR 99, PEL 00], phonotactics, morphology and prosody). As seen in the previous section on resources, multilingual audio data may be accompanied by orthographic transcriptions or phonetic segmentations (e.g. OGI-TS corpus). These make it possible to train language-specific acoustic phone models or to estimate phonotactic constraints for each language. Standard speech recognition techniques can then be applied to the LId problem. Introducing the highly informative lexical level may result in complex systems, potentially equivalent to multilingual transcription systems. Although this approach certainly guarantees the most reliable identification results, it is at the expense of high development and operating costs.

Acoustic-phonemic and phonotactic modeling approaches raise the question of the phone sets to be used in a multilingual context. For automatic LId using phonotactic constraints, either multiple language-dependent phone sets are used by recognizers in parallel, or a single global phone set [HAZ 97, COR 97], or even a combination of both [ZIS 96a, MAT 99] have been implemented. Defining a single global phone set is an interdisciplinary research issue of its own, which ranges from phonetic and phonemic domains to multilingual speech recognition [IPA 99, ADD 03, ANT 04, SCH 02]. LId system development costs are closely linked to the number of mandatory language-specific resources, hence the success of purely acoustic modeling techniques, such as GMMs, SVMs and unsupervised phonotactic approaches [ZIS 96a, LUN 96].

8.7.1. *Acoustic front-end*

Automatic LId systems generally make use of the same acoustic features as used for speech and speaker recognition, namely MFCC(mel frequency cepstral coefficients) [DAV 80],[CHI 77] or PLP (perceptual linear prediction) features [HER 90], [HON 05]. Feature vectors (typically 10-15 parameters) are computed at a fixed rate (e.g. 10 ms). First (and second) order derivatives are generally computed from the MFCC or PLP vector flow in order to better take into account the dynamic properties of speech sounds. The dimension of the resulting feature vectors then typically ranges from 20 to 30 parameters and they capture dynamic information of a 50-80 ms time span. Some attempts have been made to develop specific acoustic front-ends for language recognition [DUT 00]. An improved feature set called *shifted delta cepstra* (SDC, see Figure 8.7) [BIE 94], was made popular by Torres-Carrasquillo et al. [TOR 02b] within a Gaussian mixture model approach. SDC vectors are an extension of delta-cepstral coefficients: they consist of stacked delta vectors of typically 7 consecutive deltas. The temporal scope of such vectors are around 250 ms, including information of at least one syllabic unit. Such long-span features allow for an implicit linguistic unit modeling and they happen to be not only language-specific, but also corpus and topic specific, the most frequent syllables being best represented. These features are hence very effective, if there is no mismatch between training and test data. Other recent studies to improve feature vectors, based on split temporal context and neural nets are described in [MAT 05].

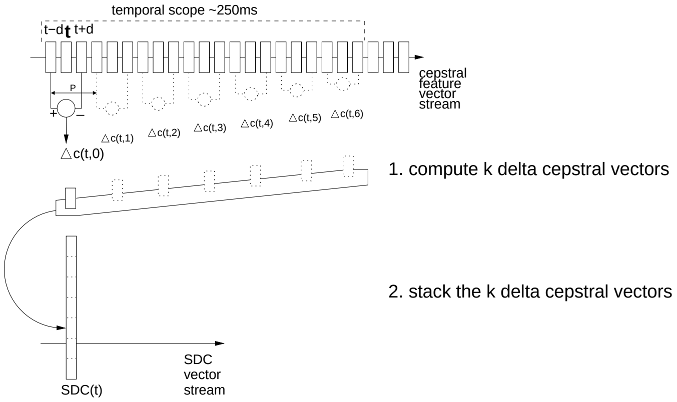

Figure 8.7. *Computation of a SDC (shifted delta cepstral) vector obtained by stacking k=7 consecutive delta cepstral vectors (deltas use +/- d=1 vectors around t), shifted by P=3.*

8.7.2. *Acoustic language-specific modeling*

Purely acoustic approaches to LId only require language-specific audio data for each of the considered languages. The advantage is that no language specific knowledge (generally linguistically informed transcriptions) is needed. Hence LId system development and further extensions to additional languages become straightforward. Early acoustic based approaches included filterbanks and LPC based approaches [LEO 74, CIM 82] as well as vector quantization based approaches [FOI 86, SUG 91]. Nowadays, Gaussian mixture models (GMM) and support vector machines (SVM) are the most commonly used approaches for acoustic modeling.

Gaussian mixture models

Gaussian mixture models (GMM) are the most popular acoustic approach to LId, in particular for the language detection task. A GMM model is estimated for each language and the only prior knowledge to train the language-specific models consists in the language identity of the audio corpus. Figure 8.8 gives a schematic representation of a Gaussian mixture model. The Gaussians, estimated from the acoustic feature vectors produced by the front-end, attempt to model the entire acoustic space. Equation 8.10 gives the acoustic likelihood $P(X|L)$ computation of the Gaussian mixture model G based on N Gaussians.

$$P(X|L) = \prod_{t=1}^{T} P(x_t|L) = \prod_{t=1}^{T} (\Sigma_{n=1}^{N} w_n G_n(x_t|L)) \tag{8.10}$$

N typically ranges from 64 to 1024. Figure 8.8 illustrates a GMM with N=5 Gaussians. As a general rule, GMMs do not tend to capture temporal dependencies very well, hence the introduction of SDC acoustic features. State of the art performance

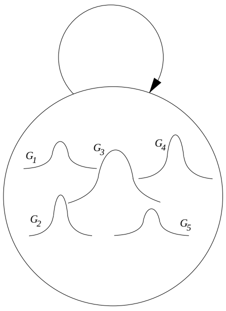

Figure 8.8. *Schematic representation of a Gaussian mixture model with 5 Gaussians representing the acoustic space of a spoken language.*

was achieved using Shifted Delta Cepstra (SDC) feature vectors [TOR 02b]. These excellent language recognition performances [TOR 04, TOR 02b, WON 02] have established GMMs as a major LId approach in addition to the successful PPRLM (parallel phone recognition followed by language modeling) [ZIS 96a].

GMM models, generally employed to measure the acoustic adequacy between the input X and the language L, can also be used to act as speech tokenizers [TOR 02a], i.e. converting the acoustic vector stream into a discrete label stream. This is done by replacing each vector x_t of X by the label of the Gaussian which produces the highest contribution in equation 8.10. This then makes it possible to apply PPRLM approaches to the LId problem, replacing the linguistic tokenizing units (phones, syllables) by acoustic vector units.

Support Vector Machines

Support vector machines (SVM) [VAP 97], a popular tool for discriminative classification [CRI 00, COL 01], have recently been introduced for language recognition [CAM 04]. The assumption here is that two languages are separable by a hyperplane, provided an appropriate acoustic feature space is used for the spoken language representation.

A support vector machine (SVM) is a two-class classifier constructed from sums of kernel functions. The SVM training process aims at modeling the boundary between the two classes as opposed to the generative Gaussian mixture models which represent language-dependent probability densities. SVMs are hence particularly adapted to language detection and verification problems (yes/no decision). Extensions to multiclass classification can be implemented as several two-class SVMs in parallel.

Figure 8.9 compares SVM and GMM results as DET curves [MAR 97] achieved by Campell et al. [CAM 04] on NIST 2003 evaluation data (Callfriend corpus, 12 languages). LId improvements correspond to a DET curve shift towards the origin (left, bottom). The curves indicate that SVMs perform slightly worse than GMMs, but SVMs carry complementary information, as the fusion of both methods improves the overall performances. Furthermore, as SVMs have been introduced for LId only recently, their potential might not be fully exploited yet and additional improvements could be expected in the future. Beyond SVMs' acoustic classification capacities, SVMs have also served as back-end classifiers of an LId system [WHI 06].

8.7.3. *Parallel phone recognition*

In contrast to the acoustic modeling approaches, LId systems based on parallel phone recognition (PPR) require an important amount of language-specific knowledge and resources, namely phone inventories and correspondingly labeled acoustic

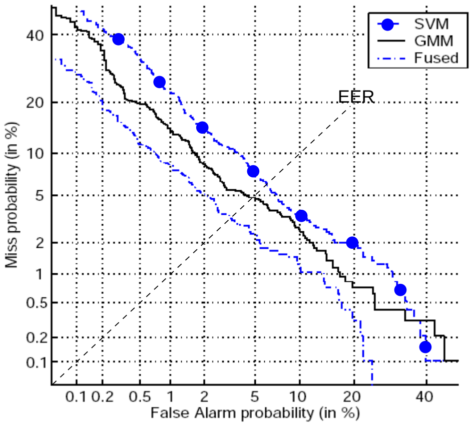

Figure 8.9. *DET curves of LId results achieved on NIST LRE 2003 test data using GMM, SVM and fusion of both (after W. M. Campbell et al. MIT-LL [CAM 04]).*

training data. However, these resources generally exist for the world's major languages, and in particular those for which automatic speech recognition systems have been developed.

In the early nineties, language-specific phone decoders in parallel were a popular approach to automatic LId [YAN 96b, MUT 93, LAM 94, ZIS 94]. The assumption here is that different languages make use of different sound inventories and, even though part of these sounds are shared among languages, their spectral realizations as well as their occurrences might differ significantly from one language to another. These differences may be taken into account by language-specific sets of phone-based hidden Markov models (HMM) and researchers have taken advantage of the progress accomplished in speech recognition to investigate HMMs for LId [UED 90, NAK 92, LAM 94, ZIS 94]. Figure 8.10 representing a phone-based HMM model illustrates the additional modeling capacities as compared to the acoustic GMM approach: improved temporal modeling (three left-to-right GMM states); different GMMs for different states and sounds (φ). The number of Gaussians per state remains limited (typically between 16 and 32) as compared to the acoustic GMM approach described before. A language can then be viewed as a source of phonemes, modeled by a fully connected Markov chain. Its higher level structures are approximated by phonotactic constraints. Figure 8.11 then shows a schematic representation of such a parallel phone recognition approach, termed PPR in Zissman's terminology [ZIS 96a]. PPR systems correspond to implementations of equation 8.5 [LAM 94, ZIS 94]. For each language L_i a phone inventory ($\{\varphi_{L_i}\}$) is defined and a set of acoustic phone models (AM_{L_i}) is

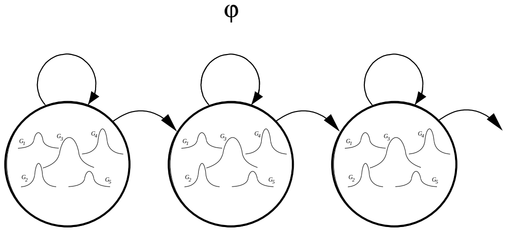

Figure 8.10. *Schematic representation of a phone-based HMM model representing the acoustic realizations of a given phoneme φ. Each state corresponds to a GMM (5 Gaussians per state).*

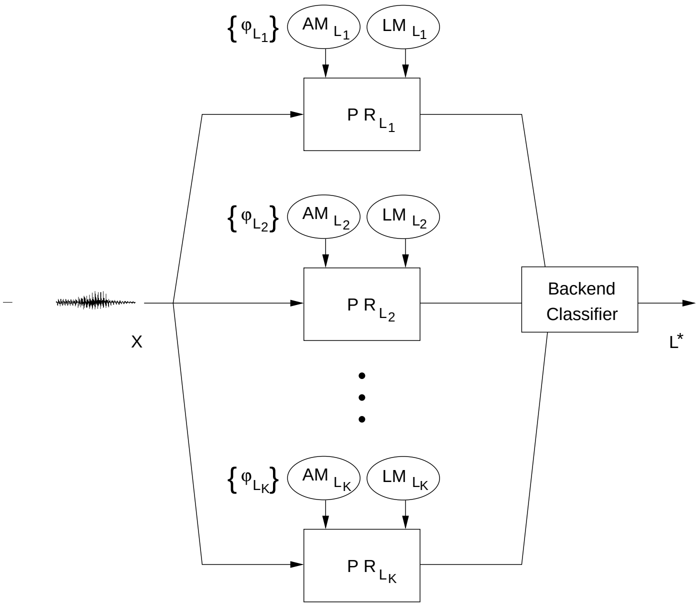

Figure 8.11. *Schematic representation of a LId system based on language-dependent phone recognizers (PR) in parallel: the PPR (parallel phone recognition) approach.*

trained from appropriately labeled acoustic training data. Phonotactic language models (LM$_{L_i}$), generally bigrams or trigrams representing phone cooccurrence specificities, are also estimated from such labeled training data. The likelihood scores resulting from PPR incorporate both acoustic and phonotactic information, the latter contributing to improved automatic phonemic transcriptions. The scores are then used by the back-end classifier to determine the identity of the language being spoken. The PPR approach can be implemented using many variants: the phone sets corresponding to the basic modeling units can be changed. The sets can be extended to include longer units such as syllables [NAG 06] or words [MAT 98]. Articulatory features may be used instead of linguistic units [KIR 02]. Acoustic HMM models may or not depend on gender, phonemic context or channel conditions and the complexity of language model constraints may be differently tuned to fit the amount of training data available. The PPR approach is expensive both in terms of implementation and of computation resources in operating mode. However it yields excellent identification results, in particular for shorter durations, as acoustic, phonemic and phonotactic information levels jointly contribute to the decision. A serious bottleneck of the approach is the extension to minority languages and spoken varieties for which the required language-specific knowledge about phonemic systems and pronunciations as well as the segmented and labeled audio data are lacking.

To address these shortcomings, a smart modification to the PPR approach [ZIS 94] exploits the fact that phonotactic constraints for one language can be estimated based on **automatically decoded phone streams** using the acoustic models of a different language, even if the quality of the produced phone label streams certainly differ from a labeling achieved by a human phonetician. The advantage is then, that phone recognizers need not be developed for all the target languages. This idea has led to the most successful parallel phone recognition followed by language modeling (PPRLM) approach, which will be developed in the *phonotactic modeling* paragraph hereafter.

8.7.4. *Phonotactic modeling*

Similarly to the PPR approach, the phonotactic approach views a language as a source of phonemes (or some other structuring units), modeled by a fully connected Markov chain. The phonotactic approach addresses the LId problem using the formulation of equation 8.7. In contrast to the PPR approach, the acoustic decoding component no longer influences the language score directly. The acoustic decoder merely serves as a speech tokenizer, producing a discrete symbol sequence H^*. The first automatic LId systems using phonotactic constraints date back to the early eighties [LI 80]. They exploit the findings of House and Neuburg [HOU 77], who experimented LId with hand-labeled broad phonemic classes corresponding to phonemic transcriptions of texts.

Figure 8.12 gives a simple overview of language identification using a phonotactic approach. Both the language-dependent phonotactic models (LM$_{L_i}$) estimated from automatically decoded language-specific training data, and the test sequence X to be identified depend on the accuracy of the acoustic model (AM) and phone recognizer (PR) [LAM 95, MAT 05]. The implementation of the speech tokenizer raises

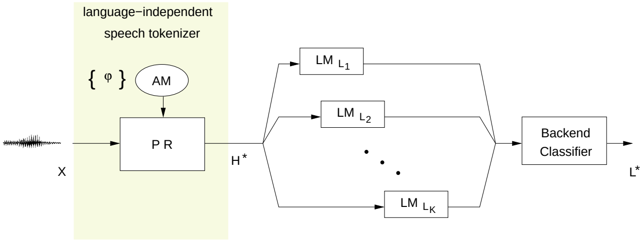

Figure 8.12. *Automatic language identification based on a phonotactic approach (PRLM: phone recognition followed by language modeling).*

many challenging research issues, in particular concerning the choice of a "language-independent" symbol inventory $\{\varphi\}$. Such inventories may be limited to broad phonetic classes which are shared among languages [HOU 77, MUT 93], IPA-like inventories [COR 97, ZHU 05, NAV 01] or the union of several language-dependent phone sets. The size of the inventory may then range from ten to several hundreds. As a general rule, results increase with the size of the phone inventory $\{\varphi\}$, provided that the amount of available training data is appropriate. Instead of using phone-like units, longer units representing syllables or sub-words have been experimented [ANT 04, ZHU 06, TUC 94, MAR 06], with the objective of achieving higher decoding accuracy and hence better language identification. However, longer units are generally performing less well. A potential explanation is that the training data remain insufficient, as the number of units drastically increases (thousands) as compared to phone-like units (tens). A set of units may also be automatically determined from a multilingual speech corpus using unsupervised clustering techniques. Whatever the approach and the resulting set of acoustic models, the speech tokenizer tends to best represent the languages included in the training data.

The phonotactic constraints are approximated by n-gram language models (LMs). Standard smoothing techniques have been shown to work well for LId [SIN 03]. The orders of the LMs are generally limited to bigrams and trigrams. Important performance gains are observed for trigrams (30-40% relative) as compared to the lower order model. Fourgrams or even higher orders prove to remain globally inefficient in test conditions. Figure 8.13 (left) illustrates performance differences for bigrams and trigrams achieved by the MIT Lincoln Lab [SHE 06] in the NIST 2005 evaluation.

An important improvement to the phonotactic approach consists in replacing the 1-best hypothesis H^* of the speech tokenizer by a hypothesis lattice (see equation 8.9) [GAU 04], both for LM training and for the operational phase. This results in more reliable LMs, and more information can be extracted from the unknown test segment for automatic identification. Figure 8.13 (right) shows the achieved gains with a lattice-based approach and PPRLM systems.

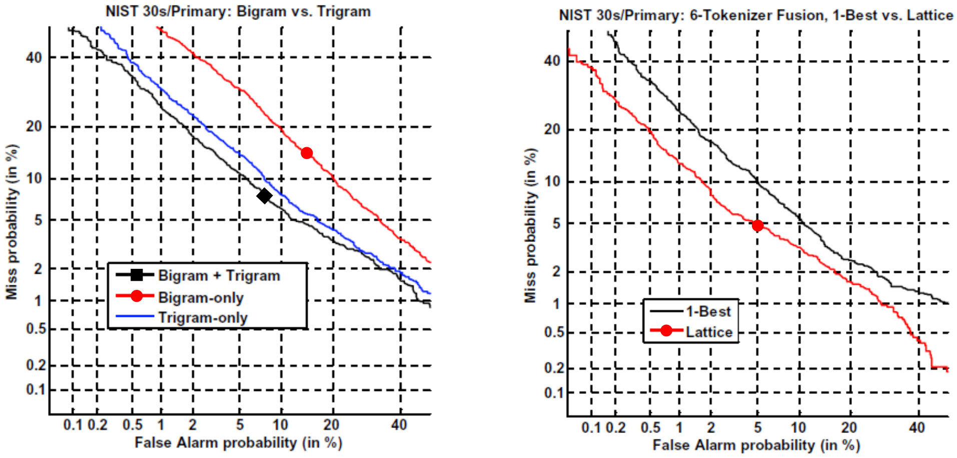

Figure 8.13. *DET curves illustrating improvements to the phonotactic approach.* **Left***: phonotactic constraints implemented as trigrams vs bigrams.* **Right***: standard 1-best versus lattice-based PPRLM approach (Figures after W. Shen at al. MIT-LL [SHE 06]).*

Parallel phone recognition followed by language modeling

Instead of progressively enhancing models, methods and techniques to optimize the performance of a given PRLM system, LId results can be improved by running several different PRLM systems in parallel. This corresponds to the most popular PPRLM (parallel phone recognition followed by language modeling) approach [ZIS 94, YAN 96a, ZIS 96a] as represented in Figure 8.14. The same decision rule formulation as for PRLM applies (equation 8.7). However, for multiple recognizers, care must be taken to apply the decision rule across LM scores from different recognizers with bias normalization.

The PPRLM approach is the most efficient *stand-alone* method [SHE 06]. During the LRE05 evaluation the best MIT-LL subsystem was based on PPRLM with a 4.9% EER (see Figure 8.13 right), whereas the best system (combination of GMM, SVM and PPRLM) was below 3%.

In contrast to the PPR approach, language-specific resources are required only to train the speech tokenizers. An arbitrary large number of languages may potentially be identified, exclusively relying on audio data to train the language-specific models. To avoid using prior knowledge corresponding to linguistic unit sequences, Torres-Carrasquillo and colleagues [TOR 02a, TOR 02b] proposed to use H^* tokens obtained via GMMs, resulting in a “phonotactic” approach requiring exclusively language-specific audio data.

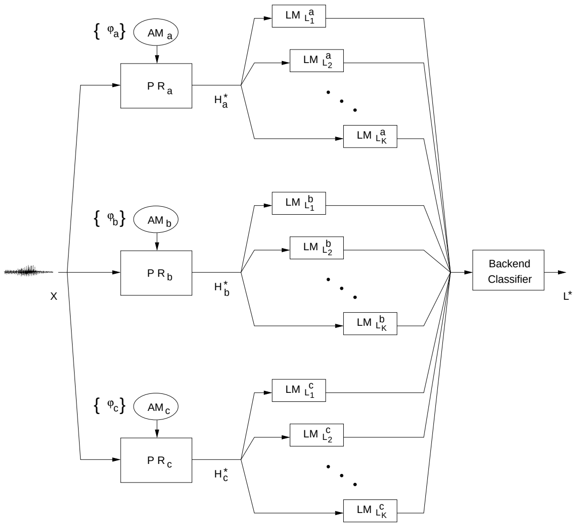

Figure 8.14. *Automatic language identification with a PPRLM approach with 3 phone recognizers (a,b,c) in parallel.*

Prosodic modeling

Beyond different sound inventories and sound combinations, different languages may also be characterized by different intonation or rhythm patterns, for which measurable acoustic correlates are: fundamental frequency $f0$ variations, energy and duration [RAM 99, FAR 01, ROU 05]. The importance of prosodic information in recognizing speech or in discriminating between languages has long been acknowledged. In section 8.3, we have seen that humans are able to identify languages, although with limited accuracy, based on prosodic information [NAV 01, RAM 99]. However, this information is often ignored in LId systems. Some early studies including prosodic information to automatic LId examined pitch variation, duration and syllabic rate with marginal success [MUT 93, THY 96, HAZ 93, HAZ 97].

In the NIST evaluations, prosodic information is largely ignored by most competing systems. However, renewed interest in prosodic cues for LId can be found in recent study on prosody [ADA 03, OBU 05] for LId. In [ADA 03], the authors make use of temporal trajectories of prosodic cues, i.e. fundamental frequency and short-term energy to segment and label the speech signal into a small set of discrete units. A prosodic labeling system is derived which distinguishes 10 different classes, depending on segment duration (2 classes) combined with either classes of falling and rising $f0$ and energy (4 combinations) for voiced segments and an unvoiced class. The number of modeled classes thus remains much lower than in more traditional phone-based or GMM approaches. The approach is evaluated using the NIST Language Identification task and achieves promising results even though significantly worse than the more standard approaches. It demonstrates however that the prosody dynamics can capture language-dependent information.

System combination and further LId extensions

System combination, exploited for years within the PPRLM approach, tends to emerge as a standard practice for present LId systems. The fusion of heterogeneous systems makes it possible to efficiently cope with individual systems' limitations. Whereas the best NIST LRE 1996 systems made use of PPRLM approaches, in 2003 and 2005 the trend was to combine different approaches in parallel [MAR 03, NIS 05]. In 2003, the best performance ($<$ 3% EER on 30 second segments) was achieved by fusing the scores of phonotactic PPRLM, acoustic GMM (Gaussian mixture models) and discriminative SVM (support vector machines) based systems, with a decoding time of roughly 15 times real-time. Therefore, gains were mainly due to an optimal combination of different known approaches, rather than very significant progress for one specific approach. Heterogeneous system combinations recall the hypothesis of multiple partly redundant information levels with respect to language identification previously formulated for human perception.

As for prosodic modeling, many different attempts have been made to extend phone-based approaches to exploit syllable-level knowledge to LId and dialect identification [LI 94, BER 98, BER 99, MAR 06, ANT 04, ZHU 06]. Adopted approaches range from syllabic spectral features [LI 94], to syllabotactic modeling together with multilingual syllabic inventories [ANT 04, ZHU 06], as well as syllable-like phone triplets [MAR 06]. Whereas these approaches are individually less efficient than optimized phone-based PPRLM systems, their fusion with already existing systems generally tends to achieve perfomance gains. Including lexical information into LId systems [HIE 96, MEN 96, SCH 96, MAT 98] has also been proposed. However such approaches, beyond being extremely resource-greedy, are also difficult to extend to any additional spoken language.

8.7.5. *Back-end optimization*

LId results can be significantly improved by an optimized back-end decision module. The simple maximum picking used in the earlier PPRLM systems can be replaced by linear or non-linear combinations of different scores, generally normalized log-likelihood (LLR) scores, stemming from multiple, often heterogeneous LId scoring modules [YAN 95b, YAN 96b, ZIS 97, GAU 04]. Score normalization allows to reduce the bias created by different LId scoring modules. The impact of back-end optimization on LId results may range from very low up to 40% relative improvement.

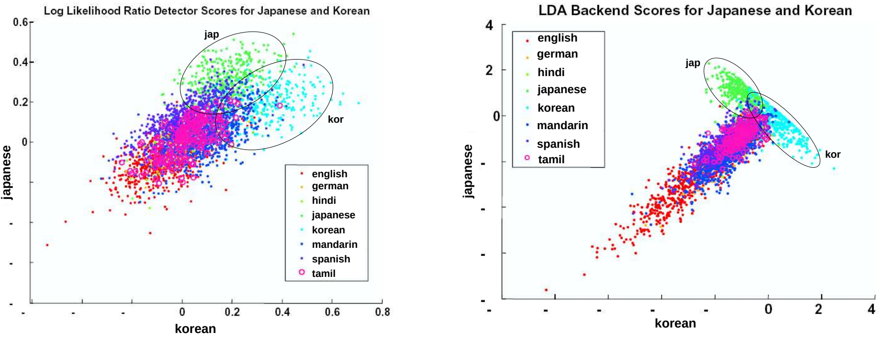

Figure 8.15. *Figures after W. Shen at al. MIT-LL [SHE 06].*

For example, a comparative analysis of the different MIT-LL systems developed for the NIST LRE 2005 evaluations [SHE 06] revealed important gaps in achieved accuracy for given back-ends depending on the LId system configurations (standard vs lattice-based PPRLM). The authors provide a thorough analysis of different back-end classifiers, and Figure 8.15 illustrates these differences by plotting the correlation between Japanese and Korean language detector scores for the target and impostor languages' test samples. In the left plot, LLR scores are used, whereas the right plot is achieved by applying a linear discriminant analysis (LDA)-based Gaussian back-end classification. The high cross-class correlation observed in the left plot appears to be significantly reduced in the right plot, thus allowing for improved detection capacities. The difference between both back-ends remains low for the standard PPRLM approach, but it proved to become very efficient for the lattice-based PPRLM system.

8.8. Discussion

Most successful approaches to automatic LId rely on a statistical modeling framework, requiring appropriate observation sets in order to estimate models accurately. The progress in LId research over the last twenty years has been strongly linked to the production and the public availability of multilingual speech corpora. OGI has provided an important pioneering contribution with their multilingual telephone speech corpus, collected and distributed in the early 90s.

A number of different approaches have been implemented for LId making use of various degrees of language-specific information, ranging from purely acoustic ones, such as GMMs [ZIS 93, TOR 02a, TOR 02b, WON 02] and SVM [CAM 04], to lexically informed systems [ITA 93, MAT 98, SCH 96]. Phonotactic approaches have achieved a very interesting trade-off between LId accuracy, implementation ease and cost: competitive results together with very low classification times; the extension to additional languages does not require any language-specific resources, besides audio, as unsupervized phonemic labeling can be carried out.

Many of the approaches applied to spoken LId are derived either from speech recognition, or from speaker recognition systems. Whereas the best 1996 LId systems made use of PPRLM approaches (banks of parallel phone tokenizers followed by phonotactic language modeling), in 2003 the trend was to combine different approaches in parallel, fusing the scores of phonotactic PPRLM, acoustic GMM (Gaussian mixture models) and discriminative SVM based systems. Thus, gains were mainly due to an optimal combination of different known approaches, rather than to a new approach or to an important advance of one single approach. An important exception might be the lattice-based approach [GAU 04], which demonstrated significant improvements to PPRLM, achieving similar results as the best MIT-LL system combination, with a significantly lower decoding time.

Independently from the adopted approach, the decision accuracy largely depends on the stimulus length (in general a maximum duration of several tens of seconds). Results are much better for 30-second excerpts than 3-second ones, and phonotactic approaches tend to need longer speech samples than approaches including acoustic-based scores. As human identification performance behaves asymptotically for much shorter durations of speech utterances, an important amount of information is still lacking in current approaches. This gap on short speech excerpts may be progressively bridged by improving existing methods, by using a larger variety of acoustic parameters, by combining more and more basic LId system variants in parallel, thus producing richer input to optimized back-ends, and last but not least, by using larger audio corpora for model estimation. The gap between humans and machines may also be related to under-represented or missing information levels, such as prosody, morphology and voice quality or to a lack of discriminating capacity between fine acoustic-phonetic differences. The comparison of machine and human performances is informative to gain insight into the achieved modeling accuracy and to guide future research. Perception experiments suggest that humans exploit multiple partly redundant information levels to achieve language identification or at least a broad classification. Heterogeneous system combinations can then be considered as an implementation of this human capacity, rather than as a second best implementation of the complex LId problem.

Today, automatic LId systems are able to process a large number of languages (typically several tens) and this is certainly more than an over-average gifted human being. Automatic LId then entails the question of how to define a language, and how to distinguish between variations within and across languages. Present research efforts increasingly focus on subtler identification problems, such as dialect and accent identification. NIST LRE evaluations have started addressing dialect identification since 2005 and appropriate audio data are still being collected (Callfriend2-LVDID [CAL2]). The labeling of these speech data in terms of language, dialect and accent may be problematic, and may interact with measured identification errors. Detailed analyses of the 2005 NIST evaluation [NIS 05] revealed problems due to over-simplified labels for Chinese dialects and due to Indian-accented English often misclassified as Hindi. Accented speech collections are also good candidates to include a non negligible part of code mixing and code switching [NIE 06], representing challenges to LId research for the forthcoming years. In the future, research will increasingly address dialects, regional varieties, and foreign accents, the challenge for the latter being the identification of both the spoken language (L2) and the speaker accent, related to his/her native language (L1). These issues will benefit from accurate acoustic-phonetic models to improve automatic transcription of accented speech [BAR 07], and more generally rich transcription in a multilingual framework.

Further research directions include language identification of audio-visual and multimodal speech [BEN 00]. Beyond the information delivered by human faces, language-specific information might be found in visemes and in accompanying gestures, potentially contributing to a multimodal LId component. The results of LId research, beyond multilingual indexing, public speech-driven services and intelligence applications, may cross-fertilize other research domains, such as foreign language acquisition/training, corpus-based language typology and characterization as well as dialectal studies.

8.9. Bibliography

[ADA 03] Adami A., Hermansky H., "Segmentation of Speech for Speaker and Language Recognition", *Proc. of Eurospeech-03*, Geneva, September, 2003.

[ADD 03] Adda-Decker M., Antoine F., Boula de Mareüil P., Vasilescu I., Lamel L., Liénard J.S., Vaissière J., Geoffrois E., "Phonetic Knowledge, Phonotactics and Perceptual Validation for Automatic Language Identification", *Proc. of ICPhS-03*, Barcelona, 2003.

[ANT 04] Antoine F., Zhu D., Boula de Mareüil P., Adda-Decker M., "Identification des langues par unités phonémiques et syllabiques", *Proc. of JEP-04*, Fès, Morocco, 2004.

[BAR 99] Barkat M., et al, "Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects", *Proc. of Eurospeech-99*, Budapest, Hungary, September 1999.

[BAR 07] Bartkova K, Jouvet D., "On using units trained on foreign data for improved multiple accent speech recognition", *Speech Communication*, Vol.49, **10-11**, pp.836-846, 2007.

[BEN 00] Benoît C., Martin J.C., Pelachaud C., Schomaker L., Suhm B., "Audio-visual and multimodal speech-based systems". In D. Gibbon, I. Mertins, R. Moore (Eds.), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*, pp.102-203, Boston Kluwer Academic Publishers, 2000.

[BER 98] Berkling K., Zissman M., Vonwiller J., Cleirigh C., "Improving Accent Identification Through Knowledge of English Syllable Structure", *Proc. of ICSLP-98*, Sydney, 1998.

[BER 99] Berkling K., Reynolds D., Zissman M., "Evaluation of Confidence Measures for Language Identification", *Proc. of Eurospeech-99*, Budapest, Hungary, September 1999.

[BIE 94] Bielefeld B. "Language Identification using shifted delta cepstrum", XIVth Annual Speech Research Symposium, Baltimore, 1994.

[BIM 08] Bimbot F. "Automatic Speaker Recognition", chapter in this volume.

[BON 01] Bonastre, J.F., Magrin-Chagnolleau, I., Euler, S., Pellegrino, F., André-Obrecht, R., Mason, J., Bimbot, F., "Speaker and Language characterization (SpLC): A Special Interest Group (SIG) of ISCA", *Proc. of Eurospeech-01*, pp.1145-1148, Aalborg, Danemark, 2001.

[BOU 04] Boula de Mareüil P., Marotta G., Adda-Decker M., "Contribution of prosody to the perception of Spanish/Italian accents", *Proc. of Speech Prosody*, Nara, Japan, March, 2004.

[BOU 06] Boula de Mareüil P., Vieru-Dimulescu B., "The Contribution of Prosody to the Perception of Foreign Accent", *Phonetica*, Int. Journal of Phonetic Science, pp.247-267, vol. 63, n°4, December, 2006.

[BOY 91] Boysson-Bardies B. de, Vihman M., "Adaptation to language: evidence from babbling and first words from four languages", *Language* 67 (2) 1991, pp. 297-319, 1991.

[CAM 02] Campbell W., "Generalized linear discriminant sequence kernels for speaker recognition", *Proc. of ICASSP-02*, pp.161-164, Orlando, May, 2002.

[CAM 04] Campbell W., Singer E., Torres-Carrasquillo P., Reynolds D., "Language Recognition with Support Vector Machines", *Proc. Odyssey-04*, Toledo, June, 2004.

[CAM 06] Campbell W., et al, "Advanced Language Recognition using Cepstra and Phonotactics: MIT-LL System Performance on the NIST 2005 Language Recognition Evaluation", *Proc. of Odyssey-06*, Puerto Rico, June, 2006.

[CAS 98] Caseiro D., Trancoso I., "Spoken Language Identification Using the Speechdat Corpus", *Proc. of ICSLP-98*, Sydney, Australia, December, 1998.

[CHI 77] Childers D.G., et al, "The Cepstrum: a guide to processing", *Proc. of the IEEE*, Vol.65, n°10, pp.1428-1443, October, 1977.

[CHO 77] Chomsky N., Halle M., *The Sound Pattern of English*, MIT Press, Cambridge, Massachusetts, London, England, 1977.

[CIM 82] Cimarusti C., Ives R., "Development of an automatic identification system of spoken languages: phase I", *Proc. of ICASSP-82*, pp.1661-1663, May 1982.

[CRI 00] Cristianini N., Shawe-Taylor J., *Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[COL 01] Collobert R., Bengio S., "SVMTorch: Support vector machines for large-scale regression problems", *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[COM 90] Comrie B., *The World's Major Languages*, Oxford University Press, Oxford UK, 1990.

[COM 01] Compernolle (van) D., "Speech recognition by Goats, Wolves, Sheep and ... Nonnatives", *Speech Communication*, Volume 35., 2001, pp.71-79.

[COR 97] Corredor-Ardoy C., Gauvain J.L., Adda-Decker M., Lamel L., "Language identification with language-independent acoustic models". *Proc. Eurospeech-97*, Rhodos, Greece, September, 1997.

[CRY 93] Crystal D. *The Cambridge Factfinder*, Cambridge University Press, Cambridge UK, 1993.

[CRY 97] Crystal D. *The Cambridge Encyclopedia of Language*, Cambridge University Press, Cambridge UK, 1997.

[DAL 96] Dalsgaard P., Andersen O., Hesselager H., Petek B., "Language-Identification using Language-Dependent Phonemes and Language-Independent Speech Units", *Proc. of ICSLP-96*, Philadelphia,USA, October, 1996.

[DAN 96] Daniels P., Bright, W. *The World's Writing Systems*, Oxford University Press, Oxford 1996.

[DAV 80] Davis S., Mermelstein P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, nO.4, pp.357-366, August 1980.

[DEL 65] Delattre P., *Comparing the phonetic features of English, Spanish, German and French*. Julius Gross Verlag. Heidelberg 1965.

[DUT 00] Dutat M., Magrin-Chagnolleau I., Bimbot F., "Language Recognition Using Time-Frequency Principal Component Analysis And Acoustic Modeling", *Proc. of ICSLP-00*, Beijing, China, Octobre 2000.

[EAR 05] EARS: Effective affordable reusable speech-to-text,
http://projects.ldc.upenn.edu/EARS/

[ELD] ELDA: European Language Resources Distribution Agency. http://www.elda.org

[FAR 01] Farinas J., Pellegrino F.,"Automatic rhythm modeling for Language Identification", *Proc. of Eurospeech-03*, September, 2001.

[FLE 03] Flege J.E., et al, "Interaction between the native and second language phonetic subsystems", *Speech Communication*, Volume 40, (**4**), June, 2003.

[FOI 86] Foil, J. T., "Language identification using noisy speech", *Proc. of ICASSP-86*, pp.861-864, 1986.

[GAU 04] Gauvain J.L., Messaoudi A., Schwenk H., "Language recognition using phone lattices", *Proc. of ICSLP 04*, Jeju Island, South Korea 2004.

[GEN 07] Gendrot C., Adda-Decker M., "Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages", *Proc. of ICPhS-07*, Saarbrücken, August, 2007.

[GEO 04] Geoffrois E., "Identification automatique des langues: techniques, ressources et évaluations", *Proc. of MIDL-04*, November, 2004.

[ETH 05] Gordon R.G., (ed.), 2005. *Ethnologue: Languages of the World*, 15th edition. Dallas, Tex. SIL International. Online version: http://www.ethnologue.com/

[GRE 78] Greenberg J., Ferguson C., Moravcsik E. (Ed.) 1978 *Universals of Human Languages: Method and Theory*. Stanford University Press.

[HAL 91] Hallé P.A., de Boysson-Bardies B., Vihman M., "Beginnings of prosodic organization: intonation and duration patterns of disyllables produced by Japanese and French infants", *Language and Speech* 34, pp.299-318, 1991.

[HAZ 93] Hazen T., Zue V., "Automatic language identification using a segment-based approach", *Proc. of Eurospeech-03*, Berlin, Germany, September, 1993.

[HAZ 97] Hazen T., Zue V., "Segment-Based Automatic Language Identification", *Journal of the Acoustical Society of America*, (JASA), Vol.101, No. 4, pp.2323-2331, April 1997.

[HER 90] Hermansky H., "Perceptual lineair predictive (PLP) analysis of speech", *Journal of Acoustical Society of America*, (JASA), Vol.87, no. 4, pp.1738-1752., April 1990.

[HIE 94] Hieronymous J., "ASCII phonetic symbols for the world's languages: Worldbet", *Technical Report AT&T Bell Labs*, 1994.

[HIE 96] Hieronymous J., Kadambe S., "Spoken Language Identification Using Large Vocabulary Speech Recognition", *Proc. of ICSLP-96*, Philadelphia, USA, 1996.

[HIR 99] Hirst D., Di Cristo A. (ed.), *Intonation Systems: A Survey of Twenty Languages*, Cambridge University Press, 1999.

[HOM 99] Hombert J.M., Maddieson I., "The Use of 'Rare' Segments for Language Identification", *Proc. of Eurospeech-99*, Budapest, Hungary, September, 1999.

[HOM 06] Hombert J.M. (dir.) *Aux Origines des langues et du langage - Towards origins of languages and language*, directed by Jean-Marie Hombert, Fayard ed. 2005.

[HON 05] Hönig F., et al, "Revising Perceptual Linear Prediction (PLP)", *Proc. of Eurospeech-05*, September, 2005.

[HOU 77] House A.S., Neuburg E.P., "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations", *Journal of the Acoustical Society of America*, (JASA), Vol.62, n°3, pp.708-713, September, 1977.

[IPA 99] IPA association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[ITA 93] Itahashi S., et al, "Language Identification with phonological and lexical models", *Proc. of Eurospeech-95*, Madrid, Spain, September, 1995.

[JEL 98] Jelinek F., *Statistical methods for speech recognition*, ISBN 0-262-10066-5, The MIT Press, 1998.

[KIR 02] Kirchhoff K., Parandekar S., Bilmes J., "Mixed-memory Markov Models for Automatic Language Identification", *Proc. of ICASSP-02*, Orlando, 2002.

[KIR 06] Kirchhoff K., "Language Characteristics", in *Multilingual speech processing*, T. Schultz & K. Kirchhoff eds., Elsevier, 2006.

[KUM 96] Kumpf K., King R.W., "Automatic accent classification of foreign Accented Australian English Speech", *Proc. of ICSLP-96*, Philadelphia, October, 1996.

[KWA 95] Kwan H.K., Hirose K., "Recognized phoneme-based n-gram modeling in automatic language identification", *Proc. of Eurospeech-95*, Madrid, Spain, September, 1995.

[LAB 04] Labov W., "Phonological atlas of North America", 2004.
http://www.ling.upenn.edu/phono_atlas/home.html

[LAD 96] Ladefoged P., Maddieson I., *The Sounds of the World's Languages*, Blackwell, 1996.

[LAM 94] Lamel L., Gauvain J.L., "Language Identification Using Phone-based Acoustic Likelihoods", *Proc. of ICASSP-94*, Adelaide, April, 1994.

[LAM 95] Lamel L., Gauvain J.L., "A phone-based approach to non-linguistic speech feature identification", *Computer Speech and Language*, January, 1995.

[LAM 98] Lamel L., Adda G., Adda-Decker M., Corredor-Ardoy C., Gangolf J.J., Gauvain J.L. "A multilingual corpus for language identification", *Proc. of LREC-98*, Granada, May, 1998.

[LAN 95] Lander T., et al., "The OGI 22 language telephone speech corpus", *Proc. of Eurospeech-95*, Madrid, Spain, September, 1995.

[LAV] Laval University, "Les grandes familles linguistiques du monde",
http://www.tlfq.ulaval.ca/AXL/monde/familles.htm

[LDC 07] LDC, http://www.ldc.upenn.edu/Catalog/

[CAL 96] LDC-Callfriend, "CallFriend Corpus," Linguistic Data Consortium, 1996.

[CAL2] LDC-Callfriend2 LVDID: Language, Variety and Dialect Identification,
http://www.ldc.upenn.edu/CallFriend2

[LEO 74] Leonard R., Doddington G., "Automatic language identification", *Technical report* RADC-TR-74-200, Air Force Rome Air Development Center, August, 1974.

[LEV 91] Levitt A., Wang Q. "Evidence for language-specific rhythmic influences in the reduplicative babbling of French-and-English-learning infants", *Language and Speech* 34, pp.235-249, 1991.

[LI 80] Li K., Edwards T., "Statistical models for automatic language identification", *Proc. of ICASSP-80*, pp.884-887, Denver, Colorado, 1980.

[LI 94] Li K., Edwards T., "Automatic Language Identification Using Syllabic Spectral Features", *Proc. of ICASSP-94*, Adelaide, Australia, April, 1994.

[LUN 96] Lund M., Ma K., Gish H., "Statistical Language Identification Based on Untranscribed Training", *Proc. of ICASSP-94*, Adelaide, Australia, April, 1994.

[MAD 02] Maddieson I., Vasilescu I., "Factors in Human Language Identification", *Proc. of ICSLP-02*, Denver, 2002.

[MAR 98] Mariani J., Paroubek P. "Human language technologies evaluation in the European framework", *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, pp.237-242, 1998.

[MAR 05] Mariani J., "Developing Language Technologies with the Support of Language Resources and Evaluation Programs", *Journal of Language Resources and Evaluation*, Springer Netherlands, 2005.

[MAR 13] Markov A.A., "An example of statistical investigationin the text of *Eugen Onyegin* - Illustrating the coupling of tests in chains", *Proc. Acad. of St. Petersburg*, Vol.7, pp.153-162, 1913.

[MAR 97] Martin A., et al, "The Det Curve in Assessment of Detection Task Performance ", *Proc. of EUROSPEECH-97*, Rhodos, Greece, September, 1997.

[MAR 03] Martin A., Przybocki M., "NIST 2003 Language Recognition Evaluation", *Proc. of Eurospeech-03*, pp.1341-1344, Geneva, September, 2003.

[MAR 06] Martin T., Baker B., Wong E., Sridharan S., "A syllable-scale framework for language identification", *Computer Speech & Language*, Vol.20, pp.276-302, 2006.

[NIS 05] Martin A., Le A., "The Current State of Language Recognition: NIST 2005 Evaluation Results", *Proc. of Odyssey-06*, Puerto Rico, June, 2006.

[MAT 05] Matejka P., et al, "Phonotactic Language Identification using High Quality Phoneme Recognition", Proc. of the 9th European Conference on Speech Communication and Technology (Eurospeech), Lisboa, Portugal, September, 2005.

[MAT 06] Matejka P., et al, "Brno University of Technology System for Nist 2005 Language Recognition Evaluation", Workshop Odyssey, June, 2006.

[MAT 98] Matrouf D., Adda-Decker M., Lamel L., Gauvain J.L., "Language Identification Incorporating Lexical Information", *Proc. of ICSLP-98*, Sydney, Australia, December, 1998.

[MAT 99] Matrouf D., Adda-Decker M., Gauvain J.L., Lamel L., "Comparing Different Model Configurations for Language Identification Using a Phonotactic Approach", *Proc. of Eurospeech-99*, Budapest, Hungary, September, 1999.

[MEN 96] Mendoza S., Gillick L., Ito Y., Lowe S., Newman M., "Automatic Language Identification using Large Vocabulary Continuous Speech Recognition", *Proc. of ICASSP-96*, Atlanta, April, 1996.

[MON 06] Montero-Asenjo A., et al, "Exploring PPRLM performance for NIST 2005 Language Recognition Evaluation", *Proc. of Odyssey-06*, Puerto Rico, June, 2006.

[MOR 99] Mori K., Toba N., Harada T., Arai T., Komatsu M., Aoyagi M., Murahara Y., "Human Language Identification with Reduced Spectral Information", *Proc. of Eurospeech-99*, Budapest, Hungary, September, 1999.

[MUT 92] Muthusamy Y., Cole R., Oshika B., "The OGI multi-language telephone speech corpus", *Proc. of ICSLP-92*, Alberta, October, 1992.

[MUT 93] Muthusamy Y., "A Segmental Approach to Automatic Language Identification, A Segment Based Automatic Language Identification System", *Ph.D. thesis*, Oregon Graduate Institute of Science and Technology, July, 1993.

[NAG 06] Nagarajan T., Murthy H.A., "Language identification NAG 0620using acoustic loglikelihoods of syllable-like units", *Speech Communication*, pp.913-926, Volume 48, 2006.

[NAK 92] Nakagawa, S., Ueda, Y., Seino, T., "Speaker-independent, text-independent language identification by HMM", *Proc. of ICSLP-92*, pp.1011-1014, Alberta, October, 1992.

[NAV 01] Navrátil J., "Spoken language recognition: a step towards multilinguality in speech processing", in IEEE *Transactions on Speech and Audio Processing*, Vol.9, n°6, pp.678-685, 2001.

[NAV 06] Navrátil J., "Automatic Language Identification", in *Multilingual speech processing*, T. Schultz & K. Kirchhoff eds., Elsevier, 2006.

[NIE 06] Niesler T., Willett D., "Language identification and multilingual speech recognition using discriminatively trained acoustic models", *Proc. of MULTILING-06*, Stellenbosch, South Africa, April, 2006.

[NIS 03] NIST, *The 2003 NIST language recognition evaluation plan*, http://www.nist.gov/speech/tests/lang/index.htm

[OBU 05] Obuchi Y., et al, "Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization", *Proc. of ICASSP-05*, Philadelphia, March, 2005.

[OGI 07] OGI-22L, http://cslu.cse.ogi.edu/corpora/22lang/

[PAR 99] Parlangeau N., Pellegrino F., André-Obrecht R., "Investigating Automatic Language Discrimination via Vowel System and Consonantal System Modeling", *Proc. of ICPhS-99*, San Francisco, 1999.

[PEL 00] Pellegrino F., André-Obrecht R., "Automatic language identification: an alternative approach to phonetic modeling", *Signal Processing*, Elsevier Science North Holland, Vol.80, pp.1231-1244, July, 2000.

[PUL 96] Pullum G. K., Ladusaw W. A., "Phonetic Symbol Guide", ed. University of Chicago Press, second edition, 1996.

[RAM 99] Ramus F., Mehler J., " Language identification with suprasegmental cues: A study based on speech resynthesis", Journal of Acoustical Society of America, (JASA), Vol.105, 1999.

[ROU 05] Rouas J.L., "Modeling Long and Short-Term Prosody for Language Identification", *Proc. of Eurospeech-05*, Lisboa, Portugal, September, 2005.

[RUH 05] Ruhlen, M., "Taxonomy, typology, and historical linguistics". In James W. Minett and William S.-Y. Wang, editors, *Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics*, City University of Hong Kong Press, 2005.

[SAU 15] Saussure, F. de, 1915. *Cours de linguistique générale*. Payot, Paris.

[SCH 96] Schultz T., Rogina I., Waibel A., "LVCSR-Based Language Identification", *Proc. of ICASSP-96*, Atlanta, USA, May, 1996.

[SCH 02] Schultz T., "Globalphone: a multilingual text and speech database developed at Karlsruhe University", *Proc. of ICSLP-02*, Denver 2002.

[SCH 06] Schultz T., et al, *Multilingual speech processing*, ISBN 13: 978-0-12-088501-5, Elsevier, 2006.

[SCH 97] Schwartz J.L., Boë L.J., Vallée N., Abry C., "Major trends in vowel system inventories", *Journal of Phonetics*, Vol.25, n°3, pp.233-253, 1997.

[SHE 06] Shen W., Campbell W., Gleason T., Reynolds D., Singer E., "Experiments with Lattice-based PPRLM Language Identification", *Proc. of Odyssey-06*, Puerto Rico, June, 2006.

[SHO 80] Shoup J., "Phonological aspects of speech recognition", In Lea W. (Ed.), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, pp.125-138, 1980.

[SIN 03] Singer E., Torres-Carrasquillo P., Gleason T., Campbell W., Reynolds D., "Acoustic, Phonetic and Discriminative Approaches to Automatic Language Recognition", *Proc. of EuroSpeech-03*, Geneva, September, 2003.

[SIN 04] Singer E., Reynolds D., "Analysis of Multitarget Detection for Speaker and Language Recognition", *Proc. of Odyssey-04*, Toledo, 2004.

[SPDM] Speechdat-(M) mobile phone: EU-project LRE-63314.
http://www.phonetik.uni-muenchen.de/SpeechDat.html

[SUG 91] Sugiyama, M., "Automatic Language Recognition Using Acoustic Features," *Proc. of ICASSP-91*, pp.813-816, Toronto, Ontario, May, 1991.

[TEI 97] Teixeira C., Trancoso I., Serralheiro A., "Recognition of Non-Native Accents", *Proc. of Eurospeech-97*, Rhodos, Greece, September, 1997.

[THY 96] Thymé-Gobbel A.E., Hutchins S.E., "On using prosodic cues in automatic language identification", *Proc. of ICSLP-96*, Philadelphia, October, 1996.

[TOR 02a] Torres-Carrasquillo P., Reynolds D., Deller J., "Language Identification Using Gaussian Mixture Model Tokenization," *Proc. of ICASSP-02*, Orlando, 2002.

[TOR 02b] Torres-Carrasquillo P., Singer E., Kohler M., Greene R., Reynolds D., Deller J., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", *Proc. of ICSLP-02*, Denver, 2002.

[TOR 04] Torres-Carrasquillo P., Gleason T., Reynolds D., "Dialect identification using Gaussian Mixture Models", *Proc. of Odyssey-04*, Toledo, 2004. Spanish dialects

[TUC 94] Tucker R., Carey M., Parris E., "Automatic Language Identification Using Sub-Word Models", *Proc. of ICASSP-94*, Adelaide, Australia, April, 1994.

[UED 90] Ueda, Y., Nakagawa, S., "Prediction for phoneme/syllable/word category and identification of language using HMM", *Proc. of ICSLP-90*, pp.1209-1212, Kobe, 1990.

[VAL 99] Vallée N., Schwartz J.L., Escudier P.,"Phase spaces of vowel systems: a typology in the light of the Dispersion-Focalisation Theory", *Proc. of ICPhS-99*, San Francisco, 1999.

[VAP 97] Vapnik V.N. "Support vector learning machines - Tutorial at NIPS'97", Denver (CO), December, 1997.

[VAS 05] Vasilescu I., Candea M., Adda-Decker M., "Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages", *Proc. of Eurospeech-05*, Lisboa, Portugal, September, 2005.

[WAN 99] Wanneroy R., Bilinski E., Barras C., Adda-Decker M., Geoffrois E., "Acoustic-Phonetic Modeling of Non-Native Speech for Language Identification", *Proc. of MIST-99*, Leusden, The Netherlands, September, 1999.

[WEL 97] Wells J., "SAMPA computer readable phonetic alphabet", In Gibbon D., Moore R., Winsky R. (eds), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, New York - Berlin, 1997.

[WHI 06] White C., Shafran I., Gauvain J.L., "Discriminative classifiers for language recognition", *Proc. of ICASSP-06*, Toulouse, France, May, 2006.

[WIT 99] Witt, S., Young, S., "Off-line acoustic modelling of non-native accents", *Proc. of Eurospeech-99*,, pp.1367-1370, Budapest, Hungary, September, 1999.

[WON 00] Wong E., Pelecanos J., Myers S., Sridharan S., "Language identification using efficient Gaussian mixture model analysis", *Proc. of Australian Int. Conference on Speech Science and Technology*, 2000.

[WON 02] Wong E., Sridharan S., "Methods to Improve Gaussian Mixture Model Based Language Identification System", *Proc. of ICSLP-02*, Denver, 2002.

[YAN 95a] Yan Y., Barnard E., "An approach to automatic language identification based on language-dependent phoneme recognition", *Proc. of ICASSP-95*, pp.3511-3514, Madrid, Spain, September, 1995.

[YAN 95b] Yan Y., Barnard E., "A comparison of neural net and linear classifier as the pattern recognizer in automatic language identification", *Proc. of ICNNSP-95*, Nanjing, 1995.

[YAN 96a] Yan Y., Barnard E., "Experiments with Conversational Telephone Speech for Language Identification", *Proc. of ICASSP-96*, Atlanta, USA, May, 1996.

[YAN 96b] Yan Y., Barnard E., Cole R., "Development of an Approach to Language Identification based on Phone Recognition", *Computer Speech & Language*, **10**, pp.37-54, 1996.

[ZHU 05] Zhu D., et al, "Different Size Multilingual Phone Inventories and Context-Dependent Acoustic Models for Language Identification", *Proc. of Eurospeech-05*, Lisboa, Portugal, September, 2005.

[ZHU 06] Zhu D., Adda-Decker M., "Language identification using lattice-based phonotactic and syllabotactic approaches", *Proc. of Odyssey-06*, June, 2006.

[ZIP 49] Zipf G.K., "Human behaviour and the principle of least effort", Addison-Wesley Publishing, Reading, MA, 1949.

[ZIS 93] Zissman M., "Automatic Language Identification using Gaussian Mixture and Hidden Markov Models", *Proc. of ICASSP-93*, pp.399-402, Minneapolis, April, 1993.

[ZIS 94] Zissman M., Singer E., "Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modeling", *Proc. of ICASSP-94*, Vol.1, pp.305-308, Adelaide, Australia, April, 1994.

[ZIS 96a] Zissman M., "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", IEEE *Transactions on Speech and Audio Processing*, SAP-4(1), pp.31-44, January, 1996.

[ZIS 96b] Zissman M., Gleason T., Rekart D., Losiewicz B., "Automatic Dialect Identification of Extemporaneous Conversational, Latin American Spanish Speech", *Proc. of ICASSP-96*, Atlanta, Georgia, 1996.

[ZIS 97] Zissman M., "Predicting, diagnosing and improving automatic language identification performance", *Proc. of Eurospeech-97*, pp.51-54, Rhodos Greece, September, 1997.

[ZIS 99] Zissman M., Berkling K., "Automatic Language Identification", *Speech Communication*, pp.115-124, Vol. 35, Issues 1-2, August, 2001.

Chapter 9

Index