40. Spoken Language Characterization

M. P. Harper, M. Maxwell

This chapter describes the types of information that can be used to characterize spoken languages. Automatic spoken language identification (LID) systems, which are tasked with determining the identity of the language of speech samples, can utilize a variety of information sources in order to distinguish among languages. In this chapter, we first define what we mean by a language (as opposed to a dialect). We then describe some of the language collections that have been used to investigate spoken language identification, followed by discussion of the types of features that have been or could be utilized by automatic systems and people. In general, approaches used by people and machines differ, perhaps sufficiently to suggest building a partnership between human and machine. We finish with a discussion of the conditions under which textual materials could be used to augment our ability to characterize a spoken language.

As we move into an increasingly globalized society, we are faced with an ever-growing need to cope with a variety of languages in computer-encoded text documents, documents in print, hand-written (block and cursive) documents, speech recordings of various qualities, and video recordings potentially containing both speech and textual components. A first step in coping with these language artifacts is to identify their language (or languages).

The scope of the problem is daunting given that there are around 7000 languages spoken across the world, as classified by the Ethnologue [40.1]. Indeed it can be difficult to collect materials for all of these languages, let alone develop approaches capable of discriminating among them. Yet the applications that would be supported by the ability to effectively and efficiently identify the language of a text or speech input are compelling: document (speech and text) retrieval, automated routing to machine translation or speech recognition systems, spoken dialog systems (e.g., for making travel arrangements), data mining systems, and systems to route emergency calls to an appropriate language expert.

In practice, the number of languages that one might need to identify is for most purposes much less than 7000. According to the Ethnologue, only about 330 languages have more than a million speakers. Thus, one might argue that in practice, there is a need for language identification (ID) of a set of languages that number perhaps in the hundreds. For example, Language Line Services [40.2] provides interpreter services to public and private clients for 156 languages, which they claim represents around 98.6% of all their customer requests for language services. Of course, for some purposes, the set of languages of interest could be far smaller.

This chapter discusses the knowledge sources that could be utilized to characterize a spoken language in order to distinguish it automatically from other languages. As a first step, we define what we mean by a language (as opposed to a dialect). We then describe some of the language collections that have been used to investigate spoken language identification, followed by a discussion of the types of linguistic features that have been or could be used by a spoken language identification (LID) system to determine the identity of the language of a speech sample. We also describe the cues that humans can use for language identification. In general, the approaches used by people and machines differ, perhaps sufficiently to point the way towards building a partnership between human and machine. We finish with a discussion of the conditions under which textual materials could be used to augment our ability to characterize a spoken language.

40.1 Language versus Dialect

When talking about language identification, it is important to define what we mean by a language. At first glance, the definition of *language* would seem to be simple, but it is not. Traditionally, linguists distinguish between languages and dialects by saying that dialects are mutually intelligible, unlike distinct languages. But in practice, this distinction is often unclear: mutual intelligibility is relative, depending on the speaker's and listener's desire to communicate, the topic of communication (with common concepts being typically easier to comprehend than technical concepts or concepts which happen to be foreign to one or the other participant's culture), familiarity of the hearer with the speaker's language (with time, other *accents* become more intelligible), degree of bilingualism, etc.

Political factors can further blur the distinction between dialect and language, particularly when two peoples want – or do not want – to be considered distinct. The questionable status of Serbo-Croatian is an obvious example; this was until recently considered to be more or less a single unified language. But with the breakup of Yugoslavia, the Serbian and Croatian languages (and often Bosnian) have been distinguished by many observers, despite the fact that they are largely mutually intelligible. The status of the various spoken varieties of Arabic is an example of the opposite trend. While many of these varieties are clearly distinct languages from the standpoint of mutual intelligibility, the desire for Arab unity has made some claim that they are merely dialects [40.3, 4].

Writing systems may also cause confusion for the distinction between languages and dialects. This is the case for Hindi and Urdu, for example, which in their spoken form are for the most part mutually intelligible (differing slightly in vocabulary). But they use radically different writing systems (Devanagari for Hindi, and a Perso-Arabic script for Urdu), in addition to being used in different countries, and so are generally treated as two different languages [40.5]. Hence, the distinction between language and dialect is often unclear, and so in general it is better viewed as a continuum rather than a dichotomy.

At a higher level, one can characterize a language by its language family, which is a phylogenetic unit such that all members are descended from a common ancestor (languages that cannot be reliably classified into a family are called language isolates). For example, Romance languages (e.g., French, Spanish), Germanic languages (e.g., German, Norwegian), Indo-Aryan languages (e.g., Hindi, Bengali), Slavic languages (e.g., Russian, Czech), and Celtic languages (e.g., Irish and Scots Gaelic) among others are believed to be descended from a common ancestor language some thousands of years ago, and hence are grouped into the Indo-European language family. Language families are often subdivided into branches, although the term family is not restricted to one level of a language tree (e.g., the Germanic branch of Indo-European language family is often called the Germanic family). As a result of their common descent, the languages of a language family or branch generally share characteristics of phonology, vocabulary, and grammar, although these shared resemblances may be obscured by changes in a particular language, including borrowings from languages of other language families. Thus, while English is a Germanic language, its grammar is substantially different from that of other Germanic languages, and a large portion of its vocabulary is derived from non-Germanic languages, particularly French [40.6].

The issues of language, dialect, and language family have repercussions for systems that try to assign a language tag to a piece of text or a sample of speech, and in particular for the International Organization for Standardization (ISO) standard language codes. The original standard, ISO 639-1, listed only 136 codes to distinguish languages. This allowed for the identification of most *major* languages, but not minor languages. In fact some of the codes did not represent languages at all. The code for *Quechua*, for example, corresponds to an entire language family consisting of a number of mutually unintelligible languages; arguably worse, the code for *North American Indian* refers to a number of language families, most of which include several distinct languages. For an analysis of some of the problems with ISO 639-1 and its revision, ISO 639-2, see [40.7].

At the other end of the spectrum of language classification is the Ethnologue [40.1], a listing of nearly 7000 languages. This work takes an explicitly linguistic view of language classification, i. e., it attempts to assign distinct names to all and only mutually unintelligible varieties. Many observers have accused the Ethnologue of being a *splitter*, i. e., of claiming too many distinctions. This is a debatable point, but it does serve to highlight a fundamental problem of classifying a language sample as belonging to this language or that: it is hard to classify a language artifact if we cannot agree ahead of time what the possibilities are. And indeed, for some purposes, the finer-grained classification like the Ethnologue may be superfluous, while for others it may be crucial. One example of the need for fine-grained classification would be Arabic *dialect* identification, where it may be desirable to determine which variety of Arabic someone is speaking.

To some extent, the problem of conflicting classification criteria (ISO versus Ethnologue) has been resolved by unifying the two systems as ISO 639-3 [40.8]. In addition to the set of 7000 languages already listed in the Ethnologue, ISO 639-3 adds various other languages, including extinct languages and artificial languages (such as Klingon) and so-called *macro languages*, which are really language families (such as Arabic). However, even given that the standard can now be agreed on – the set of languages in ISO 639-3 – there does remain a difficulty for language identification, namely the level of granularity for language classification. The reader should remember that the problem exists, and that identifying a text or speech recording as *Arabic* may be adequate for some purposes, but insufficient for others.

The continuum from language to dialect has another implication for language identification: the existence of significant differences among dialects of a single language can make identification of that language more difficult, particularly in its spoken form (written forms of languages tend to be more standardized [40.9]). Dialectal differences can occur in all linguistic aspects: lexicon, grammar (syntax and morphology), and phonology; but it will probably be the phonology that causes the most problems for language identification in speech, since this is the level of representation that is commonly used in existing systems.

In English, for example, the differences between rhotic and non-rhotic dialects are well known [40.10]. The Scots dialects of English demonstrate even greater differences from other dialects of English in their vowel systems, where they have largely lost the distinction between so-called *long* and *short* vowels; and among the consonants, Scots English has retained the voiceless velar fricative and the voiceless labiovelar, both of which have merged with other phonemes in other dialects of English [40.11]. See the SCOTS project [40.12] for some recorded speech examples. Likewise, spoken Mandarin is strongly influenced by the native dialect spoken in a region, as well as other factors such as age [40.13].

Dialects that differ primarily in their phonology are often called *accents*, although this term is also used to refer to pronunciation by a non-native speaker of a language. In this regard, while foreign accents might be dismissed as irrelevant to language identification for some purposes, in the case of major languages, there may be significant communities of non-native speakers who use the major language as a trade language. For example, in India there are 21 (currently) official languages, but English is defined by the Constitution of India, as well as by later laws, as one of the languages of communication for the federal government (the other being Hindi). The result of this and other factors is that English is spoken non-natively by a large portion of the Indian population, and it has acquired a distinctly Indian pronunciation as a consequence [40.14]. Indian English lacks voiceless aspirated stops; what would be alveolar consonants in other varieties of English are often retroflexed; and the stress patterns are altered with the effect that vowels that would be reduced in other Englishes to schwa appear instead in their nonreduced forms.

English is used as a trade language or lingua franca in many other parts of the world as well, and these local versions of English are often significantly different from the English spoken in countries where it is a native language (see [40.15] for descriptions of some of these varieties). The same is true of French, Hausa, and many other languages that are used as trade languages (cf. [40.16] and the Ethnologue [40.1]). Such lingua franca varieties can differ significantly from *standard* varieties in ways which would likely impact language identification.

Additionally, there is a substantial amount of variability in the spoken realization of a particular language [40.17–19] due to a variety of factors, including:

- mispronunciation
- individual speaking style
- genre (e.g., conversations versus formal presentations)
- variations in speaking rate
- the speaker's psychological state
- the speaker's language repertoire
- the social and economic background of the speaker
- the speaker's first language (where the speaker is speaking in a second language)
- channel characteristics

This variability creates a challenge when identifying the language of a speech sample. A greater understanding of language, dialect, and accent is an important first step in developing an approach for language identification.

40.2 Spoken Language Collections

Engineering research in spoken language identification has been based on a very small sampling of languages from the 7000 identified by the Ethnologue. The first multilingual speech collection targeted for language identification system evaluation, the Oregon Graduate Institute (OGI) multilanguage corpus, was released in 1994 and consisted of spoken responses to prompts recorded over telephone lines by speakers of 12 languages (see [40.20–22]). The set was selected to contain both unrelated languages (e.g., German, Vietnamese, and Tamil), as well as more closely related languages (e.g., English and German). It covers languages with various types of prosodic phenomena that occur in spoken languages, such as tone (e.g., Mandarin and Vietnamese) and pitch accents (Japanese), as well as languages with various levels of complexity at the syllable level. See appendix A in [40.23] for a family language tree for the languages in this collection and appendix B for a table comparing the phone inventories for these languages. Perhaps one of the most important requirements for the languages selected was the ability to access sufficient numbers of speakers of the language in the United States, a factor that has been important in collections used for evaluation.

A second collection, the Linguistic Data Consortium (LDC) CallFriend corpus (see the list under the LID heading at [40.24]), was released in 1996 and contains telephone conversations among speakers of 14 languages or dialects, most of which appeared in the OGI multilanguage corpus. This collection was used in National Institute of Standards and Technology (NIST) evaluations of LID systems in 1996 and 2003. The CallHome collection, which was released between 1996 and 1997, is an additional multilingual resource containing six languages that were collected for large-vocabulary speech recognition; however, it does not expand the number of languages beyond what was available from CallFriend. Given the limited sampling of languages in these collections, it is not surprising that researchers developing spoken LID systems have not investigated whether languages naturally cluster into groups that parallel the classification of languages into families and branches.

Additional speech collections have been developed with an increased number of languages. The Center for Spoken Language Understanding (CSLU) 22 languages corpus, which was initially collected from 1994 through 1997, contains spoken utterances in 21 languages. It has gone through several versions, resulting in an increased number of transcribed utterances, and was released through LDC in 2005 (see [40.25]). LDC is also currently collecting the MIXER corpus which will contain telephone calls over 24 languages or dialects [40.26]. The 2005 NIST evaluation data contained speech from the Mixer and CallFriend corpora, as well as some data collected at the Oregon Health and Science University. The data in all of these collections will hopefully stimulate new methods for constructing speech-based LID systems. However, even 20 or so languages is far from the number of languages that would need to be identified to support Language Line Services.

The collection of significant quantities of comparable telephone speech in multiple languages has become more challenging in recent years. There is an implicit assumption that the training and development data are collected under conditions that are comparable to the evaluation materials. Data must also be collected in such a way that it is impossible to identify language based on speaker, gender distribution, domain, channel characteristics, etc. Finally, due to the falling cost of long-distance telephone calls, incentives such as free long-distance calls are less attractive to potential participants. This difficulty in collecting comparable speech corpora for a large number of languages has obvious implications for speech-based LID; we return to this issue later.

40.3 Spoken Language Characteristics

A speech-based LID model for a particular language is trained to represent the corresponding language and to differentiate it from others. Hence an important challenge to speech-based LID systems is the effective incorporation of discriminative knowledge sources into their models. Some systems use only the digitized speech utterances and the corresponding true identities of the languages being spoken for training, whereas others require additional information such as phonetic and/or orthographic transcriptions, which can be expensive to produce. During the language recognition phase, a new audio sample is compared to each of the language-dependent models, and the language of the closest matching model (e.g., using maximum likelihood) is selected. We will discuss various levels of knowledge that can be used to identify a language in this section, touching upon research that utilizes the representation where appropriate.

Automatically Derived Features from the Speech Signal. Languages can be identified based on features that are automatically derived from the speech signal itself. Systems utilizing this type of information are motivated by the observation that different languages are made up of a variety of different sounds; hence, feature vectors automatically extracted from the speech signal over short time frames (segments) can be used to discriminate among the languages. Such systems typically use a multistep process (including modules to remove silence from the samples, to reduce channel effects, etc.) to convert the digitized speech signal into a feature vector representation. Given feature extraction and knowledge of which training samples correspond to each language, a classifier is constructed for each language based on language-dependent patterns of feature vectors. Methods include approaches that model only the static distribution of acoustic features given the language (e.g., [40.27]) and approaches that also utilize the patterns of change of these feature vectors over time (e.g., [40.28]). A variety of computational models have been investigated, including Gaussian mixture models [40.29], Hidden Markov models [40.29, 30], artificial neural networks [40.22], and support vector machines [40.31]. Although these acoustic-based systems do not require training data that is labeled with explicit linguistic units such as phonemes or words, they also do not perform as accurately as systems that use linguistic knowledge; however, improved accuracy can be achieved by integrating acoustic- and linguistic-based knowledge sources [40.23, 32].

Phonological Information. Phonology is a branch of linguistics that studies sound systems of human languages [40.33]. In a given language, a phoneme is a symbolic unit at a particular level of representation; the phoneme can be conceived of as representing a family of related phones that speakers of a language think of as being categorically the same. While the notion of phonemes has been controversial among linguists (see [40.34] for some history), it has proven a useful abstraction for speech processing. *Ladefoged* [40.35] speculates that 'there are probably about 600 different consonants' (p. 194) across the languages of the world; vowels and suprasegmental distinctions (such as tone) are not quite as numerous.

Phonetic symbols provide a way to transcribe the sounds of spoken languages. There are several phonetic alphabets; one commonly used by linguists is the international phonetic alphabet (IPA). This alphabet, including a set of diacritic modifiers, was established as a standard by the International Phonetic Association (IPA) in order to provide an accurate representational system for transcribing the speech sounds of all languages (the IPA website is at [40.36]). The goal for the IPA is to provide a representation for all of the phonemes expressed in all human languages, such that separate symbols are used for two sounds only if there exists a language for which these two sounds are distinguished phonemically. For the most recent update of the IPA, see [40.37].

Given the availability of a phonetic or phonemic representation for language sounds, there are various ways in which this information can help to differentiate among languages, including:

1. *Phonemic inventory:* It is possible to distinguish some languages from one another based on the presence of a phoneme or phone that appears in one language but not the other. Phoneme inventories range from a low of 11 (for Rotokas, a Papuan language; see [40.38]) to a high of a hundred or more (in certain Khoisan languages of southern Africa [40.39]). It is likely that no two languages share exactly the same phoneme inventory. Furthermore, even if two languages were to share a common set of phonemes, the phonemes themselves will likely differ in their relative frequency patterns [40.22].
2. *Broad class inventory:* Patterns of broad phonetic categories (e.g., vowels, fricatives, plosives, nasals, and liquids) have also been utilized to distinguish among languages in an attempt to avoid the need for fine phonetic recognition. For example, *Muthusamy* [40.22] evaluated the use of seven broad phonetic categories (vowel, fricative, stop, closure or silence, prevocalic sonorant, intervocalic sonorant, and postvocalic sonorant). However, *Hazen* [40.23] found that, even though the broad class phone

recognizers tend to be more accurate than the more-traditional finer-grained phone recognizers, using broad class inventories leads to a lower-accuracy language identification system.

3. *Phonotactics:* Phonotactics refers to the arrangements of phones or phonemes within words. Even if two languages were to share a common phoneme inventory, it is likely that they would differ in phonotactics. The first proponents of using phonotactic features were *House* and *Neuburg* [40.40], who believed accurate language identification could be achieved by making use of the statistics of the linguistic events in an utterance, in particular, language-specific phonetic sequence constraints. (They suggested using a sequence of broad phonetic classes as a way of obtaining more-reliable feature extraction across languages, although this has been found to result in less accurate language identification systems [40.23].) An implementation using this approach would typically involve several steps, where the first is to map an utterance to a sequence of phonetic labels (i. e., tokenization into phones), which would then be used to identify the language based on the observed n-grams.

   There are several variants of this approach to audio based language identification. Phone-based LID uses a single-language phone recognizer trained for some arbitrary language with sufficient resources (not necessarily one of the target languages) to tokenize the speech input into *phones* for that language, followed by the use of n-gram probabilistic language models (one for each target language) to calculate the likelihood that the symbol sequence was produced in each of the target languages, with the highest likelihood language being selected. The parallel phone recognition followed by language model (PPRLM) is similar except that it uses phone recognizers from several languages, together with some method to normalize and combine the results from the parallel streams [40.41].

   A third variant would be to train a phone recognizer on a broad-coverage phonetic database (such as that in [40.35]). To reduce the time and cost to develop speech systems in a new language, researchers have been investigating the development and use of a multilingual phone set that represents sounds across the languages to be modeled. *Schultz* and *Waibel* [40.42] in their research on multilingual speech recognition defined a phone set covering 12 languages. They assume that *the articulatory representations of phonemes are so similar across languages, that phonemes can be considered as units which are independent from the underlying language* (p. 1) [40.43].

   Hazen and *Zue* [40.23, 32] developed a LID system that was based on 87 language-independent phone units that were obtained by hand clustering approximately 900 phone labels found in training transcriptions. Researchers at the computer sciences laboratory for mechanics and engineering sciences (LIMSI) have also been investigating the use of a *universal* phone set [40.44, 45] and have attempted to identify objective acoustic criteria for clustering the language-dependent phones. *Corredor-Ardoy* et al. [40.46], found that their LID system using a language-independent set (based on clustering) performed as well as their best methods using language-dependent phones. *Ma* and *Li* [40.47] evaluated the use of a universal sound recognizer to transcribe utterances into a sequence of sound symbols that act as a common phone set for all of the languages to be identified. They then used statistics related to the large-span co-occurrence of the sound patterns, which they dubbed the bag-of-sounds approach, to identify the language of the utterance.

4. *Articulatory features:* Speech can be characterized by parallel streams of articulatory features which are used in concert to produce a sequence of phonemes. These features could be exploited to differentiate one language from another. For example, the phoneme /t/ can be realized either with or without aspiration, with a dental or alveolar closure, and with lips rounded or not [40.23]. *Kirchhoff* and *Parandekar* [40.48] utilized a set of pseudoarticulatory classes that were designed to capture characteristics of the speech production process, including: manner of articulation, consonantal place of articulation, vocalic place of articulation, lip rounding, front–back tongue position, voicing, and nasality. They developed an alternative approach to audio-based language identification that was based on the use of parallel streams of these subphonemic events together with modeling of some of the statistical dependencies between the streams.

Syllable Structure. A syllable is a unit of pronunciation that is larger than a single sound, composed of a peak of sonority (usually a vowel, but sometimes a sonorous consonant), bordered by troughs of sonority (typically consonants) [40.33]. Languages can be characterized by common syllable types, typically defined in terms of sequences of consonants (C) and vowels (V). However, it should be noted that breaking sequences of Cs and

Vs into syllables – the process of syllabification – is often difficult, and even controversial [40.49, 50]. However, since languages generally allow or disallow certain types of syllable structures (e.g., Slavic languages often have complex consonant clusters in contrast to Asian languages), this representation may help in discriminating among languages. For example CCCCVC is a valid syllable type in Russian, but not in most other languages.

Zhu et al. [40.51] developed a LID system whose acoustic decoder produces syllable streams, which they then used in syllable-based (rather than phone-based) *n*-gram language models. Accents of foreign speakers of English manifest themselves differently given their position within the syllable, a fact that has been used to improve accent identification [40.52].

Prosodic Information. The duration, pitch, and stress patterns in one language often differ from another. For example, different languages have distinct intonation patterns. In stress languages, pitch is often one correlate of stress used to mark syllable prominence in words (and in pitch accent languages, such as Swedish, the primary correlate); whereas, in tone languages, a change in the meaning of a word is signalled by the tone on the syllables or other tone-bearing units (e.g., Mandarin Chinese or Thai). For stress languages, patterns of stress can provide an important cue for discriminating between two languages. Some of the stress patterns are initial stress (e.g., Hungarian), penultimate stress (e.g., Polish or Spanish), final stress (e.g., French or Turkish), and mixed stress (e.g., Russian or Greek). Prosodic cues of duration can also be potentially useful. For example, some languages (such as Finnish) distinguish long and short vowels and/or consonants.

One ostensibly useful typology investigated by both linguists and psychologists involves the rhythm of a language, where a distinction is made between stress-timed languages (in which stressed syllables are longer than unstressed syllables, all else being equal, e.g., English), syllable-timed (each syllable has comparable time duration, e.g., French), and mora-timed (each mora has essentially constant duration, e.g., Japanese) languages [40.53]. Although this classification remains controversial [40.54], rhythmic modeling has been investigated by *Rouas* et al. [40.55, 56] for language identification. Their rhythm model was able to discriminate fairly accurately between languages on a read speech corpus, but less well on a spontaneous speech corpus [40.55]. Indeed, extracting reliable prosodic cues from spontaneous speech is a challenge due to its variability.

In language identification systems that utilize prosodic features, these features are typically combined with other knowledge sources to achieve reasonable accuracy. *Muthusamy* [40.22] was able to incorporate pitch variation, duration, and syllable rate features into his LID model. *Hazen* and *Zue* [40.23, 32] integrated duration and pitch information into their LID model, with the duration model being more accurate on its own than the pitch model. *Tong* et al. [40.57] successfully integrated prosodic features (i. e., duration and pitch) with spectrum, phonotactic, and bag-of-sounds features, where pitch variation and phoneme duration were especially useful for short speech segments.

Lexical Information. Each language has its own vocabulary, which should help in identifying a language more reliably. For speech inputs, this would require the availability of a speech recognition system for each of the candidate languages, along with the requisite training, tuning, and evaluation materials needed to ensure the speech model is adequate. *Schultz* et al. [40.58, 59] developed a LID system for four languages based on large-vocabulary continuous speech recognition (LVCSR). They found that word-based systems with trigram word language modeling significantly outperform phone-based systems with trigram phone modeling on the four-language task, suggesting that the lexical level provides language discrimination ability, even when word error rates are fairly high. *Matrouf* et al. [40.60] found that incorporating lexical information with a phone-based approach yielded relative error reductions of 15–30%, and that increasing lexical coverage for a language had a positive effect on system performance. *Hieronymus* and *Kadambe* [40.61] constructed a LID system based on LVCSR for five languages (English, German, Japanese, Mandarin Chinese, and Spanish), obtaining 81% and 88% correct identification given 10 and 50 second utterances, respectively, without using confidence measures and 93% and 98% correct with confidence measures. Although each language clearly has its own vocabulary that enables language identification systems to discriminate among a candidate set of languages more effectively, for spoken languages, this information is generally quite expensive to obtain and use.

Morphology. Morphology is a branch of grammar that investigates the structure of words [40.33]. The field of morphology is divided into two subfields: inflectional morphology, which investigates affixes that signal grammatical relationships that do not change the grammatical class of a word (e.g., affixes marking tense, number, and case) and derivational morphology, which focuses on word formation involving affixes, such as *-ment*, that can be used to create a new word form with a possibly different grammar class, such as the noun *amendment* derived from the verb *amend*. Since languages form words in a variety of different ways, morphology could provide an excellent cue for automatic language identification. For example, one could use common suffixes to discriminate among some of the Romance languages (e.g., *-ment* in French, *-miento* in Spanish, *-mento* in Portugese, and *-mente* in Italian). Although in speech the morphology of a word is covered in part by phonotactics, with morphological knowledge of its candidate languages, a LID system could focus on specific portions of words when discriminating between two languages.

Syntax. Languages and dialects also differ in the ways that words are arranged to create a sentence. They differ in the presence or absence of words with different parts of speech, as well as in the ways that words are marked for various types of roles in a sentence. In conjunction with other kinds of information (e.g., accents), errors in grammatical usage could provide a helpful cue for identifying the first language of someone speaking a second language. For example, someone who learned Mandarin as a first language would tend to make determiner (deletion, substitution, and insertion) and agreement errors in English or German.

Languages also often differ from each other in the word order of a sentence's subject (S), verb (V), and object (O). For example, English is considered to be an SVO language because the subject typically appears before the verb, which occurs before the object; whereas, Japanese is an SOV language. Even if two languages have the same word form, it is likely that the word would appear in very different word contexts across languages. Although words may be sufficient to distinguish among languages, syntax could play an especially important role for discriminating among dialects of a language.

Language and dialect identification systems that are based on LVCSR would utilize an acoustic model, a dictionary, and a language model for each language or dialect in the candidate set. The language models utilize word co-occurrence statistics that capture some aspects of the candidate language's syntactic structure. These systems could be expanded to utilize syntax more directly by using structured language models (e.g., [40.62, 63]).

Other Information. Higher-level knowledge sources such as semantics and pragmatics are rarely used by audio-based LID systems, although this type of knowledge could potentially help. In addition, information about the source of the speech data (e.g., country of origin of a broadcast news show) could be used to narrow down the language choices.

Research on automatic language identification suggests that the more knowledge that goes into a decision about which language corresponds to an audio sample, the greater the accuracy; hence, knowledge integration is important. However, there is a trade-off between accuracy and efficiency, and furthermore, there is a need for resources to support the knowledge brought into the automatic system. Much of the work has struck an engineering balance in addressing this problem; they use the resources that can be obtained simply and reliably for the set of languages to be discriminated among.

It is important to note that the length of an audio sample (the amount of speech available) will impact the knowledge sources that can be reliably used in determining its language. The smaller the samples used, the less likely that a key piece of higher-level knowledge will be available to discriminate a particular language from the others.

40.4 Human Language Identification

A human who knows the language being spoken is capable of positively identifying short samples of speech quickly and accurately. Even if they do not know the language, people can make sound decisions about the identity of a language given some exposure to the language, and with some training about cues that differentiate a set of candidate languages, this capability can be improved and expanded.

There have been several experiments reported in the literature that consider human ability in a scenario where the decision is based on a combination of their prior knowledge about certain languages (a variable that is difficult to control) and a limited amount of online training for the languages being identified. *Muthusamy* et al. [40.22, 64] had human subjects listen to short samples of 10 different languages and guess the language of the sample. In general, he found that familiarity with a language was an important factor affecting accuracy, as was the length of the speech sample (with longer samples leading to greater accuracy). Subjects were able to improve their LID accuracy only slightly over time with feedback, but they were quite aware of the cues they used for language discrimination (e.g., phonemic inventory, word spotting, and prosody).

Maddieson and *Vasilescu* [40.65] examined the effect of exposure to academic linguistics on language identification accuracy of five language. Subjects with more than *passing* familiarity with the language were excluded from the study. They found that prior casual exposure to a language and linguistic education level (ranging from no linguistic training to a PhD) were not effective predictors of performance on a five-language identification task; however, they found that linguistic training did predict improved performance on a language discrimination task (in which subjects were asked to decide if a sample was one of the five target languages, similar to one of those languages, or unlike them).

Several human studies have been conducted in an attempt to determine what types of information people can effectively utilize when making decisions about the identity of a language. These experiments involve the presentation of speech stimuli that were obtained by modifying the speech samples presented to the subjects. For example, *Mori* et al. [40.66] found that their subjects were able to identify two languages (Japanese and English) fairly reliably even when segmental information was reduced using signal editing techniques. They argue that other cues such as intensity and pitch are being used to make these judgments. *Navratil* [40.67] evaluated the importance of various types of knowledge, including lexical, phonotactic, and prosodic, by humans asked to identify the language (Chinese, English, French, German, or Japanese) of a speech sample. Subjects were presented unaltered speech samples, samples containing randomly ordered syllables from speech samples, and samples for which the spectral shape was flattened and vocal-tract information removed (leaving F_0 and amplitude). Navratil found that humans on six second samples were far more accurate at identifying unaltered speech samples (96%) than samples with shuffled syllables (73.9%), and were more accurate with the shuffled syllable samples than samples with only prosodic cues remaining (49.4%). Based on these experiments, it appears that the lexical and phonotactic information provides discriminative information that is used more reliably by humans.

None of the subjects in these listening experiments were explicitly trained to identify cues to discriminate one language from another, hence, these experiments did not explore the full range of human capability that could be achieved with training on a particular set of languages to be identified. People can identify the language of an audio input fairly reliably when they do not speak/understand the language by being taught to use a variety of cues that are discriminative for a language or language family. Some combination of the following sorts of clues can be used to identify a particular language or to narrow the possibilities down to a smaller set of languages:

- general impression (‘gestalt’), i. e., what a given language sounds like, which may help narrow the language down to a geographic area or a language family
- stress patterns, where these may be most reliably discerned at pause boundaries or on assimilated loan words
- vowel and/or consonant durations
- the presence or absence of nasalization on (some) vowels
- the presence or absence of lexical tone
- syllable structure, particularly the presence or absence of consonant clusters
- the presence of unusual sounds, such as front rounded vowels, glottalized consonants, clicks, or retroflexed consonants
- reduplication (particularly full-word reduplication)
- the presence of common words, particularly short high-frequency words that are easily recognizable (e.g., determiners, prepositions)

Some of these features would obviously be difficult for computers to use, including the ‘gestalt’ of the language or detecting reduplication. Other features are similar to what programs doing spoken language identification already utilize, e.g., the use of consonant clusters (phonotactics). Still others of these features suggest possible directions for future work in automated language ID, for example recognizing unusual sounds.

There is also a more-general difference in the methodology used by humans and speech-based LID algorithms. Most speech-based LID systems do not utilize the high-level knowledge that people tend to use. Also, most of the automatic LID algorithms tend to process and combine evidence from the entire speech stream when making a decision about the identity of a language; whereas, humans rely on very specific cues taken from small portions of the sample to refine their hypotheses.

For example, two of the cues used by human experts for LID on texts are the pairings of diacritics with base characters and the presence of short (but common) words. Diacritics are especially difficult for image-based language identification systems, which can have difficulty differentiating them from noise in the image. In contrast, the primary cue used by computers for computer encoded text or in speech is the statistical frequency character or phone *n*-grams. *n*-grams would be hard for people to learn, with the exception of unusual single letters or sounds and morphemes which are cognate with English morphemes; anything more probably requires the user to refer to a *cheat sheet*. Even harder for people is computing the statistics of *n*-grams; judgments with a finer granularity than *frequent* or *rare* would be difficult for people to make.

In summary, while it is possible for computers to make use of more of the characteristics of languages than humans use to identify them, it is probably not practical to teach people to use the methodology used by automatic LID systems to do language identification.

40.5 Text as a Source of Information on Spoken Languages

Given the difficulty in building comparable speech corpora for a significant number of languages, and in particular for rare languages, another source of information that might be mined to learn more about the spoken form of a language is its written form. (While most unwritten languages have small speaker populations, and the total number of speakers of unwritten languages is much smaller than the number of speakers of written languages, there are still significant numbers of languages of the world – perhaps more than half – which are unwritten.) It is comparatively easy to build a text corpus for a written language; however, there are several issues that affect whether the textual data will be useful in characterizing the spoken language. An important issue is how close the written language is to its typical spoken form. There are many aspects to this question, but two of the most significant issues are diglossia and complex orthographies.

A diglossic language situation exists when two (or more) forms of a language coexist, and the forms diverge to a significant degree; typically one form is perceived as *high* (correct), and the other as *low* (vulgar). For our purposes, diglossia is relevant when the high and low varieties correspond to the written and the spoken languages respectively, where they differ significantly in style and vocabulary. Tamil is a typical example; the written form of the language is considered classical, and the spoken forms (there is more than one dialect) are considered low. In such a situation, written corpora may not be representative of the spoken form of the language.

Arabic is another important example of a diglossic situation: the written form, known as modern standard Arabic (MSA), is taught in schools; whereas, the spoken varieties – of which the Ethnologue [40.1] lists nearly 40 – are strikingly different from MSA in vocabulary, morphology, and phonology (the latter with reference to how MSA is generally read out loud, for instance on news broadcasts).

As for the complex orthography issue, for our purposes a complex orthography is one in which the written forms of words do not have a direct mapping to the spoken forms. English is a notorious example of this, and so to a lesser extent is French. A related issue is orthographies which undermark phonemic distinctions in the spoken language. Written Arabic is an example, since the short vowels are not normally written, resulting in significant ambiguity: multiple morphological analyses, each corresponding to different pronunciations, are possible for a large percentage of the words in running Arabic (MSA) text.

Many orthographies that are complex today are so only because the spoken language has changed faster than the written language. Orthographies that have been developed in the recent past tend to be less complex, i. e., they usually map more or less directly to the phonemes of the spoken language (or to a standard dialect of the language), and can therefore be said to be *phonemic*. However, some otherwise phonemic modern orthographies fail to discriminate a subset of the phonemic contrasts of the language, whether in practice or in principle. For example, while Yoruba is supposed to be written with tone marks for the high and low tones (and optionally for the mid tone), in practice these are often omitted, as are the dots under the Yoruba letters ‘e’ and ‘o’, intended to indicate a more-open vowel than the same letters without the dot. Finally, some orthographies omit certain phonemic distinctions (e.g., tone) on the principle that the distinction in question is not important enough in that language.

Assuming that a language's orthography is close to phonemic, it is also necessary to know the mapping from individual characters (or sequences of characters) to a phonemic representation. This is because orthographies do not always use the letters in a standard way. In Hungarian, for example, the letter 's' represents an alveopalatal (like the digraph 'sh' in English), and the digraph 'sz' represents an alveolar fricative (like the English letter 's' in most words); Polish represents these sounds in exactly the opposite way. Fortunately, this information about alphabets can generally be obtained.

In summary, then, it would be possible at least in principle to expand the number of spoken languages that can be investigated and characterized linguistically by using more readily obtainable text corpora in place of speech corpora. Whether these corpora can be used to enhance speech-based LID systems is an open question.

40.6 Summary

In this chapter, we have discussed a variety of knowledge sources that could be utilized to characterize a spoken language in order to distinguish it automatically from other languages. We discussed differences between a language and a dialect, and also described some of the variability in a language that could potentially challenge LID algorithms. Currently available corpora for evaluating language identification systems have only *scratched the surface* of the possible space of languages that could be investigated. Most state-of-the-art systems are able to detect tens of languages as opposed to the 100 or more that would be required by Language Line Services. It remains to be seen whether acoustic- and phonotactic-based systems will effectively scale up to handle these one hundred plus languages. As the number of languages and dialects increase, it is likely that systems will need to utilize more linguistic insight to achieve accuracies comparable to those obtained over a smaller set of languages. These larger systems could utilize more lexical information to achieve target accuracies; however, this knowledge source comes with a high cost for system development.

We also discussed several experiments reported in the literature that investigated a person's ability to accurately identify a spoken language. None of these studies involved a situation where participants were trained to accurately identify a language based on salient language-specific high-frequency cues. We enumerated some cues that have been commonly used, some of which overlap with features used by automatic LID systems. However, since some of the cues that humans use would be difficult to incorporate into an automatic LID system (e.g., a general impression of the language), it is interesting to contemplate whether there would be some way to build a partnership between trained humans and LID systems.

We ended this chapter by discussing conditions under which textual materials could potentially augment our knowledge of a spoken language, in particular, a rare language. There are a number of factors that impact the correspondence between spoken and written language forms; however, if there is a good correspondence, the written form could be used to gain a deeper understanding of the kinds of features that would help discriminate the language from others.

References

40.1 R.G. Gordon Jr. (ed): *Ethnologue: Languages of the World*, 15th edn. (SIL International, Dallas 2005), online version: http://www.ethnologue.com

40.2 Language Line Services, http://www.languageline.com

40.3 A.G. Chejne: *The Arabic Language Its Role in History* (University of Minnesota Press, Minneapolis 1969)

40.4 N. Haeri: *Sacred Language Ordinary People: Dilemmas of Culture and Politics in Egypt* (Palgrave MacMillan, New York 2003)

40.5 C.P. Masica: *The Indo-Aryan Languages* (Cambridge Univ. Press, Cambridge 1991)

40.6 B.A. Fennel: *A History Of English: a Sociolinguistic Approach*, Blackwell Textbooks in linguistics, Vol. 17 (Blackwell, Malden 2001)

40.7 P. Constable, G. Simons: An Analysis of ISO 639: Preparing the way for advancements in language identification standards, 20th International Unicode Conference (2002), available as http://www.ethnologue.com/14/iso639/An_analysis_of_ISO_639.pdf

40.8 ISO/DIS 639-3, http://www.sil.org/iso639-3/default.asp

40.9 D. Barton: *Literacy: An Introduction to the Ecology of Language* (Blackwell, Malden 1994)

40.10 C.W. Kreidler: *Describing Spoken English: An Introduction* (Routledge, London 1997)

40.11 A. McMahon: *Lexical Phonology and the History of English*, Cambridge Studies in Linguistics 91 (Cambridge Univ. Press, Cambridge 2000)

40.12 The SCOTS Project, http://www.scottishcorpus.ac.uk

40.13 R. Sproat, T.F. Zheng, L. Gu, D. Jurafsky, I. Shanfran, J. Li, Y. Zheng, H. Zhou, Y. Su, S. Tsakalidis, P. Bramsen, D. Kirsch: *Dialectical Chinese Speech Recognition: Final Technical Report* (Johns Hopkins University, Baltimore 2004), CLSP

40.14 J.C. Wells: *Accents of English*, Vol. 3 (Cambridge Univ. Press, Cambridge 1982)

40.15 P. Trudgill, J. Hannah: *International English: A Guide to Varieties of Standard English* (Oxford Univ. Press, New York 2002)

40.16 A. Sakaguchi: Towards a clarification of the function and status of international planned languages. In: *Status and Function of Language and Language Varieties*, ed. by U. Ammon (Walter de Gruyter, Berlin 1989) p. 399

40.17 W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H.J. Nock, M. Riley, M. Saralar, C. Wooters, G. Zavaliagkos: Pronunciation modelling using a handlabelled corpus for conversational speech recognition, Proc. ICASSP (1998) pp. 313–316

40.18 E. Fosler-Lussier, N. Morgan: Effects of speaking rate and word frequency on conversational pronunciations, Speech Commun. **29**, 137–158 (1999)

40.19 S. Greenberg: Speaking in shorthand – A syllablecentric perspective for understanding pronunciation variation, Speech Commun. **29**, 159–176 (1999)

40.20 Multi-Language Telephone Speech Corpus Distribution, http://www.ldc.upenn.edu/Catalog/readme_files/ogi readme.html, January 1994.

40.21 OGI Multilanguage Corpus, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S17

40.22 Y.K. Muthusamy: *A Segmental Approach to Automatic Language Identification* (Oregon Graduate Institute of Science and Technology, Beaverton 1993), Ph.D. Thesis

40.23 T.J. Hazen: *Automatic Language Identification Using a Segment-Based Approach* (MIT, Cambridge 1993), Masters Thesis

40.24 LDC Projects, http://www.ldc.upenn.edu/Catalog/project_index.jsp

40.25 CLSU: 22 Languages Corpus, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S26

40.26 MIXER Telephone Study, http://mixer.ldc.upenn.edu

40.27 D. Cimarusti, R.B. Ives: Development of an automatic identification system of spoken languages: Phase 1, Proc. ICASSP (1982) pp. 1661–1663

40.28 P.A. Torres-Carrasquillo, D.A. Reynolds, J.R. Deller Jr.: Language identification using Gaussian mixture model tokenization, Proc. ICASSP (2002) pp. 757–760

40.29 M.A. Zissman: Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models, Proc. ICASSP (1993) pp. 399–402

40.30 K.K. Wong, M. Siu: Automatic language identification using discrete hidden Markov model, Interspeech (2004) pp. 1633–1636

40.31 W.M. Campbell, E. Singer, P.A. Torres-Carrasquillo, D.A. Reynolds: Language recognition with support vector machines, Proceedings of Odyssey: The Speaker and Language Recognition Workshop (2004) pp. 41–44

40.32 T.J. Hazen, V. Zue: Segment-based automatic language identification, J. Acoust. Soc. Am. **101**(4), 2323–2331 (1997)

40.33 D. Crystal: *A Dictionary of Linguistics and Phonetics*, 2nd edn. (Basil Blackwell, Oxford: 1985)

40.34 M. Kenstowicz: Generative Phonology. In: *Encyclopedia of Language and Linguistics*, ed. by K. Brown (Elsevier, Amsterdam 2005), 2nd edn.

40.35 P. Ladefoged: *Vowels and Consonants: An Introduction to the Sounds of Languages* (Blackwell, Oxford 2005)

40.36 The International Phonetic Association, http://www.arts.gla.ac.uk/IPA/index.html

40.37 Reproduction of the International Phonetic Alphabet, http://www.arts.gla.ac.uk/IPA/ipachart.html , 2005.

40.38 I. Maddieson: *Sound Patterns of Language* (Cambridge Univ. Press, Cambridge 1984)

40.39 A. Traill: *Phonetic and Phonological Studies of !Xóõ bushman* (Helmut Buske, Hamburg 1985)

40.40 A.S. House, E.P. Neuberg: Toward automatic identification of the languages of an utterance: preliminary methodological considerations, J. Acoust. Soc. Am. **62**(3), 708–713 (1977)

40.41 M.A. Zissman: Comparison of four approaches to automatic language identification of telephone speech, IEEE Trans. Speech Audio Process. **4**(1), 31–44 (1996)

40.42 T. Schultz, A. Waibel: Language independent and language adaptive acoustic modeling for speech recognition, Speech Commun. **35**(1-2), 31–51 (2001)

40.43 A. Black, T. Schultz: Speaker clustering for multilingual synthesis, Proceedings of the ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (2006)

40.44 M. Adda-Decker, F. Antoine, P.B. de Mareuil, I. Vasilescu, L. Lamel, J. Vaissiere, E. Geoffrois, J.-S. Liénard: Phonetic knowledge, phonotactics and perceptual validation for automatic language identification, International Congress of Phonetic Sciences (2003)

40.45 P.B. de Mareüil, C. Corredor-Ardoy, M. Adda-Decker: Multi-lingual automatic phoneme clustering, Int. Congress Phonetic Sci. (1999) pp. 1209–1213

40.46 C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel: Language Identification with Language-independent Acoustic Models, Proc. Eurospeech (1997) pp. 355–358

40.47 B. Ma, H. Li: Spoken language identification using bag-of-sounds, International Conference on Chinese Computing (2005)

40.48 K. Kirchhoff, S. Parandekar: Multi-stream statistical language modeling with application to automatic language identification, Proceedings of the 7th European Conference on Speech Communication and Technology Proceedings of Eurospeech (2001) pp. 803–806

40.49 J. Blevins: The syllable in phonological theory. In: *The Handbook of Phonological Theory*, Blackwell Handbooks in Linguistics, Vol. 1, ed. by J.A. Goldsmith (Blackwell, Oxford 1995) pp. 206–244

40.50 M. Kenstowicz: *Phonology in Generative Grammar*, Blackwell Textbooks in Linguistics, Vol. 7 (Blackwell, Oxford 1994)

40.51 D. Zhu, M. Adda-Decker, F. Antoine: Different size multilingual phone inventories and context-dependent acoustic models for language identification, Interspeech (2005) pp. 2833–2836

40.52 K. Berkling, M. Zissman, J. Vonwiller, C. Cleirigh: Improving accent identification through knowledge of English syllable structure, Proceedings of the 5th International Conference on Spoken Language Processing (1998) pp. 89–92

40.53 E. Grabe, E.L. Low: Durational variability in speech and the rhythm class hypothesis. In: *Laboratory Phonology*, ed. by C. Gussenhoven, N. Warner (Mouton de Gruyter, Berlin 2002) pp. 515–546

40.54 R.M. Dauer: Stress-timing and syllable-timing reanalysed, J. Phonet. **11**, 51–62 (1983)

40.55 J. Rouas, J. Farinas, F. Pellegrino, R. Andre-Obrecht: Modeling prosody for language identification on read and spontaneous speech, Proc. ICASSP (2003) pp. 40–43, vol. 6

40.56 J. Rouas, J. Farinas, F. Pellegrino, R. André-Obrecht: Rhythmic unit extraction and modelling for automatic language identification, Speech Commun. **47**(4), 436–456 (2005)

40.57 R. Tong, B. Ma, D. Zhu, H. Li, E.S. Chang: Integrating acoustic, prosodic and phonotactic features for spoken language identification, Proc. ICASSP (2006) pp. 205–208

40.58 T. Schultz, I. Rogina, A. Waibel: Experiments with LVCSR based language identification, Proceedings of the Speech Symposium SRS XV (1995)

40.59 T. Schultz, I. Rogina, A. Waibel: LVCSR-based language identification, Proc. ICASSP (1996) pp. 781–784

40.60 D. Matrouf, M. Adda-Decker, L. Lamel, J. Gauvain: Language identification incorporating lexical information, Proceedings of the 5th International Conference on Spoken Language Processing (1998) pp. 181–184

40.61 J. Hieronymus, S. Kadambe: Robust spoken language identification using large vocabulary speech recognition, Proc. ICASSP (1997) pp. 779–782

40.62 C. Chelba: *Exploiting Syntactic Structure for Natural Language Modeling* (Johns Hopkins University, Baltimore 2000), Ph.D. thesis

40.63 W. Wang, M. Harper: The SuperARV language model: Investigating the effectiveness of tightly integrating multiple knowledge sources, Proceedings of Conference of Empirical Methods in Natural Language Processing (2002) pp. 238–247

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

40.65 I. Maddieson, I. Vasilescu: Factors in human language identification, Proceedings of the 5th International Conference on Spoken Language Processing (2002) pp. 85–88

40.66 K. Mori, N. Toba, T. Harada, T. Arai, M. Komatsu, M. Aoyagi, Y. Murahara: Human language identification with reduced spectral information, Proceedings of the 6th European Conference on Speech Communication and Technology (1999) pp. 391–394

40.67 J. Navratil: Spoken language recognition – A step towards multilinguality in speech processing, IEEE Trans. Speech Audio Process. **9**(6), 678–685 (2001)