INVITED PAPER

Spoken Language Recognition: From Fundamentals to Practice

This paper provides an introductory tutorial on the fundamentals and the state-of-the-art solutions to automatic spoken language recognition, from both phonological and computational perspectives. It also gives a comprehensive review of current trends and future research directions.

By Haizhou Li, *Senior Member IEEE*, Bin Ma, *Senior Member IEEE*, and Kong Aik Lee, *Member IEEE*

ABSTRACT | Spoken language recognition refers to the automatic process through which we determine or verify the identity of the language spoken in a speech sample. We study a computational framework that allows such a decision to be made in a quantitative manner. In recent decades, we have made tremendous progress in spoken language recognition, which benefited from technological breakthroughs in related areas, such as signal processing, pattern recognition, cognitive science, and machine learning. In this paper, we attempt to provide an introductory tutorial on the fundamentals of the theory and the state-of-the-art solutions, from both phonological and computational aspects. We also give a comprehensive review of current trends and future research directions using the language recognition evaluation (LRE) formulated by the National Institute of Standards and Technology (NIST) as the case studies.

KEYWORDS | Acoustic features; calibration; classifier; fusion; language recognition evaluation (LRE); phonotactic features; spoken language recognition; tokenization; vector space modeling

I. INTRODUCTION

Spoken language recognition refers to the automatic process that determines the identity of the language spoken in a speech sample. It is an enabling technology for a wide range of multilingual speech processing applications, such as spoken language translation [141], multilingual speech recognition [116], and spoken document retrieval [23]. It is also a topic of great interest in the areas of intelligence and security for information distillation.

Humans are born with the ability to discriminate between spoken languages as part of human intelligence [147]. The quest to automate such ability has never stopped [69], [70], [89], [96], [150]. Just like any other artificial intelligence technologies, spoken language recognition aims to replicate such human ability through computational means. The invention of digital computers has made this possible. A key question is how to scientifically measure the individuality of the diverse spoken languages in the world. Today, automatic spoken language recognition is no longer a part of science fiction. We have seen it being deployed for practical uses [23], [81], [116], [141].

It is estimated that there exist several thousands of spoken languages in the world [29], [30], [58]. The recent edition of the *Ethnologue*, a database describing all known living languages [71], has documented 6909 living spoken languages. Text-based language recognition has traditionally relied on distinct textual features of languages such as words or letter substrings. It was arguably established in 1967 by Gold [41] as a closed class experiment, in which human subjects were asked to classify a given test document into one of the languages. It was not until the 1990s, however, when people resorted to statistical techniques that formulate the problem as a text categorization problem [3], [22], [37]. For languages that use the Latin alphabet, text-based language recognition has attained reasonably good performance, thus it is considered a solved problem [83]. As words and letter substrings are the manifestation of lexical–phonological rules of a language, this research has led to the conjecture that a spoken

Manuscript received May 22, 2012; revised September 19, 2012; accepted December 10, 2012. Date of publication February 6, 2013; date of current version April 17, 2013.
H. Li is with the Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore 138632, and also with the University of New South Wales, Kensington, NSW 2052, Australia (e-mail: hli@i2r.a-star.edu.sg).
B. Ma and **K. A. Lee** are with the Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore 138632 (e-mail: mabin@i2r.a-star.edu.sg; kalee@i2r.a-star.edu.sg).
Digital Object Identifier: 10.1109/JPROC.2012.2237151

language can be characterized by its lexical–phonological constraints.

In practice, spoken language recognition is far more challenging than text-based language recognition because there is no guarantee that a machine is able to transcribe speech to text without errors. We know that humans recognize languages through a perceptual or psychoacoustic process that is inherent in the auditory system. Therefore, the type of perceptual cues that human listeners use is always the source of inspiration for automatic spoken language recognition [147].

Human listening experiments suggest that there are two broad classes of language cues: the prelexical information and the lexical semantic knowledge [147]. Phonetic repertoire, phonotactics, rhythm, and intonation are all parts of the prelexical information [108]. Although one may consider phones and phonotactics as segmental, and rhythm and intonation as suprasegmental, they are all very different from the lexical semantic knowledge that the words encapsulate, such as semantic meanings and conceptual representations. There is no doubt that both prelexical and lexical semantic knowledge contribute to the human perceptual process for spoken language recognition [108].

Studies in infants' listening experiments have revealed that when infants have not gained a great deal of lexical knowledge, they successfully rely on prelexical cues to discriminate against languages [108]. In fact, when an adult is dealing with two unfamiliar languages, one can only use prelexical information. But as the infant's language experience enriches or as the adult is handling familiar languages, lexical semantic information will start to play a very important, sometimes determining, role [147]. While we know that it requires a major effort to command the lexical usage of an entirely new language, in human listening studies, we have observed that human subjects are able to acquire prelexical information rapidly for language recognition purposes.

The relative importance of perceptual cues for language recognition has always been a subject of debate. Studies in automatic spoken language recognition have confirmed that acoustic and phonotactic features are the most effective language cues [97], [150]. This coincides with the findings from human listening experiments [96], [137]. While the term "acoustic" refers to physical sound patterns, the term "phonotactic" refers to the constraints that determine permissible syllable structures in a language. We can consider acoustic features as the proxy of phonetic repertoire and call it acoustic–phonetic features. On the other hand, we see phonotactic features as the manifestation of the phonotactic constraints in a language. Therefore, in this paper, we would like to introduce the spoken language recognition techniques that are mainly based on acoustic and phonotactic features, which represent the mainstream research [87], [102], [150]. In what follows, we use the term language recognition for brevity.

We will provide insights into the fundamental principles of the research problem, and a practical applicability analysis of different state-of-the-art solutions to language recognition. We will also make an attempt to establish the connection across different techniques. Advances in feature extraction, acoustic modeling, phone *n*-gram modeling, phone recognizers, phone lattice generation, vector space modeling, intersession compensation, and score calibration and fusion have contributed to the state-of-the-art performance [13], [73], [121]. To benefit from a wealth of publications related to the National Institute of Standards and Technology (NIST) language recognition evaluations (LREs) [85], [86], [88], [89], [102], we will demonstrate the working principles of various techniques using the NIST LREs as case studies.

The rest of the paper is organized as follows. We will elaborate on the principles, problems, and recent advances in language modeling in Section II. We will introduce the phonotactic approaches and acoustic approaches in Sections III and IV, respectively. Advanced topics in system developments related to Gaussian back–ends, multiclass logistic regression, score calibration and fusion, and language detection will be discussed in Section V. Section VI will be devoted to the NIST LRE paradigm, and, finally, an overview of current trends and future research directions will be discussed in Section VII.

II. LANGUAGE RECOGNITION PRINCIPLES

A. Principles of Language Characterization

The first perceptual experiment measuring how well human listeners can perform language identification was reported in [96]. It was concluded that human beings, with adequate training, are the most accurate language recognizers. This observation still holds after 15 years as confirmed again in [137], provided that the human listeners speak the languages.

For languages that they are not familiar with, human listeners can often make subjective judgments with reference to the languages they know, e.g., it sounds like Arabic, it is tonal like Mandarin or Vietnamese, or it has a stress pattern like German or English. Though such judgements are less precise for hard decisions to be made for an identification task, they show how human listeners apply linguistic knowledge at different levels for distinguishing between certain broad language groups, such as tonal versus nontonal languages.

Studies also revealed that, given only little previous exposure, human listeners can effectively identify a language without much lexical knowledge. In this case, human listeners rely on prominent phonetic, phonotactic, and prosody cues to characterize the languages [96], [137].

The set of language cues discussed above can be illustrated according to their level of knowledge abstraction,

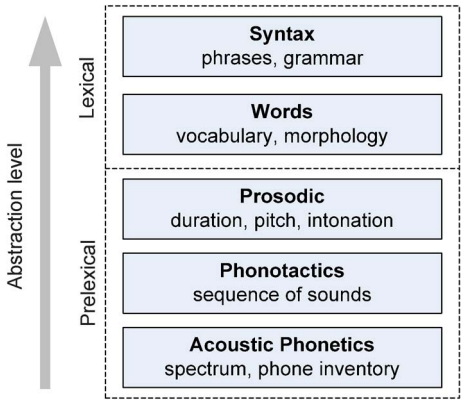

Fig. 1. ***Various levels of perceptual cues used for language recognition.***

as shown in Fig. 1. Language recognition systems are usually categorized by the features they use, such as the acoustic–phonetic approach, the phonotactic approach, the prosodic approach, and the lexical approach.

- *Acoustic phonetics*: The human speech apparatus is capable of producing a wide range of sounds. Speech sounds as concrete acoustic events are referred to as phones, whereas speech sounds as entities in a linguistic system are termed as phonemes [58]. The number of phonemes used in a language ranges from about 15 to 50, with the majority having around 30 phonemes each [28]. For example, English has a phonetic system that contains 24 consonants and 14 vowels, Mandarin has 21 consonants and ten vowels, and Spanish has 18 consonants and five vowels [147]. Phonetic repertoires differ from language to language although languages may share some common phonemes. For example, retroflex consonants are used in Mandarin but not in Spanish. The English consonant /ð/ (i.e., the voiced *th* sound) does not appear in Mandarin. These differences between phonetic repertoires imply that each language has its unique set of phonemes [58], thus acoustic–phonetic feature distributions.
- *Phonotactics*: As concluded in phonological studies, each language has its unique set of lexical–phonological rules that govern the combinations of different phonemes. The phonotactic constraints dictate the permissible phone sequences. We note that phonemes can be shared considerably across languages, but the statistics of their sequential patterns differ very much from one language to another. Some phone sequences that occur frequently in one language could be rare in another. For example, consonant clusters, e.g., /fl/, /pr/, and /str/, are commonly observed in English words, while impermissible in Mandarin. Such phonotactic constraints can be characterized by a phone *n*-gram model [55].
- *Prosody*: Prosody in general refers to suprasegmental features in running speech, such as stress, duration, rhythm, and intonation [2]. The set of interrelated prosodic features are all important characteristics of spoken languages. For example, the world's languages can be grouped into three rhythm classes: stress-timed languages such as English and other Germanic languages, syllable-timed languages such as French and Hindi, and mora-timed languages such as Japanese [109]. Lexical tone is the most prominent feature of tonal languages, such as Mandarin, Thai, and Vietnamese. Therefore, prosody appears to be useful for distinguishing between broad language classes (e.g., tonal versus nontonal languages). However, human listening experiments reported in [99], and [108] show that prosodic cues are less informative than the phonotactic one. This observation is consistent with that reported in automatic language recognition [46], [127], where it was shown that phonotactic features are far more superior to prosodic features. Furthermore, it remains a challenge to reliably extract prosodic features [100], [101]. Therefore, we do not go into details about prosodic approaches in this paper.
- *Words and syntax*: Languages contain a phonological system that governs how symbols are used to form words or morphemes, and a syntactic system that governs how words and morphemes are combined to form phrases and utterances. Each language has its unique phonological system and syntactic system that can be used for language identification, which characterizes a unique word list or a set of word *n*-grams [37]. Therefore, the lexical approach seems to be an obvious choice for language recognition. This is also encouraged by human listening experiments where humans do well when they know the languages. One idea is to run multiple large vocabulary continuous speech recognition (LVCSR) systems for different target languages in parallel [48], [92], [115]. The system that gives the highest likelihood score makes the best sense of the speech and is considered as the recognition result. The theory behind this is that, once the system knows what a person is saying, its language is obvious. This study raises an interesting question: If a multilingual LVCSR system has already recognized the language, why do we need a standalone language recognizer? In practice, just like a person would not master a language for the sake of language recognition, we are concerned about the cost effectiveness of using LVCSR for language recognition. As a result, the LVCSR-based lexical approach has not been widely used.

The use of phonetic and phonotactic cues is based on the assumption that languages possess partially

overlapping sets of phonemes. The same basis was used in constructing the international phonetic alphabet (IPA). Though there are 6909 languages in the world [71], the total number of phones required to represent all the sounds of these languages ranges only from 200 to 300 [2]. Phones that are common in languages are grouped together and given the same symbol in IPA. For instance, the GlobalPhone multilingual text and speech database [117] uses a phone set consisting of 122 consonants and 114 vowels. Three nonphonetic units (two noise models and one silence model) are also defined for modeling nonspeech events. The same phone set is used for transcribing 20 spoken languages in the GlobalPhone database.

A systematic study on the extent to which languages are separated in phonetic and phonotactic terms is possible using the GlobalPhone database, in which languages are mapped to a common set of IPA symbols. Fig. 2 compares and contrasts the phonetic histograms of two languages that are arbitrarily selected from the GlobalPhone database in the form of polar plots. The size of the pie shape in the figure indicates the normalized count of a phone in a speech utterance. The three plots in the upper panel were obtained for three different subjects speaking Czech of different contents, while the plots in the lower panel are for another three subjects speaking Portuguese of different contents. A visual inspection easily confirms that the two languages are different in terms of phonetic repertoire and the frequency count of phones, while only slight variations are observed for different instances within the same language despite different speech contents. If we treat the phonetic histogram as an empirical distribution of the language, we may apply similarity measures to provide quantifiable measurements between languages.

We can also use the GlobalPhone database to study the phonotactic differences between languages by examining how well a phone n-gram model of one language predicts the phone sequence across different languages in terms of perplexity [55]. A lower perplexity shows that a phone n-gram matches better the phone sequence, in other words, the phone sequence is more predictable. We expect a low perplexity when the n-gram model and the test phone sequence are of the same language, while high perplexity is expected otherwise. To this end, we first build a phone n-gram model [55] for each of the seven languages selected from the GlobalPhone database that have been transcribed in IPA. We then evaluate the perplexity of the phone n-gram models over some held-out test data for every language pair. Table 1 shows the perplexity measures between

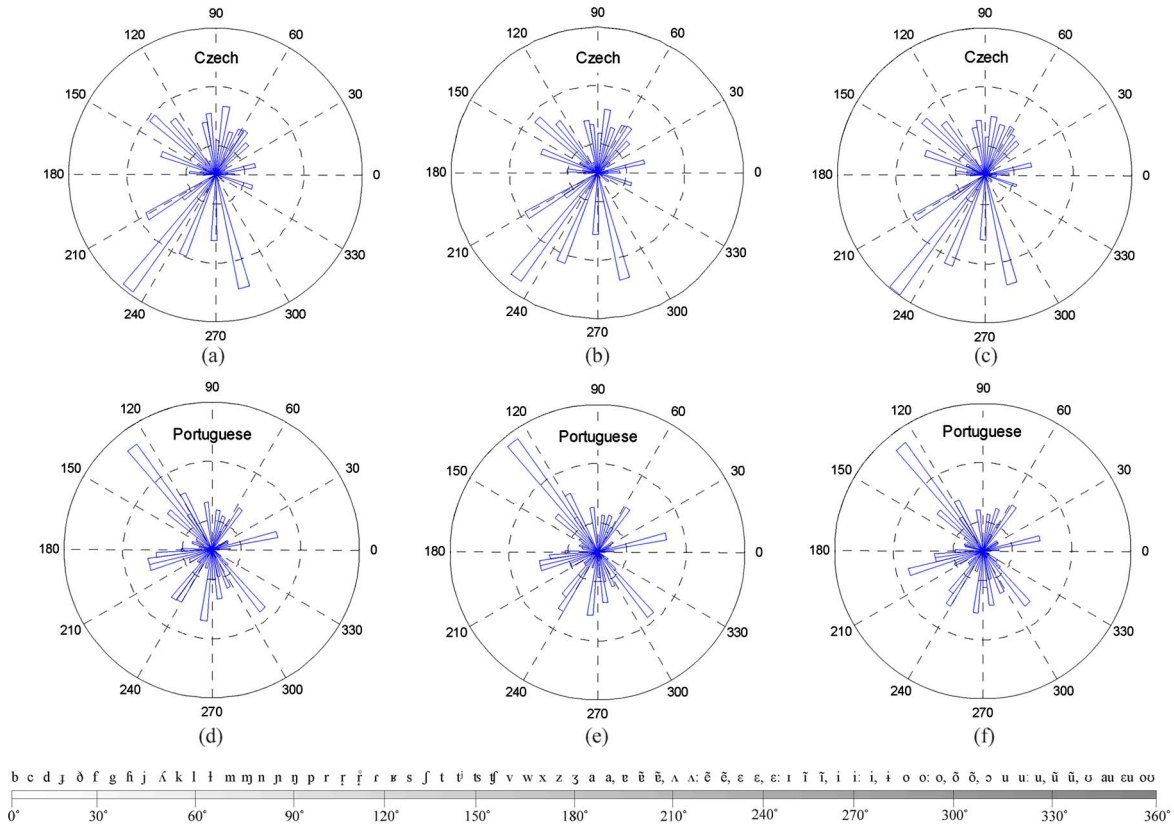

Fig. 2. ***Polar histograms showing the phone distributions of (a)–(c) Czech and (d)–(f) Portuguese utterances for three different native speakers in each language. Utterances in the same language are of different contents.***

Table 1 Perplexity Measured Between Seven Languages Selected Arbitrarily From the GlobalPhone Multilingual Database Based on Bigram and Trigram Models

	Bigram model						
Test languages	CZ	FR	GE	KO	PO	SW	TU
Czech (CZ)	16	318	356	491	383	273	555
French (FR)	456	15	115	452	155	200	335
German (GE)	586	163	15	451	447	190	370
Korean (KO)	605	582	549	16	509	548	554
Portuguese (PO)	717	321	382	490	17	450	565
Swedish (SW)	416	424	162	506	584	16	670
Turkish (TU)	985	212	214	362	429	310	13
	Trigram model						
Test languages	CZ	FR	GE	KO	PO	SW	TU
Czech (CZ)	15	210	233	253	233	212	235
French (FR)	229	12	146	248	151	200	204
German (GE)	256	185	13	254	244	187	225
Korean (KO)	258	259	260	15	258	254	259
Portuguese (PO)	254	196	230	251	15	244	248
Swedish (SW)	234	226	166	251	250	14	274
Turkish (TU)	256	176	180	230	245	232	10

languages for the cases of bigram and trigram in the upper and lower panels, respectively. The tabulated data clearly indicate that the lowest perplexity values are always observed when the phone n-gram models and the test data are from the same language. This observation confirms the differences between languages in terms of phonotactics and the effectiveness of using phone n-gram models to quantitatively measure such differences.

While the analysis of the GlobalPhone database was conducted using perfectly transcribed phone sequences, as opposed to an automatic transcription, it illustrates a quantifiable link between phonetic and phonotactic features and language recognition. For the purpose of analysis as described above, phonetic transcriptions were derived from orthographic transcriptions by means of pronunciation dictionaries, as defined in the GlobalPhone database. Furthermore, we use the add-one smoothing method [55] when training n-gram models to deal with the zero-count problem caused by unseen or out-of-set phones for a particular language (recall that the same phone set is used across different languages in the GlobalPhone database).

B. Formulation of Language Recognition

In the following, we state the problem of language recognition from an engineering perspective utilizing the linguistic knowledge sources we described in Section II-A.

Considering a speech waveform in an unknown language, we convert the audio sample into a sequence of acoustic feature vectors $\mathcal{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$, in which $\boldsymbol{o}(t)$ is extracted from the waveform at the discrete frame t and there are T such vectors. Let the set of languages under consideration $\{L_1, L_2, \ldots, L_N\}$ be equally probable. Identifying a language out of the set of N possible languages involves an assignment of the most likely language label $\hat{L}$ to the acoustic observation $\mathcal{O}$, such that

$$\hat{L} = \arg\max_{l} \; p(\mathcal{O}|L_l) \tag{1}$$

which follows a maximum-likelihood (ML) criterion. For the case where the prior probabilities of observing individual languages are not uniform, the maximum *a posteriori* (MAP) criterion could be used [64]. For mathematical tractability, the language-specific density $p(\boldsymbol{o}|L_l)$ is always assumed to follow some functional forms, for which parametric models could be used. In the simplest case, a Gaussian mixture model (GMM) [149], [150] can be used to model directly the distribution of the acoustic features, as we will see in Section IV.

To deal with phones and phonotactic knowledge sources, we assume that the speech waveform can be segmented into a sequence of phones $\widehat{\mathcal{Y}}$. Applying the ML criterion, we have now the most likely language given by

$$\hat{L} = \arg\max_{l} \; P(\widehat{\mathcal{Y}}|L_l). \tag{2}$$

In this case, $P(\widehat{\mathcal{Y}}|L_l)$ is the so-called phone n-gram model, which is a discrete probability model describing the phone occurrence and co-occurrences. One subtle difference between (1) and (2) is the stochastic model used, the former being a continuous density function such as a normal distribution, while the latter being a discrete distribution such as a multinomial distribution, depending on the nature of the features we model.

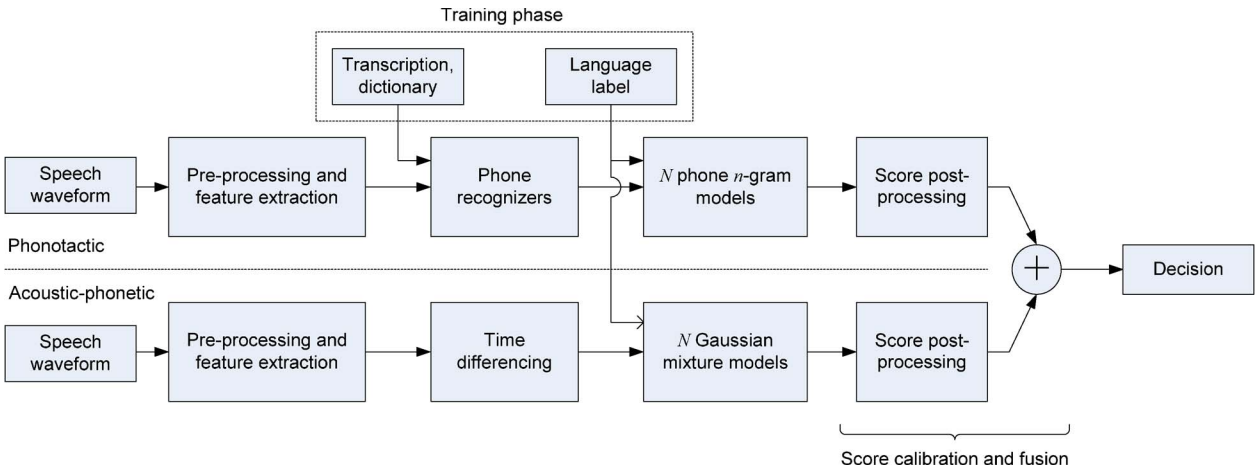

Fig. 3. ***General scheme of acoustic–phonetic and phonotactic approaches to automatic language recognition, where N indicates the number of target languages.***

In (2), we have assumed that we know the exact sequence of phones in the speech waveform $\mathcal{O}$. In practice, $\mathcal{Y}$ has to be decoded from $\mathcal{O}$ by selecting the most likely one from all possible sequences. To this end, we hinge on a set of phone models $\mathcal{M}$, such as a hidden Markov model (HMM), to decode the waveform. The most likely phone sequence could then be obtained using Viterbi decoding

$$\widehat{\mathcal{Y}} = \arg\max_{\mathcal{Y}} \; P(\mathcal{O}|\mathcal{Y}, \mathcal{M}). \tag{3}$$

Putting (2) and (3) together, and considering all possible phone sequences instead of the single best hypothesis, we obtain a more general form

$$\hat{L} = \arg\max_{l} \sum_{\forall \mathcal{Y}} P(\mathcal{O}|\mathcal{Y}, \mathcal{M}) P(\mathcal{Y}|L_l). \tag{4}$$

More details can be folded in by having $\mathcal{M}$ be language dependent or using lexical constraints in the phone decoding [40], [118]. State-of-the-art language recognizers always resort to simpler cases as in (1)–(3). In Sections III and IV, we will give a detailed mathematical formulation as to how modeling and the decision rule are implemented in practice.

C. Recent Advances

Contemporary language recognition systems operate in two phases: training and the runtime test. During the training phase, speech utterances are analyzed and models are built based on the training data given the language labels. The models are intended to represent some language-dependent characteristics seen on the training data. Depending on the information sources they model, these models could be 1) stochastic, e.g., Gaussian mixture model (GMM) [149], [150] and hidden Markov model (HMM) [98], [149]; 2) deterministic, e.g., vector quantization (VQ) [125], support vector machine (SVM) [16], [19], [31], [65], and neural network [7]; or 3) discrete stochastic, e.g., n-gram [40], [46], [66], [67]. During the test phase, a test utterance is compared to each of the language-dependent models after going through the same preprocessing and feature extraction step. The likelihood of the test utterance to each model is computed. The language associated with the most likely model is hypothesized as the language of the utterance in accordance to the ML criterion as given in (1), (2), and (4) in Section II-B.

A wide spectrum of approaches has been proposed for modeling the characteristics of languages. In this paper, we are particularly interested in the two most effective ones: the acoustic–phonetic approach and the phonotactic approach. Fig. 3 shows exemplars from these two broad categories in one diagram illustrating their common ground and differences. The advanced techniques have explored the combination of different features and approaches.

The acoustic–phonetic approach is motivated by the observation that languages differ at a very fundamental level in terms of phones and frequencies of these phones occurring (i.e., the phonetic differences between languages). More importantly, it is assumed that the phonetic characteristics could be captured by some form of spectral-based features. We then proceed to model the distribution of each language in the feature space. Notably, shifted-delta cepstrum (SDC) used in conjunction with GMM has shown to be very successful in this regard [131], where a GMM is used to approximate the acoustic–phonetic distribution of a language. It is generally believed that each

Gaussian density in a GMM captures some broad phonetic classes [110]. However, GMM is not intended to model the contextual or dynamic information of speech.

The phonotactic approach is motivated by the belief that a spoken language can be characterized by its lexical–phonological constraints. With phone transcriptions, we build N phone n-gram models for an N language task. Briefly, a phone n-gram model [55] is a stochastic model with discrete entries, each describing the probability of a subsequence of n phones (more details in Section III). Given a test utterance, each phone n-gram model produces a likelihood score. The language of the most likely hypothesis represents the classification decision. The method is referred to as the phone recognition followed by language modeling (PRLM) in the literature [150]. The languages used for the phone recognizers need not be the same, or may even be disjoint with any of those recognized. The idea is analogous to the case where a human listener uses his native language knowledge to describe unknown languages.

From a system development point of view, it is worth noting that the acoustic–phonetic approach requires only the digitized speech utterances and their language labels, while the phonotactic approach requires the phonetic transcription of speech, which could be expensive to obtain. The two broad categories as described above are by no means encompassing all possible algorithms and approaches available. Variants exist within and between these broad categories, for examples, GMM tokenization [130], parallel phone recognition [150], universal phone recognition (UPR) [73], articulatory attribute-based approach [122], [123], discriminatively trained GMM using maximum mutual information (MMI) [14], [142] or minimum classification error (MCE) [54], [105], just to name a few.

State-of-the-art language recognition systems consist of multiple subsystems in parallel, where individual scores are combined via a postprocessing back–end, as shown in Fig. 3. The motivation of score fusion is to harness the complementary behavior among subsystems provided that their errors are not completely correlated. Also appearing at the score postprocessing back–end is a calibration stage, the purpose of which is to ensure that individual scores are consistently meaningful across utterances. In particular, score calibration has been shown to be essential for a language detection task where decisions have to be made using a fixed threshold for all target languages given test segments with arbitrary in-set and out-of-set languages [85]–[87]. On the other hand, well-calibrated scores also make it easier to map unconstrained scores to a normalized range which can be viewed as confidence scores for downstream consumers. Recent works reported in [8], [9], and [12] have shown a unified framework where score calibration and fusion are performed jointly using the multiclass logistic regression. We elaborate further on score calibration and fusion in Section V.

III. PHONOTACTIC APPROACHES

The first sustained effort using the phonotactic patterns was to distinguish languages by comparing the frequency of occurrences of certain reference sounds or sound sequences with that of the target languages [69], [70]. To do this, we need to first tokenize the running speech into sound units. This can be achieved by a conventional speech recognizer that employs phonetic classes, phone classes, or phones [4], [5], [44], [45], [62], [95], [133], [143], [149] as the sound units. The first attempt with phonotactic constraints for language recognition was conducted using a Markov process to model sequences of broad phonetic classes generated by a manual phonetic transcription in eight languages [51]. The Markov process with the broad phone classes was later applied to real speech data for the language identification of five languages [77]. Next, we discuss the tokenization techniques and two common phonotactic modeling frameworks.

A. Speech Tokenization

A speech tokenizer converts an utterance into a sequence of sound tokens. As illustrated in Fig. 4, a token is defined to describe a distinct acoustic–phonetic attribute and can be of different sizes, ranging from a speech frame, a subword unit such as a phone or a syllable, to a lexical word.

The GMM tokenization technique [130] operates at the frame level. It converts a sequence of speech frames into a sequence of Gaussian labels, each of which gives rise to the highest likelihood of the speech frame. This is similar to a vector quantization process. As GMM tokenization does not attempt to relate a Gaussian label with an acoustic–phonetic event, it does not require transcribed speech data for model training. Unfortunately, the interframe

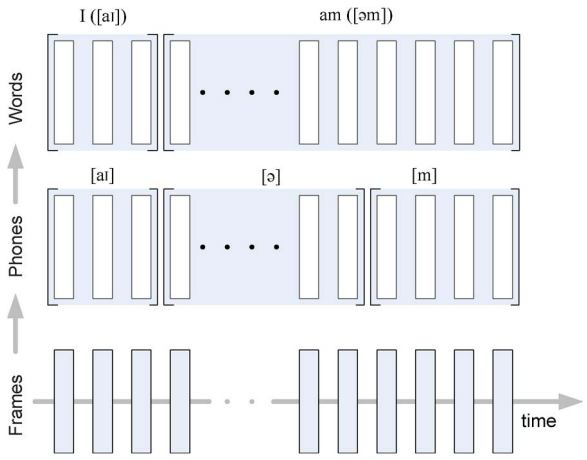

Fig. 4. *Tokenization of speech at different levels with tokens of different sizes, ranging from a speech frame, a phone, to a lexical word.*

phonotactics only describe speech dynamics at a range of tens of milliseconds, which is too short to capture the lexical–phonological information. In general, GMM tokenization is inferior to phone or subword tokenization as far as the system performance is concerned.

Another method worth mentioning is the attribute-based approach [122], [123], which tokenizes speech utterances into sequences of articulatory attributes instead of phones. An advantage of attribute-based units is that they are defined such that the same set of units is common to all languages. The training data available for different languages can, therefore, be shared to build a universal attribute recognizer (UAR). Similar objective was conceived for the UPR front–end [73] with a key difference that the UAR uses a much smaller set of speech units. For example, the work reported in [122] and [123] uses 15 attributes (five manner-of-articulation attributes, nine place-of-articulation attributes, and one silence unit to indicate the absence of articulation).

As a tradeoff between the development cost and effectiveness, the phone units are widely used in the state-of-the-art systems. A phone tokenizer is also referred to as a phone recognizer, where a phone is typically modeled by an HMM [107]. An open phone loop (or null grammar) configuration [52], whereby the transition from one phone to the others is equally probable, is often used in the phone decoding process.

We have observed in practice that a phone recognizer in a language can be used to tokenize speech of any other language. For example, a high-quality Hungarian phone recognizer based on long temporal context [90] has been widely used in the community as the phone recognizer for a variety of target languages. The role of a phone recognizer is to provide as accurate as possible phonotactic statistics for language characterization. It is, therefore, not a surprise to know that the accuracy of a phone recognizer has a great impact on the language recognition performance [73].

B. Phone n-Gram Modeling

Similar to a word n-gram that describes near-distance word occurrences, a phone n-gram captures the patterns of phone sequences using a multinomial distribution. A phone n-gram model typically employs one or multiple phone recognizers as the front–end and phone n-gram modeling for target languages as the back–end. The front–end tokenizes speech inputs based on a common phoneme inventory, while the back–end language model describes what each target language should look like in terms of phone n-gram statistics. Fig. 5 depicts the block diagram of a PRLM language recognition system with a single phone recognizer and phone n-gram models. To illustrate how it works, we take a single phone recognizer, e.g., English phone recognizer, as an example. During training, for each of the N target languages $\{L_1, L_2, \ldots, L_N\}$, the training utterances are tokenized into sequences of English phones,

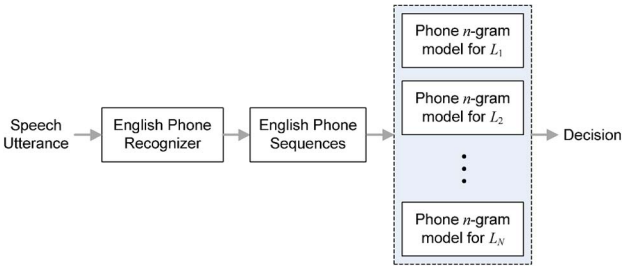

Fig. 5. *Block diagram of a PRLM language recognition system.*

which are then used to train the phone n-gram models $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. At runtime, the test utterance is tokenized by the same English phone recognizer to give a phone sequence of length J, $\widehat{\mathcal{Y}} = w_1, w_2, \ldots, w_J$. For each target language l, the log likelihood of the phone sequence can be formed as follows:

$$\log P(\widehat{\mathcal{Y}}|\lambda_l) = \sum_{j=1}^{J} \log P_{\lambda_l}\left(w_j|w_{j-1}\ldots w_{j-(n-1)}\right). \quad (5)$$

Equation (5) is also interpreted as the cross entropy between the statistics of the phone sequence $\widehat{\mathcal{Y}}$ that represents the empirical distribution of the test sample, with the phone n-gram models

$$\log P(\widehat{\mathcal{Y}}|\lambda_l) = \sum_{\hat{w}} C(\hat{w}) \log P_{\lambda_l}\left(w_j|w_{j-1}\ldots w_{j-(n-1)}\right) \quad (6)$$

where $\hat{w} = w_j w_{j-1}\ldots w_{j-(n-1)}$ and $C(\hat{w})$ is the normalized count of the n-gram $\hat{w}$ in the sequence $\widehat{\mathcal{Y}}$. The cross entropy measures how well a phone n-gram model predicts a phone sequence. By substituting (6) into (2), we can make a language recognition decision.

Working in the same principle as PRLM, parallel phone recognition followed by the phone n-gram language model or parallel phone recognition language modeling (PPR–LM) employs multiple phone recognizers, each of which is trained for a language. The multiple phone recognizers provide the statistics of the test sample from different viewpoints. As shown in Fig. 6, for each of the F phone recognizers, we train N language models for the N target languages. PPR–LM can be seen as a fusion of multiple PRLM subsystems. Given a test utterance, FN language model scores are generated through the FN phone n-gram models, $\lambda_{f,l}$ for $f = 1, 2, \ldots, F$ and $l = 1, 2, \ldots, N$. With the scores, one can devise a strategy for language recognition decision making. In the case where we need a single output score, there are many ways to summarize the FN scores into one. One possible way is to fuse the

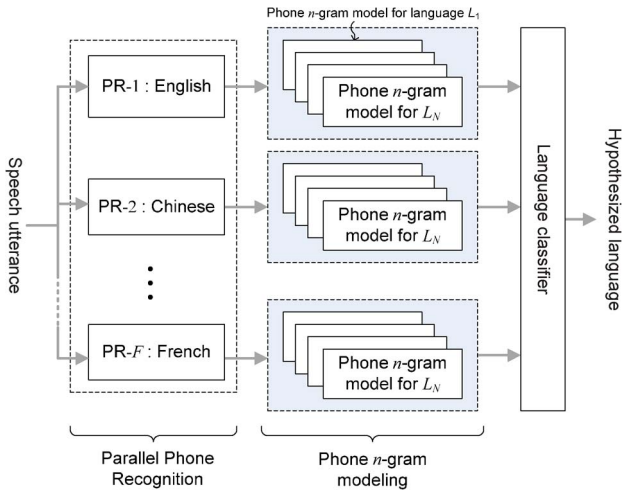

Fig. 6. *Block diagram of a PPR–LM language recognition system.*

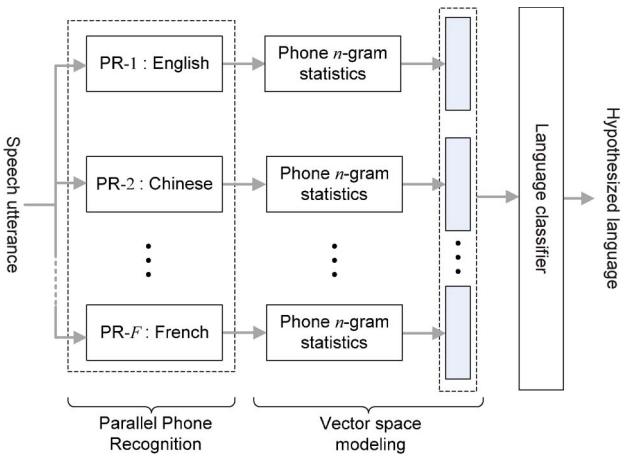

Fig. 7. *Block diagram of a PPR–VSM language recognition system.*

posterior probabilities from F parallel subsystems, as follows:

$$\log P(L_l|\mathcal{O}) = \sum_{f=1}^{F} \log \frac{P(\widehat{\mathcal{Y}}_f|\lambda_{f,l})}{\sum_{i=1}^{N} P(\widehat{\mathcal{Y}}_f|\lambda_{f,i})} \tag{7}$$

where $\widehat{\mathcal{Y}}_f$ is the phone sequence generated from speech utterance $\mathcal{O}$ by the fth phone recognizer. $P(\widehat{\mathcal{Y}}_f|\lambda_{f,l})$ is the likelihood score for the lth target language.

C. Vector Space Modeling

The *bag-of-sounds* framework [72] marks another successful attempt in language recognition. The idea is to represent the empirical distribution of the test sample in a high-dimensional vector. The *bag-of-sounds* concept is analogous to the *bag-of-words* paradigm originally formulated in information retrieval and text categorization [113]. The *bag-of-words* paradigm represents a text document as a vector of word counts. It is believed that it is not just the words, but also the co-occurrence of words that distinguish semantic domains of text documents. One can easily draw the analogy between a sound token in the *bag-of-sounds* and a word in the *bag-of-words*. The difference is that we deal with a phone sequence instead of a word sequence.

In the *bag-of-sounds* framework, we arrange the phone n-gram statistics of both training and test samples into high-dimensional vectors. Fig. 7 depicts the block diagram of a PPR–VSM architecture,[1] in which PPR is followed by vector space modeling (VSM) [73], [75]. Suppose that we have F phone recognizers with a phone inventory of $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_f, \ldots, \mathcal{V}_F\}$ and the number of phonemes in $\mathcal{V}_f$ is n_f. A speech utterance is decoded by these phone recognizers into F phone sequences. Each of these phone sequences can be expressed by a high-dimensional phonotactic feature vector with the phone n-gram statistics. The dimension of the feature vector is equal to the total number of phone n-gram patterns needed to highlight the overall behavior of the utterance. If unigram and bigram are the only concerns, we will have a vector of $n_f + n_f^2$ phonotactic elements, denoted as $\mathbf{v}_f$, to represent the utterance by the fth phone recognizer. As shown in Fig. 7, all the F phonotactic feature vectors are concatenated into a composite *bag-of-sounds* vector $\mathbf{v} = [\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_f^{\mathrm{T}}, \ldots, \mathbf{v}_F^{\mathrm{T}}]^{\mathrm{T}}$, with a dimension of $B = \sum_f (n_f + n_f^2)$, if only unigram and bigram features are included.

After a spoken utterance is vectorized in this way, language recognition can be cast as a vector-based classification problem. The simplest way is to measure the similarity between two composite vectors, one derived from the test utterance and another derived from all the training data of a target language. The similarity between two vectors in VSM can be approximated by the inner product or cosine distance [24], [114]. If we take a close look at (6), we can find that the cross entropy can be seen as an inner product between the normalized count vector of the test utterance and the vectorized log-probability from the target phone n-gram model.

Vector space modeling techniques benefit from the recent progress in SVM, which offers a low-cost solution to the classification of high-dimensional vectors. The SVM is optimized based on a structural risk minimization principle [138]. Due to its distribution-free property, it has the advantage of providing an excellent generalization capability.

Suppose that we have two *bag-of-sounds* vectors of B dimensions $\mathbf{x} = [x_1, x_2, \ldots, x_B]^{\mathrm{T}}$ and $\mathbf{y} = [y_1, y_2, \ldots, y_B]^{\mathrm{T}}$, extracted from two speech utterances. Each of the vectors

[1] A similar architecture was also studied in speaker recognition [15].

represents a discrete empirical distribution of phonotactic counts. An L^2 inner product kernel [18] is given by

$$\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{B} x_i \cdot y_i \tag{8}$$

which measures the similarity between two *bag-of-sounds* vectors.

For each target language, an SVM is trained using the *bag-of-sounds* vectors pertaining to the target language as the positive set, with those from all other languages as the negative set. In this way, N one-versus-the-rest SVM models are built, one for each target language. Given a test utterance, N SVM output scores will be generated for language recognition. Alternatively, the N SVM scores can be used to produce an N-dimensional score vector [82], and thus we are able to project the high-dimension composite feature vectors into a much lower dimension of N. The generated N-dimensional score vectors can be further modeled by a Gaussian back–end for classification decision [151].

D. Phonotactic Front–End

By carefully examining Figs. 6 and 7, we realize that phone n-gram modeling and vector space modeling are using the same phone recognition front–end, such as a PPR front–end, to derive the n-gram statistics. The difference lies in the way of phone n-gram representation. Therefore, any improvement in phone recognition front–end will benefit both modeling techniques.

The study of *bag-of-sounds* has motivated a series of works to further the design of the PPR–VSM front–end. There have been attempts to explore high-order n-grams and to select discriminative phonotactic features [72], [73], [112], [132], to redesign phone recognizers that are knowledgeable about the target languages [128], and to construct diversified acoustic models for phone recognizers to provide different acoustic perspectives [120]. While the phone n-grams are typically estimated over the one-best phone transcriptions, it was discovered that the expected n-gram counts derived from a phone lattice outperform the n-gram counts from one-best phone transcriptions. The improvement is attributed to the richer information available in the lattice than in the one-best results [40], [118].

In general, it is believed that more parallel phone recognizers, higher order phone n-grams, and more accurate phone recognizers provide more informative phonotactic features and thus lead to better performing systems.

IV. ACOUSTIC–PHONETIC APPROACHES

As discussed in Section II-A, we believe that each language has its unique phonetic repertoire. In the acoustic–phonetic approach, we attempt to model the acoustic–phonetic distribution of a language using the acoustic features.

The early efforts of acoustic–phonetic approaches to language recognition probably began in the 1980s. A polynomial classifier based on 100 features derived from linear predictive coding (LPC) analysis [27] was studied for recognition of eight languages. A VQ technique with pitch contour and formant frequency features was proposed for recognition of three languages [38], [42]. A study using multiple language-dependent VQ codebooks and a universal VQ codebook (i.e., general and language independent) was conducted over the LPC derived features for recognition of 20 languages [125]. This study suggests that acoustic features are effective in different settings. A comparative study among four modeling techniques, VQ, discrete HMM, continuous-density HMM (CDHMM), and GMM with mel-cepstrum coefficients, was carried out over four languages, showing that CDHMM and GMM offer a better performance [98].

In this section, we will discuss feature extraction, two modeling techniques, and intersession variability compensation that all affect the system performance.

A. Acoustic Feature Extraction

Mel-frequency cepstral coefficients (MFCCs) [32] are effective in most speech recognition tasks because they exploit auditory principles, whereby the mel-scale filter bank is a rough approximation to human auditory system's response [106]. It is not surprising that they work well in language recognition as well. To overcome undesired variation across sessions, compensation techniques such as mean-variance normalization (MVN) [52], RASTA [47] and vocal tract length normalization (VTLN) [68] are typically applied after the voice activity detection (VAD) process, by which silence is removed.

Typically, MFCC features are computed at each short speech segment (e.g., 10 ms) together with their first- and second-order derivatives to capture the short-term speech dynamics. We would like to mention in particular the shifted-delta-cepstral (SDC) coefficients [131], which are useful for language recognition because they capture the speech dynamics over a wider range of speech frames than the first- and second-order MFCC derivatives. The computation of the SDC features is illustrated in Fig. 8. The SDC features are specified by four parameters $\{Z, d, P, k\}$, where Z is the number of cepstral coefficients computed at each frame, d represents the time advance and delay for the delta computation, k is the number of delta-cepstral blocks whose delta-cepstral coefficients are stacked to form the final feature vector, and P is the time shift between consecutive blocks. For example, the Z static features are computed as

$$\mathbf{c}(t) = [c_0(t), c_1(t), \ldots, c_{Z-1}(t)]^{\mathrm{T}} \tag{9}$$

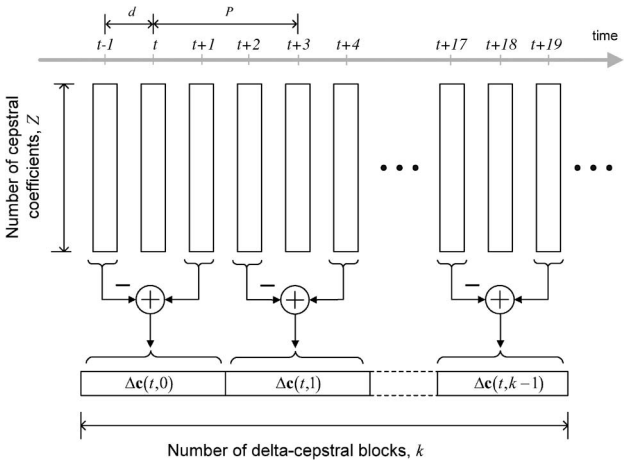

Fig. 8. ***SDC feature extraction at frame time t, illustrated for the case of*** $Z-d-P-k=7-1-3-7$.

and the ith block of delta-cepstral features is computed as

$$\Delta\mathbf{c}(t,i)=\mathbf{c}(t+iP+d)-\mathbf{c}(t+iP-d). \tag{10}$$

Finally, each SDC feature vector at time t contains Z parameters of static cepstral features and kZ delta features. A commonly adopted SDC configuration is $Z-d-P-k=7-1-3-7$ leading to $kZ=49$ delta features in addition to seven static features. Such SDC features cover temporal information over $kP=21$ consecutive frames of cepstral features.

B. Statistical Modeling

Language recognition and speaker recognition have many similarities in terms of technical formulation, methodologies, and evaluation measurement. The statistical modeling technique, with the universal-background-model-based GMM (GMM–UBM), brings the same success to language recognition that it has brought to speaker recognition [111]. One of the attractive attributes of the GMM is its ability to closely approximate any arbitrarily shaped data distributions, and its ability to model the underlying data classes by the individual Gaussian components. In language recognition, we consider the set of spectral frames from the utterances as a collection of independent samples. A GMM is used to approximate the overall acoustic–phonetic distributions of a spoken language. The GMM modeling technique is popular in language recognition due to not only its ability to model a large class of sample distributions, but also its competitive performance in practice.

For a D-dimensional feature vector $\boldsymbol{o}$ with M Gaussian mixture density functions, the likelihood function of a GMM is a weighted linear combination

$$p(\boldsymbol{o}|\lambda)=\sum_{i=1}^{M}\omega_i\mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i). \tag{11}$$

Each of the Gaussian density functions is parameterized by a D-dimensional mean vector $\boldsymbol{\mu}_i$ and a $D\times D$ covariance matrix $\boldsymbol{\Sigma}_i$. A diagonal covariance matrix is normally adopted for the sake of computation simplicity and practical consideration especially when only a limited amount of training data is available. The Gaussian density function is given as follows:

$$\mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_i)\right] \tag{12}$$

with the mixture weights summing to one, $\sum_{i=1}^{M}\omega_i=1$. We denote a GMM as $\lambda=\{\omega_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,i=1,\ldots,M\}$.

Given a collection of training feature vectors, a GMM is generally trained with the expectation–maximization (EM) algorithm [36], where the model parameters are estimated with the ML criterion. In language identification, we train a GMM for each language for the recognition task as formulated in (1).

In the GMM–UBM paradigm, where UBM is a background model that represents the world's spoken language, we usually start by training a UBM with data from all languages and adapt a GMM model for each language from the UBM using the MAP technique [39]. In practice, we often suffer from insufficient training data to build a GMM from scratch. The GMM–UBM training process offers a solution to overcome such a problem. With a simple formulation yet competitive performance, the GMM–UBM technique has become a reference system in language recognition, which we will further discuss in Section V-C.

As the GMM model operates by capturing the underlying acoustic classes as reflected in the spectral feature distributions for each language, it is vulnerable to undesired variability due to nonlanguage effects, such as speaker and channel. Several training techniques have been attempted to address session variations and to improve discriminative ability, for example, constrained maximum-likelihood linear regression (CMLLR) [119], soft margin estimation [148], and MMI training [14].

C. Vector Space Modeling

Similar to vector space modeling in the phonotactic approaches, there have been effective ways to characterize the spectral features of an utterance as a high-dimensional vector under the SVM paradigm.

The SVM is typically used to separate vectors in a binary classification problem. It projects an input vector $\mathbf{x}$ into a scalar value $f(\mathbf{x})$, as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{I} \alpha_i \kappa(\mathbf{x}, \mathbf{x}_i) + \beta \tag{13}$$

where vectors $\mathbf{x}_i$ are support vectors, I is the number of support vectors, α_i are the weights, and β is a bias. The weights imposed on the support vectors are constrained such that $\sum_{i=1}^{I} \alpha_i = 0$ and $\alpha_i \neq 0$. Function $\kappa(\cdot)$ is the kernel, subject to certain properties (the Mercer condition), so that it can be expressed as

$$\kappa(\mathbf{x}, \mathbf{y}) = \phi^{\mathrm{T}}(\mathbf{x})\phi(\mathbf{y}) \tag{14}$$

where $\phi(\mathbf{x})$ is a mapping from the input space to a possibly infinite-dimensional space. Now, the question is: How do we represent a speech utterance, thus a language, in high-dimensional vector space? We will discuss two different vector space modeling techniques that characterize a speech utterance with spectral features and GMM parameters, respectively.

After acoustic feature extraction, a speech utterance has become a sequence of spectral feature vectors. Comparing two speech utterances with the SVM, there needs to be a way of taking two such sequences of feature vectors, calculating a sequence kernel operation, and computing the SVM output. A successful approach using the sequence kernel is the generalized linear discriminative sequence (GLDS) [16]. It takes the explicit polynomial expansion of the input feature vectors and applies the sequence kernel based on generalized linear discriminants. The polynomial expansion includes all the monomials of the features in a feature vector.

As an example, for two features in the input feature vector $\boldsymbol{o} = [o_1, o_2]^{\mathrm{T}}$, the monomials with degree two are $\mathbf{b}(\boldsymbol{o}) = [1, o_1, o_2, o_1^2, o_1 o_2, o_2^2]^{\mathrm{T}}$. Let $\mathcal{O}_{\mathrm{X}}$ and $\mathcal{O}_{\mathrm{Y}}$ be two input sequences of feature vectors. The polynomial discriminant can be obtained using the mean-squared error criterion [16]. The resulting sequence kernel can be expressed as follows:

$$\kappa(\mathcal{O}_{\mathrm{X}}, \mathcal{O}_{\mathrm{Y}}) = \tilde{\mathbf{b}}_{\mathrm{X}}^{\mathrm{T}} \mathbf{R}^{-1} \tilde{\mathbf{b}}_{\mathrm{Y}} \tag{15}$$

where the vector

$$\tilde{\mathbf{b}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{b}\{\boldsymbol{o}(t)\} \tag{16}$$

is the average expansion over all the feature vectors of the input utterance $\mathcal{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$. The matrix $\mathbf{R}$ is the correlation matrix obtained from a background data set. An approximation of $\mathbf{R}$ can be applied to calculate only the diagonal terms [16].

While we consider a GLDS kernel as a nonparametric approach to the language recognition problem, the GMM supervector offers a parametric alternative. Given a UBM and a speech utterance drawn from a language, we derive a GMM supervector $\mathbf{m} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_M]^{\mathrm{T}}$ by stacking the mean vectors of all the adapted mixture components [18]. In this way, a speech utterance is mapped to a high-dimensional space using the mean parameters of a GMM. It should be noted that the MAP adaptation is performed for each and every utterance available for a particular language. For two speech utterances $\mathcal{O}_{\mathrm{X}}$ and $\mathcal{O}_{\mathrm{Y}}$, two GMMs can be derived using MAP adaptation from the UBM to obtain two supervectors $\mathbf{m}^{\mathrm{X}} = [\boldsymbol{\mu}_1^{\mathrm{X}}, \boldsymbol{\mu}_2^{\mathrm{X}}, \ldots, \boldsymbol{\mu}_M^{\mathrm{X}}]^{\mathrm{T}}$ and $\mathbf{m}^{\mathrm{Y}} = [\boldsymbol{\mu}_1^{\mathrm{Y}}, \boldsymbol{\mu}_2^{\mathrm{Y}}, \ldots, \boldsymbol{\mu}_M^{\mathrm{Y}}]^{\mathrm{T}}$. There are a number of ways to compare two speech utterances in terms of the derived mean vectors. A natural choice is the Kullback–Leibler (KL) divergence, which can be approximated by the following upper bound [18]:

$$d(\mathbf{m}^{\mathrm{X}}, \mathbf{m}^{\mathrm{Y}}) = \frac{1}{2} \sum_{i=1}^{M} w_i \left(\boldsymbol{\mu}_i^{\mathrm{X}} - \boldsymbol{\mu}_i^{\mathrm{Y}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \left(\boldsymbol{\mu}_i^{\mathrm{X}} - \boldsymbol{\mu}_i^{\mathrm{Y}}\right). \tag{17}$$

We can then find the corresponding inner product that serves as a kernel function, i.e., the so-called KL kernel [18], as follows:

$$\kappa_{\mathrm{KL}}(\mathcal{O}_{\mathrm{X}}, \mathcal{O}_{\mathrm{Y}}) = \sum_{i=1}^{M} \left(\sqrt{w_i}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_i^{\mathrm{X}}\right)^{\mathrm{T}} \left(\sqrt{w_i}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_i^{\mathrm{Y}}\right). \tag{18}$$

Note that the KL kernel only accommodates adaptation of GMM to mean vectors while leaving the covariance matrices unchanged. A Bhattacharyya kernel [145] was proposed that allows for adaptation of covariance matrices, showing an improved performance. Similar to the KL kernel, the Bhattacharyya kernel represents speech utterances as the adapted mean vectors of GMMs

$$\kappa_{\mathrm{BHATT}}(\mathcal{O}_{\mathrm{X}}, \mathcal{O}_{\mathrm{Y}}) = \sum_{i=1}^{M} \left[\left(\frac{\boldsymbol{\Sigma}_i^{\mathrm{X}} + \tilde{\boldsymbol{\Sigma}}_i}{2}\right)^{-1/2} \left(\boldsymbol{\mu}_i^{\mathrm{X}} - \tilde{\boldsymbol{\mu}}_i\right)\right]^{\mathrm{T}} \times \left[\left(\frac{\boldsymbol{\Sigma}_i^{\mathrm{Y}} + \tilde{\boldsymbol{\Sigma}}_i}{2}\right)^{-1/2} \left(\boldsymbol{\mu}_i^{\mathrm{Y}} - \tilde{\boldsymbol{\mu}}_i\right)\right]. \tag{19}$$

The difference lies at the adapted covariance matrices $\mathbf{\Sigma}_i^{\mathrm{X}}$ and $\mathbf{\Sigma}_i^{\mathrm{Y}}$, which appear as the normalization factors to the adapted mean vectors [145]. Here, $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\mathbf{\Sigma}}_i$ are, respectively, the mean vectors and covariance matrices of the UBM. In [19], the KL kernel was extended to include covariance matrices into the supervector. An interesting interaction between the GMM supervector and the SVM is that SVM parameters can be pushed back to GMM models that allow for a more effective scoring.

Finally, for each target language, a one-versus-the-rest SVM can be trained in the vector space with the target language being the positive set and all other competing languages being the negative set. A decision strategy can be devised to summarize the outputs of multiple one-versus-the-rest SVMs during the runtime test, as suggested in Section III-C.

D. Intersession Variability

In language recognition, we face several sources of nonlanguage variability, such as speaker, gender, channel, and environment, in the speech signals. For simplicity, we refer to all such variabilities as intersession variability, that is, the variability exhibited by a given language from one recording session to another. The class of subspace methods for model compensation, such as joint factor analysis (JFA) [21], [57], [140] and nuisance attribute projection (NAP) [124], which originated from speaker recognition research, is useful to address the variability issues. The idea of the subspace techniques is to model and eliminate the nuisance subspace pertaining to intersession variability, thereby reduce the mismatch between training and test. Recent efforts have brought the subspace methods from model domain to feature domain, such as the feature-level latent factor analysis (fLFA) [56], [139] and feature-level NAP (fNAP) [20] that have shown superior performance in language recognition. Feature domain compensation is not tied to a specific model assumption, thus, it could be used for a wider range of modeling schemes. We explain the working principle of fLFA as follows.

We assume that the GMM supervector $\mathbf{m}$, derived for a speech utterance, can be decomposed into the sum of two supervectors

$$\mathbf{m} = \tilde{\mathbf{m}} + \mathbf{U}\mathbf{z} \tag{20}$$

where $\tilde{\mathbf{m}}$ is the UBM mean supervector, $\mathbf{U}$ is a low-rank matrix projecting the latent factor subspace into the supervector model domain, and $\mathbf{z}$ is a low-dimensional vector holding the latent factors for the current speech utterance and language [139]. Matrix $\mathbf{U}$ can be estimated using an EM training based on the principal component analysis [126], and the latent factors $\mathbf{z}$ are estimated for each session using the probabilistic subspace adaptation method based on MAP estimation [80]. In order to conduct feature-domain compensation, (20) can be rewritten as

$$\boldsymbol{\mu}_i = \tilde{\boldsymbol{\mu}}_i + \mathbf{U}_i\,\mathbf{z} \tag{21}$$

where $\boldsymbol{\mu}_i$ and $\tilde{\boldsymbol{\mu}}_i$ are the mean vectors of the ith Gaussian component of the GMM and the UBM, respectively. Submatrix $\mathbf{U}_i$ is the intersession compensation offset related to the ith Gaussian component. The compensation of feature vector $\boldsymbol{o}(t)$ is obtained by subtracting a weighted sum of the intersession compensation bias vector

$$\tilde{\boldsymbol{o}}(\boldsymbol{t}) = \boldsymbol{o}(\boldsymbol{t}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} \gamma_{\boldsymbol{i}}\mathbf{U}_{\boldsymbol{i}}\mathbf{z} \tag{22}$$

where γ_i is the Gaussian occupation probability.

Most recently, another factor analysis technique, i-vector [34], has become very popular in speaker recognition and has been also introduced to language recognition [35]. The i-vector paradigm provides an efficient way to compress GMM supervectors by confining all sorts of variabilities (both language and nonlanguage) to a low-dimensional subspace, referred to as the total variability space. The generative equation is given by

$$\mathbf{m} = \tilde{\mathbf{m}} + \mathbf{T}\mathbf{w} \tag{23}$$

where matrix $\mathbf{T}$ is the so-called total variability matrix. The latent variable $\mathbf{w}$ is taken to be a low-dimensional random vector with a standard normal distribution. For a speech utterance $\mathcal{O}$, its i-vector is given by the posterior mean of the latent variable $\mathbf{w}$, i.e., $E\{\mathbf{w}|\mathcal{O}\}$. Since $\mathbf{T}$ is always a low-rank rectangular matrix, the dimensionality of the i-vector is much smaller compared to that of the supervector $\mathbf{m}$.

One subtle difference between the total variability space $\mathbf{T}$ and the session variability space $\mathbf{U}$, as given in (20), is that the former captures both the intersession variability as well the intrinsic variability between languages, which is not represented in the latter. As such, matrix $\mathbf{U}$ is used to remove the unwanted intersession variability, whereas the purpose of matrix $\mathbf{T}$ is to preserve the important variability in a much more compact form than the original supervector. Compensation techniques are then applied on the i-vectors to cope with the intersession variability. To this end, the linear discriminant analysis has been shown to be effective [35]. It is worth noting that, though labeled data are not required for training the total variability subspace pertaining to the i-vectors, the subsequent compensation techniques can

only be effective when training data for the intended intersession variability are available.

V. TOPICS IN SYSTEM DEVELOPMENTS

Most state-of-the-art language recognition systems consist of multiple subsystems, which include various forms of acoustic or phonotactic approaches, as described earlier. Each subsystem provides an expert view about the input speech. To make a balanced decision that reflects a mixture of expert views, we rely on an effective fusion technique. By combining the individual outputs, we aim at a higher accuracy than that of the best subsystem [59]. The ways in which subsystems can be combined are numerous [49], [53], [59]. While fusion can take place at the feature, model, or score levels, studies have shown that fusion at the score level is the most effective in delivering increased accuracy [104]. In this section, we show the latest advances in decision fusion whereby output scores from component subsystems are combined at the score level through the use of Gaussian back–end and multiclass logistic regression. The combined scores for the overall system could then be used for the purpose of language identification and verification.

A. The Magic of Gaussian Back–End

The notion of a Gaussian back–end was first introduced and adopted in [151] for the fusion of likelihood scores from phone *n*-grams. It was then extended to a general fusion device taking scores from recognizers of any kind, be it phonotactics or acoustics based. The importance of the Gaussian back–end is twofold: fusion and calibration.

We consider the output scores from subsystems as elements of a score vector. If there are K subsystems and N target languages, then there are KN elements in the score vector. That is, the score vectors from the component subsystem, each producing N number of scores $s(k,l)$, are stacked to form the score vector $\mathbf{s}=[\mathbf{s}_1^{\mathrm{T}},\mathbf{s}_2^{\mathrm{T}},\ldots,\mathbf{s}_K^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{s}_k=[s(k,1),s(k,2),\ldots,s(k,N)]^{\mathrm{T}}$. During the training phase, the collection of score vectors associated with a given target language are used to train a multivariate normal distribution, one for each of the N target languages. With the ML criterion, the mean vectors of the Gaussian models are given by the sample means of the score vectors based on the language labels. The covariance matrices are estimated in a similar manner, subject to constraints, like, tied, diagonal, full, or structured covariance, by which model complexity could be controlled.

Let $\{\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l,l=1,\ldots,N\}$ be the set of Gaussian back–end parameters consisting of class-dependent mean vectors and covariance matrices. The back–ended scores, s_l' for $l=1,\ldots,N$, are obtained by evaluating the log Gaussian likelihood $\log\mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)$ for each target language $L_l\in\{L_1,L_2,\ldots,L_N\}$. For the case where all Gaussians share a common covariance matrix, we form what is called the *linear back–end*, as follows:

$$s_l'=\log\mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_l,\boldsymbol{\Sigma})-\beta=\left(\boldsymbol{\mu}_l^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\mathbf{s}-\frac{1}{2}\boldsymbol{\mu}_l^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l \quad (24)$$

where terms common to all classes are consolidated as $\beta=(\mathbf{s}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{s})/2+C$ and discarded. Here, C accounts for the normalization factor in the Gaussian function. Clearly, the simplification leads to scaling of the Gaussian likelihoods consistently for all l with a common scaling factor of $\exp(-\beta)$, which essentially disappears when the likelihood ratio or posterior probabilities are computed.

The back–end function in (24) is made linear with respect to the input raw scores $\mathbf{s}$ by tying the Gaussian models with a common covariance matrix. The covariance matrix is trained by pooling data across all languages after removing the mean. Besides parameter tying, we could also resort to diagonal or full covariance leading to a linear back–end with different complexity. To compromise between these two, we may consider using the factor analysis or the probabilistic principle component analysis for modeling the covariance. In [17] and [132], the linear discriminant analysis was used on the score vectors coming out from each core recognizer. The resulting score vector is, therefore, decorrelated justifying the use of a diagonal covariance matrix.

Another form of the back–end is obtained by imposing a heteroscedastic assumption on the score space, as opposed to the homoscedastic assumption where covariance matrices are tied. As the covariance matrices become different for each target language, an additional quadratic term $\mathbf{s}^{\mathrm{T}}\boldsymbol{\Sigma}_l^{-1}\mathbf{s}$ appears in the back–end function. A quadratic back–end is particularly useful for modeling out-of-set languages, which generally exhibit larger variation in the score space [11].

The essential element of back–end processing is to recognize languages in the score space by using the output from one or more subsystems (e.g., PPR–LM, SVM, GMM) as features. These subsystems could, therefore, be viewed as score generators [144]. The scores produced by the subsystems can be in different forms, for instance, log-likelihood, class posterior, or any deterministic distance measure (e.g., Mahalanobis distance). Using a generative score-vector model for each class in the form of a multivariate normal distribution per class, the back–ended scores are calibrated into well-behaved log-likelihood scores from which application-dependent decisions can be made systematically [8], [134]. At the same time, the back–end effectively performs fusion of the recognizers.

Another interesting byproduct of using the Gaussian back–end is that we can now use it for new languages for which we have only a limited amount of development data. With K subsystems for N target languages, we can train a

Gaussian model using score vectors for the $(N+1)$th language, a new target language, as long as we are given some developmental data from this language. This is also particularly useful to model *out-of-set* languages. To this end, the score vectors associated with nontarget languages are pooled and used to train a Gaussian model. Out-of-set languages are then treated as an additional class representing the *none-of-the-above* hypothesis.

B. Multiclass Logistic Regression

Another way of combining likelihood scores from multiple subsystems is the product rule [104], [150], or equivalently, the sum of log likelihoods as follows:

$$s'_l = \sum_{k=1}^{K} \log S(\mathcal{O}|\lambda_{k,l}) = \sum_{k=1}^{K} s(k,l) \tag{25}$$

for $l = 1, \ldots, N$. The underlying assumption in (25) is that the output scores $S(\mathcal{O}|\lambda_{k,l})$ from the models $\lambda_{k,l}$ can be interpreted as likelihood measures and the subsystems are independent of each other. No training is required where the log-likelihoods are essentially summed with equal weights. A more general form could be obtained by introducing additional control parameters as follows:

$$s'_l = \sum_{k=1}^{K} \alpha_k s(k,l) + \beta_l \tag{26}$$

for $l = 1, \ldots, N$. Note that the weights α_k assigned to subsystems are used across all target languages, while biases β_l are made dependent on the targets. The fusion function is linear with respect to the log-likelihood scores. The weights and biases could be found by an exhaustive grid search optimizing some application-dependent cost functions [74]. A more systematic way is to use a multiclass logistic regression model [43] for the class posterior

$$P\left(L_l|s'_l\right) = \frac{\exp\left(s'_l\right)}{\sum_{i=1}^{N} \exp\left(s'_i\right)} \tag{27}$$

and to find the parameters that maximize the log-posterior probability on the development data, as follows [8], [10], [12], [134]:

$$Q(\alpha_1, \ldots, \alpha_K, \beta_1, \ldots, \beta_N) = \sum_{m=1}^{M} \sum_{l=1}^{N} \gamma_{ml} \log P\left(L_l|s'_l\right). \tag{28}$$

In the above equation, M is the number of training samples and γ_{ml} is the training label, i.e.,

$$\gamma_{ml} = \begin{cases} w_l, & \text{if labeled as language } L_l \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

Here, w_l for $l = 1, 2, \ldots, N$ are weighting factors used to normalize the class proportion in the training data. If there are an equal number of training samples for each target language (or normalization is not intended), then we set $w_l = 1$. With $w_l = M/(M_l N)$, we normalize out the effect of class population M_l from the cost function, where N is the number of target languages. No closed-form solution exists to maximize (28). A conjugate gradient–descent optimization technique was found to be the most efficient among several other gradient-based numerical optimization techniques [10], [93].

Notice that calibration is achieved jointly with score fusion via (26)–(29), where subsystem scores are combined in the first term on the right-hand side of (26), while score calibration is achieved by scaling and shifting of the input scores. When $K = 1$, (26) reduces to a score calibration device. Such a form of calibration is referred to as the *application-independent calibration* in [8], [9], and [135], the purpose of which is to allow a threshold to be set on the calibrated scores based on the cost and prior (i.e., the application parameters). Setting of threshold is crucial especially for language detection and open-set language identification tasks. The rationale behind this, as given in [12], is as follows. To make a decision with uncalibrated scores, one needs to probe the subsystem to understand the behavior of its output scores and, thus, to learn where to put the threshold. In contrast, given well-calibrated scores, one can calculate the risk expectation in a straightforward way and, thus, make minimum risk decisions with no previous experience of the subsystem.

C. Identification and Verification

In a similar way as humans perceive the problem, it is intuitive to see automatic language recognition as an *identification* task. Given a spoken utterance, we match it against a predefined set of target languages and make a decision regarding the identity of the language spoken in the segment of speech. Another closely related recognition problem is language *verification* or *detection*. Here, we are given a speech segment and the task is to decide between two hypotheses—whether the speech segment is from a target (or claimed) language.

Language recognition could be best manifested as a multiclass recognition problem where the input is known to belong to one from a set of discrete classes. The objective is to recognize the class of the input. In what follows, we adopt the formulation proposed in [8] and [10] to further illustrate the problem.

We are given a list of N target classes, $\{L_1, L_2, \ldots, L_N\}$, for each of which a prior probability is given (a flat prior could be assumed in a general application-independent setting). In the closed-set scenario, $\{L_1, L_2, \ldots, L_N\}$ represents N different explicitly specified languages. In the open-set case, $L_1, L_2, \ldots, L_{N-1}$ are explicitly specified languages, and L_N denotes any of the yet unseen out-of-set languages. Such an open-set scenario could be handled by introducing an artificial "none-of-the-above" class into the target set, for instance, at the Gaussian back–end. Given a spoken utterance $\mathcal{O}$ and the set of target languages, we have the following different, but closely related, tasks [8], [10].

- *Language identification*: Which of the N languages does $\mathcal{O}$ belong to?
- *Language verification*: Does $\mathcal{O}$ belong to language L_l or to other languages (i.e., one of the other $N-1$ languages)?

For the identification task, we compute the likelihood given each language and select the language hypothesis that yields the highest likelihood. The language verification or detection is a binary classification problem, where a decision has to be made between two hypotheses with respect to a decision threshold. In Section VI, we look into more details regarding threshold setting and performance assessment following the language detection protocol as established in NIST LREs.

VI. THE NIST LRE PARADIGM

A. Corpora for LREs

The availability of sufficiently large corpora has been the major driving factor in the development of speech technology in recent decades [25], [26], [63], [76], [87], [94]. To model a spoken language, we need a set of speech data for the language. To account for the within-language variability, which we also call intersession variability, such as speaker, content, recording device, communication channel, and background noise, it is desirable to have sufficient data that include the intended intersession effects.

The first large-scale speech data collection for the purpose of language recognition research was carried out by OGI in the early 1990s. Their efforts resulted in the OGI-11L and OGI-22L corpora [63], [94], which are now distributed through the Linguistic Data Consortium (LDC). As its name implies, the OGI-11L is a multilanguage corpus of 11 languages. The number of languages was expanded to 22 in OGI-22L. Similar data collections were organized by LDC leading to the CallHome and CallFriend corpora, consisting of 6 and 12 languages, respectively. The aforementioned corpora were collected over the telephone network, which reaches out to speakers of different languages easily over a wide geographical area [94]. Both OGI-11L and OGI-22L are monologue speech corpora, where the callers answer to the questions prompted by a machine. The later collection of CallHome, CallFriend, and Mixer [25] corpora was motivated by a more challenging task aiming at conversational speech using dialog between individuals.

Another drive that has contributed to advances in speech research is the effort toward standard protocols for performance evaluation. The paradigm of formal evaluation was established by NIST in response to this need by providing the research community with a number of essential elements, such as manually labeled data sets, well-defined tasks, evaluation metrics, and a postevaluation workshop [87].

LREs were conducted by NIST in 1996, 2003, 2005, 2007, 2009, and 2011. It is evident that more languages are being included from year to year, as can be observed in Table 2, which summarizes the language profile. Another highlight is that NIST LRE has set out to focus on a detection task since its inception, as both closed-set and open-set identification problems can be easily formulated as applications of the detection task. Furthermore, performance measures such as identification accuracy, which depends on the number of target languages, could be factored out in the detection task.

NIST LREs have also sought to examine dialect detection capabilities, besides language detection. While the term dialect may have different linguistic definitions, it means a variety of language that is a particular characteristic of speakers from a specific geographical region. For instance, the dialects of interest include variants of Chinese, English, Hindustani, and Spanish in LRE 2007, as shown in Table 2. Dialect detection is generally believed to be more challenging [6], [79] than language detection, as the precise boundaries between dialects are not always clearly defined. In addition to the specified target languages (and dialects), several out-of-set languages (marked as "O") have also been included.

The emphasis of NIST LREs has been on conversational telephone speech (CTS), since most of the likely applications of the technology involve signals recorded from the public telephone system [87]. In order to collect speech data of more languages in a cost-effective way, NIST has adopted broadcast narrowband speech (BNBS) lately in LRE 2009 and LRE 2011 [26]. BNBS data are excerpts of call-in telephone speech embedded in broadcast and webcast. The call-in excerpts are used as we could expect them to cover as many speakers as possible. Broadcast entities like the Voice of America (VOA) broadcasts in more than 45 languages. Alternatively, the British Broadcast Company (BBC) also produces and distributes programs in a large number of languages. They have become an invaluable source of multilingual speech data.

B. Detection Cost and Detection Error Tradeoff (DET) Curve

In the detection task as defined in NIST LREs, system performance is evaluated by presenting the system with a set of trials, each consisting of a test segment and a

Table 2 Languages Included in the NIST Series of LRE. Target Languages Are Marked as "X" While Out-of-Set Languages Are Marked as "O"

Languages		LRE 1996	LRE 2003	LRE 2005	LRE 2007	LRE 2009	LRE 2011
Egyptian Arabic	Arabic	X	X		X	O	
Iraqi Arabic							X
Levantine Arabic							X
Modern Standard Arabic							X
Maghrebi Arabic							X
American English	English		X	X	X	X	X
General American		X					
Southern American		X					
Indian English				X	X	X	X
Amharic						X	
Azerbaijani						O	
Belorussian						O	
Bengali					X	O	X
Bosnian						X	
Bulgarian						O	
Cantonese	Chinese				X	X	
Mandarin			X		X	X	X
Mainland Mandarin		X		X	X		
Taiwan Mandarin		X		X	X		
Shanghai-Wu					X	O	
Shanghai-Min					X	O	
Creole (Haitian)						X	
Czech							X
Croatian						X	
Dari						X	X
French		X	X		O	X	
Georgian						X	
German		X	X	O	X		
Hausa						X	
Hindustani					X		
Hindi		X	X	X	X	X	X
Urdu					X	X	X
Indonesian					O		
Italian					O	O	
Japanese		X	X	X	X	O	
Korean		X	X	X	X	X	
Lao (Laotian)							X
Pashto						X	X
Farsi/Persian		X	X		X	X	X
Polish							X
Portuguese						X	
Punjabi					O	O	X
Romanian						O	
Russian			O		X	X	X
Slovak							X
Spanish			X	X	X	X	X
Caribbean Spanish		X			X		
Non-Caribbean Spanish		X			X		
Swahili						O	
Tagalog					O	O	
Tamil		X	X	X	X	O	X
Thai					X		X
Tibetan						O	
Turkish						X	X
Ukrainian						X	X
Uzbek						O	
Vietnamese		X	X		X	X	

hypothesized target language. The system has to decide for each trial whether the target language was spoken in the given segment.

Let N_{T} be the number of test segments and N be the number of target languages as defined earlier. By presenting each test segment against all target languages, there are N_{T} number of trials for each target and the system under evaluation should produce $N \times N_{\mathrm{T}}$ number of true or false decisions, one for each trial. The primary evaluation measure is the *average detection cost*, defined as follows:

$$C_{\mathrm{avg}} = \frac{1}{N}\sum_{l=1}^{N} C_{\mathrm{DET}}(L_l) \tag{30}$$

where $C_{\mathrm{DET}}(L_l)$ is the *detection cost* for the subset of N_{T} trials for which the target language is L_l

$$C_{\mathrm{DET}}(L_l) = C_{\mathrm{miss}} P_{\mathrm{tar}} P_{\mathrm{miss}}(L_l) + C_{\mathrm{fa}}(1-P_{\mathrm{tar}})\frac{1}{N-1}\sum_{m\neq l} P_{\mathrm{fa}}(L_l, L_m). \tag{31}$$

The *miss probability* (or false rejection rate) P_{miss} accounts for the error when a test segment of language L_l is rejected as being spoken in that language (i.e., classifying a target trial as a nontarget trial). The *false alarm probability* (or false acceptance rate) $P_{\mathrm{fa}}(L_l, L_m)$ accounts for the error when a test segment of language L_m is accepted as being spoken in language L_l (i.e., classifying a nontarget trial as a target trial). The probabilities are computed as the number of errors divided by the total number of trials in each subset.

The application parameters $\{C_{\mathrm{miss}}, C_{\mathrm{fa}}, P_{\mathrm{tar}}\}$ were set to the values $\{1, 1, 0.5\}$ in past evaluations, where the costs for making both types of errors C_{miss} and C_{fa} are set to be equal, and P_{tar} is the prior probability of a target foreseen for a particular application. In Section VI-C, we show how the P_{tar} can be used to set the decision threshold. Putting together these components, the average detection cost can be expressed as

$$C_{\mathrm{avg}} = C_{\mathrm{miss}} P_{\mathrm{tar}} \underbrace{\frac{1}{N}\sum_{l=1}^{N} P_{\mathrm{miss}}(L_l)}_{P_{\mathrm{miss}}(\theta_{\mathrm{DET}})} + C_{\mathrm{fa}}(1-P_{\mathrm{tar}}) \underbrace{\frac{1}{N}\sum_{l=1}^{N}\left[\frac{1}{N-1}\sum_{m\neq l} P_{\mathrm{fa}}(L_l, L_m)\right]}_{P_{\mathrm{fa}}(\theta_{\mathrm{DET}})}. \tag{32}$$

Notice that the miss probabilities $P_{\mathrm{miss}}(L_l)$ are computed separately for each target language, and for each l the false

alarm probabilities $P_{\text{fa}}(L_l, L_m)$ are computed for each target/nontarget language pairs. The direct implication of such a form of averaging computation is that the number of target trials per language is no longer of influence to the resulting C_{avg} measure. This fact was first recognized in LRE 2005 and, since then, it has been used in subsequent LREs. For the case of an open-set test, one or more unseen languages are used as *out-of-set* languages L_{oos}. This could be accounted for with an addition component $P_{\text{fa}}(L_l, L_{\text{oos}})$ in the second term of the detection cost in (32). See [88], [89], and [102] for details.

The hard decisions used for the computation of the detection cost in (32) are usually obtained based on whether the scores produced by a system exceed a chosen threshold $\theta = \theta_{\text{DET}}$. The intrinsic assumption with such a threshold-based decision is that the higher score supports more the target hypothesis, while the lower score supports the alternative. Another way to assess the system performance is to allow the threshold θ to change across a range of possible operating points. The resulting plot of $P_{\text{miss}}(\theta)$ against $P_{\text{fa}}(\theta)$ for different values of θ is referred to as the detection-error-tradeoff (DET) curve. An example of a DET plot is shown in Fig. 9. Notice that the axes of the DET plot are warped according to the *probit* function, i.e., the inverse cumulative density function of a standard Gaussian distribution, making it different from the traditional receiver-operating-characteristic (ROC) curve [43] in several ways. Notably, the curve becomes a straight line if the target and nontarget scores are normally distributed [84]. The direct consequence is a less cluttered plot than a set of ROC curves when several DET curves are produced on the same graph.

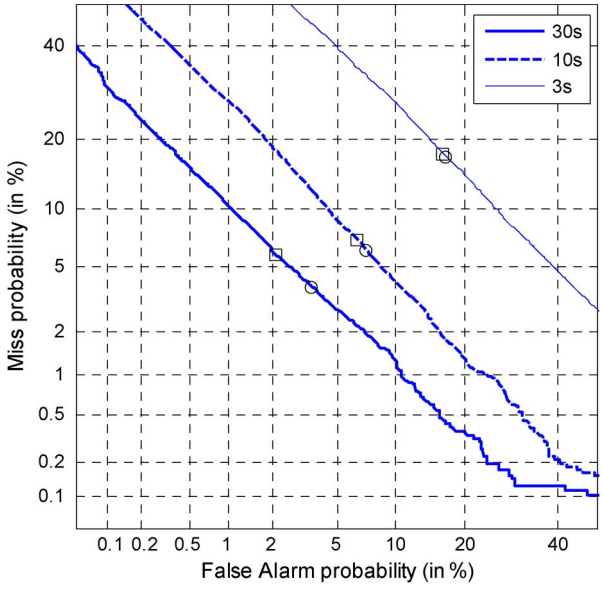

Fig. 9. ***DET curve showing the fusion system performance on the NIST LRE 2011 language detection task for test duration of 30, 10, and 3 s. Circles and squares indicate the minimum cost and actual decision points, respectively.***

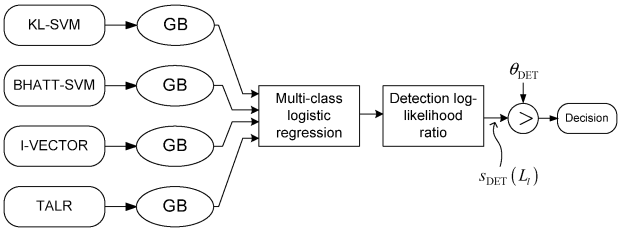

Fig. 10. ***Generative Gaussian back-end (GB) followed by multiclass logistic regression for calibration and fusion.***

The DET curve shows what happens as the decision threshold θ traverses across different operating points by which the error probabilities $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$ change in opposite directions. That is, there is an inherent tradeoff between the two errors. Another point to note is that the error probabilities are first computed separately for each target language and averaged to produce $P_{\text{miss}}(\theta)$, similarly for $P_{\text{fa}}(\theta)$. This is not the same as computing directly the error probabilities using the pooled scores (see [136] for a detailed analysis of why the latter should be avoided). Another visual advantage of DET is that we could show the actual decision operating point on the curve to see how "far" it is from the operating point producing the minimum cost.

C. Performance Assessment on NIST LRE Corpora

Since 1996, NIST has conducted six evaluations of the automatic language recognition technology, most recently in 2011. For the language detection tasks, the participants were required to provide two outputs for each trial, a hard decision and a soft score that represents the degree of support for the target hypothesis with respect to the alternative hypothesis. The hard decisions are used to compute the average detection cost C_{avg} in (32), which serves as the primary evaluation measure. The soft scores are used to assess the system performance across a wide range of operating points θ leading to the DET curve.

Fig. 9 shows the DET plots for the detection task with test segments of three different durations, namely, 30, 10, and 3 s, as designated in the LRE 2011. The DET plots show the results for a four-subsystem fusion in Fig. 10, which consists of three acoustic–phonetic subsystems and a phonotactic subsystem using the target-aware lattice rescoring (TALR) technique [73], [128], [129]. The three acoustic–phonetic subsystems are SVM with the Kullback–Leibler kernel (KL–SVM) [16], [18], SVM with Bhattacharyya kernel (BHATT–SVM) [145], [146]; and *i*-vector [34], [35]. The results as shown are comparable with other top

[2]The KL–SVM, BHATT–SVM, and TALR subsystems were taken from the IIR submission, while the i-vector subsystem was part of Brno276 submission in LRE 2011.

performing systems reported on the same LRE 2011 data set [13], [121].[2] Notice that the closer the curve is to the origin, it indicates the better performance. It is apparent that language recognition performance becomes increasingly challenging as the length of the segments decreases. Also marked on the DET plots are the minimum cost and actual decision points indicated, respectively, with a circle and a square on each curve. We obtain the minimum point by sweeping through the DET curve to find the operating point which produces the lowest cost. Ideally, this should coincide with the actual decision point $\theta = \theta_{\text{DET}}$ if the threshold was set properly.

The need to set a threshold is application dependent. Given the application parameters $\{C_{\text{miss}}, C_{\text{fa}}, P_{\text{tar}}\}$, a threshold is set to minimize the detection cost as in (32). Traditionally, this was done by probing the recognizer using a development data set and setting the threshold at the operating point that gives the lowest cost. The drawback of this approach is that the same procedure has to be repeated when a new set of application parameters is given. A more systematic approach, which has been widely accepted in the community, is to postprocess the scores via the so-called application-independent calibration (see Section V-B). Let $\{s'_l, l = 1, \ldots, N\}$ be the log-likelihood scores obtained from the calibration. For the detection task, we form the detection log-likelihood ratio [8], [10], [12]

$$s_{\text{DET}}(L_l) = \log \frac{\exp(s'_l)}{\frac{1}{N-1}\sum_{k \neq l} \exp(s'_k)}. \tag{33}$$

The detection threshold can then be set for any particular set of application parameters [8], [10], [12]

$$\theta_{\text{DET}} = -\log\left[\frac{C_{\text{miss}} P_{\text{tar}}}{C_{\text{fa}}(1 - P_{\text{tar}})}\right]. \tag{34}$$

In essence, the detection cost is minimized indirectly via the optimization of the surrogate cost in (28) formulated on the multiclass regression model.

Fig. 10 shows the schematic diagram of the four-subsystem fusion, the DET curve of which is depicted in Fig. 9. A common strategy, which was shown to be useful in several reports, is to use a Gaussian back–end (see Section V-A) to summarize each individual subsystem, and to fuse (and calibrate) the resulting Gaussian back–end score together with those from other subsystems via a multiclass logistic regression model. Fig. 11 shows the minimum and actual detection cost for the individual subsystems. In general, the acoustic–phonetic subsystems (i.e., the KL–SVM, BHATT–SVM, and i-vector) are equally competitive as their phonotactic TALR counterparts. It is

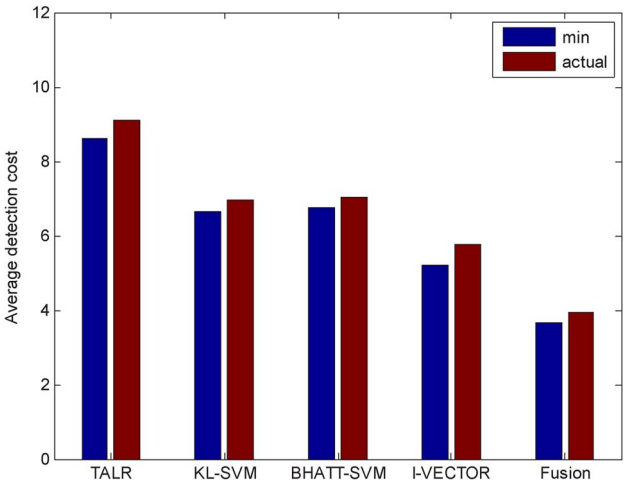

Fig. 11. *Performance of individual acoustic–phonetic and phonotactic subsystems together with their fusion in terms of actual and minimum (average) detection cost (LRE 2011, 30 s).*

generally believed that the major factor for attaining classifier fusion with a better accuracy is the diversity in the subsystems [61]. Such subsystems complement each other through the fusion process to produce the best performance. It is worth noting that the small gap between the actual and minimum costs in Fig. 11 shows the effectiveness and the need of score calibration.

VII. DISCUSSION AND FUTURE DIRECTIONS

In this paper, we have discussed several key aspects of language recognition, covering language characterization, modeling techniques, and system development strategies. While we have seen tremendous progresses in the past decades, language recognition is still far from perfect. As far as language characterization is concerned, we have not been able to effectively venture beyond acoustic–phonetic and phonotactic knowledge, despite the fact that there exists strong evidence in human listening experiments that prosodic information, syllable structure, and morphology are useful knowledge sources. Nonetheless, recent advances reported in [33] and [60] have shown a revival of interest in prosodic features as well as their effective modeling techniques.

The study of phonetics and linguistics has a long history, which provides insights into the origin of language families and the lexical–phonological relationship between languages. However, we have yet to benefit from a wealth of such discoveries [1]. Nonetheless, with a better understanding of the fundamental problems, we have seen several encouraging directions in which we could further the research. We briefly name a few in the following.

In phonotactic approaches, it is clear that good performance relies on effective phonotactic characterization of

speech. As shown in Section II-A, perfect phone transcription will provide distinctive phonotactic statistics for languages. An obvious option is to improve the phone recognizers in the tokenization front–end with a better modeling technique, such as the left-context–right-context FeatureNet [90] and deep neural networks [50], [78]. Nonetheless, we also realize that we will never be able to achieve perfect phone transcription. An alternative is to find a better way in extracting the phone *n*-gram statistics using existing phone recognizers, while continuing the efforts to improve the phone recognition accuracy. Several attempts have shown positive results along this direction. The lattice-based phone *n*-gram statistics [40] make use of the posterior probability as the soft counts as opposed to phone *n*-gram counts. The target-oriented phone tokenizer [128] and target-aware lattice rescoring [129] suggest different ways to make use of the same phone recognizer to create diverse statistics from the same source. The cross-decoder (or cross-recognizer) phone co-occurrences *n*-gram [103], for the first time, explores phonotactic information across multiple phone sequences from parallel phone recognizers. These studies have shown that it is worth the effort to explore phonotactic information from different perspectives.

In acoustic approaches, the latent factor modeling technique, as used in JFA [57], [140] and i-vector [34], is unleashing its potential in language recognition [35]. We envisage its use for hierarchical modeling of languages. In particular, a latent factor model could be used to capture the majority of useful variability among languages from a distinct language cluster, using separate subspaces for each language cluster. Discrimination between languages can then be done locally in each subspace corresponding to a language cluster. From a modeling perspective, the idea can be seen as an extension to the i-vector approach (see Section IV-D), where we used multiple subspaces, instead of one total variability space, each corresponding to a language cluster.

In system development, we have shown the usefulness of the Gaussian back–end for decision fusion and calibration. One subtle problem that a system designer may encounter is the duration-dependent nature of the Gaussian back–end. The score vectors produced by the front–end recognizer exhibit larger variation for the test segments with longer durations. This problem is traditionally dealt with by having a separate back–end for each of the nominal durations the system designers could foresee. In [91], a simple parametric function is used to scale the log-likelihood scores so that a single Gaussian back–end could be used for all durations. A more elegant, but challenging, way could be to model directly the covariance matrix as a function of the test segment duration, so that the covariance of the Gaussian back–end could be updated on the fly according to the test duration.

The need for fast, efficient, accurate, and robust means of language recognition is of growing importance for commercial, forensic, and government applications. The Speaker and Language Characterization (SpLC), a special interest group of the International Speech Communication Association (ISCA), started the Odyssey Speaker and Language Recognition Workshop in 1994 as a forum to foster interactions among researchers and to promote language recognition research. Other scientific forums include the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) and the Annual Conference of the International Speech Communication Association (INTERSPEECH), where the latest findings are reported. ■

REFERENCES

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 82–108, May 2011.

[2] M. Ashby and J. Maidment, *Introducing Phonetic Science*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[3] T. Baldwin and M. Lui, "Language identification: The long and the short of the matter," in *Proc. Annu. Conf. North Amer. Chapter ACL Human Lang. Technol.*, Los Angeles, CA, USA, pp. 229–237.

[4] K. M. Berkling and E. Barnard, "Analysis of phoneme-based features for language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, 1994, pp. 289–292.

[5] K. M. Berkling and E. Barnard, "Language identification of six languages based on a common set of broad phonemes," in *Proc. Int. Conf. Spoken Lang. Process.*, Yokohama, Japan, 1994, pp. 1891–1894.

[6] F. Biadsy, H. Soltauy, L. Manguy, J. Navratily, and J. Hirschberg, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," in *Proc. IEEE Odyssey: Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 263–270.

[7] J. Braun and H. Levkowitz, "Automatic language identification with perceptually guided training and recurrent neural networks," in *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, 1998, pp. 289–292.

[8] N. Brummer and D. Leeuwen, "On calibration of language recognition scores," in *Proc. IEEE Odyssey: Speaker Lang. Recognit. Workshop*, San Juan, Puerto Rico, 2006, DOI: 10.1109/ODYSSEY.2006.248106.

[9] N. Brummer and J. Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, no. 2, pp. 230–275, 2006.

[10] N. Brummer, *Focal Multi-Class—Tools for Evaluation, Calibration and Fusion of, and Decision-Making with, Multi-Class Statistical Pattern Recognition Scores*, Jun. 2007. [Online]. Available: http://sites.google.com/site/nikobrummer/

[11] N. Brummer, L. Burget, O. Glembek, V. Hubeika, Z. Jancik, M. Karafiat, P. Matejka, T. Mikolov, O. Plchot, and A. Strasheim, "BUT-AGNITIO system description for NIST language recognition evaluation 2009," in *Proc. NIST Lang. Recognit. Eval. Workshop*, Baltimore, MD, USA, 2009.

[12] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Dept. Electr. Electron. Eng., Stellenbosch Univ., Stellenbosch, South Africa, 2010.

[13] N. Brümmer, S. Cumani, O. Glembek, M. Karafiát, P. Matějka, J. Pešán, O. Plchot, M. Soufifar, E. de Villiers, and J. Černocký, "Description and analysis of the Brno276 system for LRE2011," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, Singapore, 2012, pp. 216–223.

[14] L. Burget, P. Matejka, and J. Cernocky, "Discriminative training techniques for acoustic language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Toulouse, France, 2006, pp. 209–212.

[15] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-level speaker verification with support vector machines," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Montreal, QC, Canada, 2004, pp. 73–76.

[16] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 210–229, 2006.

[17] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation," in *Proc. IEEE Odyssey: Speaker Lang. Recognit. Workshop*, San Juan, Puerto Rico, 2006, DOI: 10.1109/ODYSSEY.2006.248097.

[18] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–310, May 2006.

[19] W. M. Campbell, "A covariance Kernel for SVM language recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, USA, 2008, pp. 4141–4144.

[20] W. M. Campbell, D. E. Sturim, P. Torres-Carrasquillo, and D. A. Reynolds, "A comparison of subspace feature-domain methods for language recognition," in *Proc. Interspeech Conf.*, Brisbane, Australia, 2008, pp. 309–312.

[21] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 1969–1978, Sep. 2007.

[22] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proc. 3rd Annu. Symp. Document Anal. Inf. Retrieval*, Las Vegas, NV, USA, 1994, pp. 161–175.

[23] C. Chelba, T. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 39–49, May 2008.

[24] J. Chu-Carrol and B. Carpenter, "Vector-based natural language call routing," *Comput. Linguist.*, vol. 25, no. 3, pp. 361–388, 1999.

[25] C. Cieri, J. P. Campbell, H. Nakasone, D. Miller, and K. Walker, "The mixer corpus of multilingual, multichannel speaker recognition data," in *Proc. Int. Conf. Lang. Resources Eval.*, Lisbon, Portugal, 2004, pp. 24–30.

[26] C. Cieri, L. Brandschain, A. Neely, D. Graff, K. Walker, C. Caruso, A. Martin, and C. Greenberg, "The broadcast narrow band speech corpus: A new resource type for large scale language recognition," in *Proc. Interspeech Conf.*, Brighton, U.K., 2009, pp. 2867–2870.

[27] D. Cimarusti and R. Ives, "Development of an automatic identification system of spoken languages: Phase I," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Paris, France, 1982, pp. 1661–1663.

[28] H. P. Combrinck and E. C. Botha, "Automatic language identification: Performance vs. complexity," in *Proc. 8th Annu. South Africa Workshop Pattern Recognit.*, Grahams Town, South Africa, 1997.

[29] B. Comrie, *The World's Major Languages*. Oxford, U.K.: Oxford Univ. Press, 1990.

[30] D. Crystal, *The Cambridge Factfinder*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[31] S. Cumani and P. Laface, "Analysis of large-scale SVM training algorithms for language and speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1585–1596, Jul. 2012.

[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[33] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2095–2103, Sep. 2007.

[34] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[35] N. Dehak, P. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. Interspeech Conf.*, Florence, Italy, 2011, pp. 857–860.

[36] A. Dumpster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[37] T. Dunning, "Statistical identification of language," Comput. Res. Lab (CRL), New Mexico State Univ., Las Cruces, NM, USA, Tech. Rep. MCCS-94-273, 1994.

[38] J. Foil, "Language identification using noisy speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Tokyo, Japan, 1986, pp. 861–864.

[39] J. L. Gauvain and C.-H. Lee, "Maximum a posterior estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[40] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju Island, Korea, 2004, pp. 1283–1286.

[41] E. M. Gold, "Language identification in the limit," *Inf. Control*, vol. 10, no. 5, pp. 447–474, 1967.

[42] F. J. Goodman, A. F. Martin, and R. E. Wohlford, "Improved automatic language identification in noisy speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, 1989, vol. 1, pp. 528–531.

[43] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2008.

[44] T. J. Hazen and V. W. Zue, "Automatic language identification using a segment-based approach," in *Proc. Eurospeech Conf.*, Berlin, Germany, 1993, pp. 1303–1306.

[45] T. J. Hazen and V. W. Zue, "Recent improvements in an approach to segment-based automatic language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, 1994, pp. 1883–1886.

[46] T. J. Hazen and V. W. Zue, "Segment-based automatic language identification," *J. Acoust. Soc. Amer.*, vol. 101, no. 4, pp. 2323–2331, 1997.

[47] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[48] J. Hieronymus and S. Kadambe, "Spoken language identification using large vocabulary speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, USA, 1996, pp. 1780–1783.

[49] G. Hinton, "Products of experts," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, Edinburgh, U.K., 1999, vol. 1, pp. 1–6.

[50] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[51] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *J. Acoust. Soc. Amer.*, vol. 62, no. 3, pp. 708–713, 1977.

[52] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[53] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithms," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[54] B. H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.

[55] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2000.

[56] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, Toledo, Spain, 2004, pp. 219–226.

[57] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[58] K. Kirchhoff, "Language characteristics," in *Multilingual Speech Processing*, T. Schultz and K. Kirchhoff, Eds. Amsterdam, The Netherlands: Elsevier, 2006.

[59] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Anal. Appl.*, no. 1, pp. 18–27, 1988.

[60] M. Kockmann, L. Ferrer, L. Burget, and J. Černocký, "iVector fusion of prosodic and cepstral features for speaker verification," in *Proc. Interspeech Conf.*, Florence, Italy, 2011, pp. 265–268.

[61] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, Feb. 2002.

[62] L. F. Lamel and J. L. Gauvain, "Language identification using phone-based acoustic likelihoods," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, 1994, vol. 1, pp. 293–296.

[63] T. Lander, R. Cole, B. Oshika, and M. Noel, "The OGI 22 language telephone speech corpus," in *Proc. Eurospeech Conf.*, Madrid, Spain, 1995, pp. 817–820.

[64] C.-H. Lee, "Principles of spoken language recognition," in *Springer Handbook of Speech Processing and Speech Communication*, J. Benesty, M. M. Sondhi, and A. Huang, Eds. New York, NY, USA: Springer-Verlag, 2008.

[65] K. A. Lee, C. You, and H. Li, "Spoken language recognition using support vector machines with generative front-end," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, USA, 2008, pp. 4153–4156.

[66] K. A. Lee, C. H. You, H. Li, T. Kinnunen, and K. C. Sim, "Using discrete probabilities with Bhattacharyya measure for SVM-based speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 861–870, May 2011.

[67] K. A. Lee, C. H. You, V. Hautamäki, A. Larcher, and H. Li, "Spoken language recognition in the latent topic simplex," in *Proc. Interspeech Conf.*, Florence, Italy, 2011, pp. 2933–2936.

[68] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Atlanta, GA, USA, 1996, vol. 1, pp. 353–356.

[69] R. G. Leonard and G. R. Doddington, "Automatic language identification," Air Force Rome Air Develop. Cntr., Tech. Rep. RADC-TR-74-200, Aug. 1974.

[70] R. G. Leonard and G. R. Doddington, "Automatic language identification," Air Force Rome Air Develop. Cntr., Tech. Rep. RADC-TR-75-264, Oct. 1975.

[71] M. P. Lewis, Ed., *Ethnologue: Languages of the World*, 16 ed. Dallas, TX, USA: SIL International, 2009.

[72] H. Li and B. Ma, "A phonotactic language model for spoken language identification," in *Proc. Assoc. Comput. Linguist.*, Ann Arbor, MI, USA, 2005, pp. 515–522.

[73] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 271–284, Jan. 2007.

[74] H. Li, B. Ma, K. C. Sim, R. Tong, K. A. Lee, H. Sun, D. Zhu, C. You, M. Dong, and X. Wang, "Institute for Infocomm Research system description for the language recognition evaluation 2007 submission," in *Proc. NIST Lang. Recognit. Eval. Workshop*, Orlando, FL, USA, 2007.

[75] H. Li, B. Ma, and C.-H. Lee, "Vector-based spoken language classification," in *Springer Handbook of Speech Processing and Speech Communication*, J. Benesty, M. M. Sondhi, and A. Huang, Eds. New York, NY, USA: Springer-Verlag, 2008.

[76] H. Li and B. Ma, "TechWare: Speaker and spoken language recognition resources," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 139–142, Nov. 2010.

[77] K. P. Li and T. J. Edwards, "Statistical models for automatic language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Denver, CO, USA, 1980, pp. 884–887.

[78] X. Li, L. Deng, and J. Bilmes, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio Speech Lang. Process.*, accepted for publication.

[79] B. P. Lim, H. Li, and B. Ma, "Using local and global phonotactic features in Chinese dialect identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Philadelphia, PA, USA, 2005, pp. 577–580.

[80] S. Lucey and T. Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Proc. Eurospeech Conf.*, Geneva, Switzerland, 2003, pp. 2021–2024.

[81] B. Ma, C. Guan, H. Li, and C.-H. Lee, "Multilingual speech recognition with language identification," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, USA, 2002, pp. 505–508.

[82] B. Ma, H. Li, and R. Tong, "Spoken language recognition with ensemble classifiers," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2053–2062, Sep. 2007.

[83] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[84] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech Conf.*, Rhodes, Greece, 1997, vol. 4, pp. 1895–1898.

[85] A. F. Martin and M. A. Przybocki, "NIST 2003 language recognition evaluation," in *Proc. Eurospeech Conf.*, Geneva, Switzerland, 2003, pp. 1341–1344.

[86] A. F. Martin and A. N. Le, "The current state of language recognition: NIST 2005 evaluation results," in *Proc. IEEE Odyssey: Speaker Lang. Recognit. Workshop*, San Juan, Puerto Rico, 2006, DOI: 10.1109/ODYSSEY.2006.248104.

[87] A. F. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *Proc. IEEE Workshop Signal Process. Appl. Public Security Forensics*, Washington, DC, USA, 2007, pp. 1–7.

[88] A. F. Martin and A. N. Le, "NIST 2007 language recognition evaluation," presented at the Odyssey: Speaker Lang. Recognit. Workshop, Stellenbosch, South Africa, 2008, paper 016.

[89] A. F. Martin and C. Greenberg, "The 2009 NIST language recognition evaluation," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, Brno, Czech Republic, 2010, pp. 165–171.

[90] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proc. Interspeech Conf.*, Lisbon, Portugal, 2005, pp. 2237–2240.

[91] A. McCree, F. Richardson, E. Singer, and D. Reynolds, "Beyond frame independent: Parametric modeling of time duration in speaker and language recognition," in *Proc. Interspeech Conf.*, Brisbane, Australia, 2008, pp. 767–770.

[92] S. Mendoza, L. Gillick, Y. Ito, S. Lowe, and M. Newman, "Automatic language identification using large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Atlanta, GA, USA, 1996, vol. 2, pp. 785–788.

[93] T. P. Minka, "Algorithms for maximum-likelihood logistic regression," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. 738, 2001.

[94] Y. K. Muthusamy, R. Cole, and B. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. Int. Conf. Spoken Lang. Process.*, BanffABCanada, 1992, pp. 895–898.

[95] Y. K. Muthusamy, K. M. Berkling, T. Arai, R. A. Cole, and E. Barnard, "A comparison of approaches to automatic language identification using telephone speech," in *Proc. Eurospeech Conf.*, Berlin, Germany, 1993, pp. 1307–1310.

[96] Y. K. Muthusamy, N. Jain, and R. A. Cole, "Perceptual benchmarks for automatic language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, 1994, vol. 1, pp. 333–336.

[97] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 33–41, Oct. 1994.

[98] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text-independent language identification by HMM," in *Proc. Int. Conf. Spoken Lang. Process.*, Banff, AB, Canada, 1992, pp. 1011–1014.

[99] J. Navratil, "Spoken language recognition—A step toward multilinguality in speech processing," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 678–685, Sep. 2001.

[100] R. W. M. Ng, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Prosodic attribute model for spoken language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 5022–5025.

[101] R. W. M. Ng, C.-C. Leung, V. Hautamäki, T. Lee, B. Ma, and H. Li, "Towards long-range prosodic attribute modeling for language recognition," in *Proc. Interspeech Conf.*, Chiba, Japan, 2010, pp. 1792–1795.

[102] *NIST Language Recognition Evaluations.* [Online]. Available: http://nist.gov/itl/iad/mig/lre.cfm

[103] M. Penagarikano, A. Varona, L. J. Rodriguez-Fuentes, and G. Bordel, "Improved modeling of cross-decoder phone co-occurrences in SVM-Based phonotactic language recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 8, pp. 2348–2363, Nov. 2011.

[104] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to fusion of the NIST'99 1-speaker submissions," *Digital Signal Process.*, vol. 10, pp. 237–248, 2000.

[105] D. Qu and B. Wang, "Discriminative training of GMM for language identification," in *Proc. ISCA IEEE Workshop Spontaneous Speech Process. Recognit.*, Tokyo, Japan, 2003, pp. 67–70.

[106] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[107] L. R. Rabiner, "A tutorial on hidden Markov models and selected publication in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[108] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech re-synthesis," *J. Acoust. Soc. Amer.*, vol. 105, no. 1, pp. 512–521, 1999.

[109] R. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[110] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models,"

IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72–83, Jan. 1995.

[111] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[112] F. S. Richardson and W. M. Campbell, "Language recognition with discriminative keyword selection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, USA, 2008, pp. 4145–4148.

[113] G. Salton, *The SMART Retrieval System*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[114] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[115] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Atlanta, GA, USA, 1996, vol. 2, pp. 781–784.

[116] T. Schultz and A. Waibel, "Language independent and language adaptive," *Speech Commun.*, vol. 35, no. 1–2, pp. 31–51, 2001.

[117] T. Schultz, "Globalphone: A multilingual text and speech database developed at Karlsruhe University," in *Proc. Interspeech Conf.*, Denver, CO, USA, 2002, pp. 345–348.

[118] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," in *Proc. IEEE Odyssey: Speaker Lang. Recognit. Workshop*, San Juan, Puerto Rico, 2006, DOI: 10.1109/ODYSSEY.2006.248100.

[119] W. Shen and D. A. Reynolds, "Improved GMM-Based language recognition using constrained MLLR transforms," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, USA, 2008, pp. 4149–4152.

[120] K. C. Sim and H. Li, "On acoustic diversification front-end for spoken language identification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 5, pp. 1029–1037, Jul. 2008.

[121] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, "The MITLL NIST LRE2011 language recognition system," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, Singapore, 2012, pp. 209–215.

[122] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Proc. Interspeech Conf.*, Brighton, U.K., 2009, pp. 168–171.

[123] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition," in *Proc. Interspeech Conf.*, Chiba, Japan, 2010, pp. 2718–2721.

[124] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Philadelphia, PA, USA, 2005, pp. 629–632.

[125] M. Sugiyama, "Automatic language recognition using acoustic features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Toronto, ON, Canada, 1991, vol. 2, pp. 813–816.

[126] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysis," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, 1999.

[127] R. Tong, B. Ma, D. Zhu, H. Li, and E.-S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Toulouse, France, 2006, pp. 205–208.

[128] R. Tong, B. Ma, H. Li, and E. Chng, "A target-oriented phonotactic front-end for spoken language recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 7, pp. 1335–1347, Sep. 2009.

[129] R. Tong, B. Ma, H. Li, and E. Chng, "Target-aware lattice rescoring for dialect recognition," in *Proc. Interspeech Conf.*, Florence, Italy, 2011, pp. 733–736.

[130] P. A. Torres-Carrasquillo, D. A. Reynolds, and R. J. Deller, Jr., "Language identification using Gaussian mixture model tokenization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, USA, 2002, pp. 757–760.

[131] P. Torres-Carrasquillo, E. Singer, M. Kohler, R. Greene, D. Reynolds, and J. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, USA, 2002, pp. 89–92.

[132] P. Torres-Carrasquillo, E. Singer, W. Campbell, T. Gleason, A. McCree, D. Reynolds, F. Richardson, W. Shen, and D. Sturim, "The MITLL NIST LRE 2007 language recognition system," in *Proc. Interspeech Conf.*, Brisbane, Australia, 2008, pp. 719–722.

[133] R. C. Tucker, M. J. Carey, and E. S. Paris, "Automatic language identification using sub-words models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Adelaide, Australia, 1994, pp. 301–304.

[134] D. A. van Leeuwen and N. Brummer, "Channel-dependent GMM and multi-class logistic regression models for language recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recognit. Workshop*, San Juan, Puerto Rico, 2006, DOI: 10.1109/ODYSSEY.2006.248094.

[135] D. A. van Leeuwen and N. Brümmer, "An introduction to application independent evaluation of speaker recognition systems," in *Speaker Classification*, vol. 4343, R. Müller, Ed. Berlin, Germany: Springer-Verlag, 2007.

[136] D. A. van Leeuwen and K. P. Truong, "An open-set detection evaluation methodology applied to language and emotion recognition," in *Proc. Interspeech Conf.*, Antwerp, Belgium, 2007, pp. 338–341.

[137] D. A. van Leeuwen, M. Boer, and R. Orr, "A human benchmark for the NIST language recognition evaluation 2005," presented at the Odyssey: Speaker Lang. Recognit. Workshop, Stellenbosch, South Africa, 2008, paper 012.

[138] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[139] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recognit. Workshop*, San Juan, Puerto Rico, 2006, DOI: 10.1109/ODYSSEY.2006.248117.

[140] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Comput. Speech Lang.*, vol. 22, pp. 17–38, 2008.

[141] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, no. 8, pp. 1181–1190, Aug. 2000.

[142] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 25–47, 2002.

[143] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Detroit, MI, USA, 1995, vol. 5, pp. 3511–3514.

[144] Y. Yan, E. Barnard, and R. Cole, "Development of an approach to language identification based on phone recognition," *Comput. Speech Lang.*, vol. 10, pp. 37–54, 1996.

[145] C. H. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1300–1312, Aug. 2010.

[146] C. H. You, H. Li, and K. A. Lee, "A GMM-supervector approach to language recognition with adaptive relevance factor," in *Proc. EUSIPCO*, Aalborg, Denmark, 2010, pp. 1993–1997.

[147] J. Zhao, H. Shu, L. Zhang, X. Wang, Q. Gong, and P. Li, "Cortical competition during language discrimination," *NeuroImage*, vol. 43, pp. 624–633, 2008.

[148] D. Zhu, B. Ma, and H. Li, "Soft margin estimation of Gaussian mixture model parameters for spoken language recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 4990–4993.

[149] M. A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Minneapolis, MN, USA, 1993, vol. 2, pp. 399–402.

[150] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[151] M. A. Zissman, "Predicting, diagnosing and improving automatic language identification performance," in *Proc. Eurospeech Conf.*, Rhodes, Greece, 1997, pp. 51–54.

ABOUT THE AUTHORS

Haizhou Li (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively.

Currently, he is the Principal Scientist and Department Head of Human Language Technology, Co-Director of Baidu-I^{2}R Research Centre, Institute for Infocomm Research, Agency for Science, technology, and Research (A*STAR), Singapore. He is also a conjoint Professor at the School of Electrical Engineering and Telecommunications, University of New South Wales, Kensington, NSW, Australia. He has worked on speech and language technology in academia and industry since 1988. He taught at the University of Hong Kong (1988–1990), South China University of Technology (1990–1994), and Nanyang Technological University (2006–present). He was a Visiting Professor at CRIN in France (1994–1995). He was appointed as Research Manager at the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), and Vice President in InfoTalk Corp., Ltd., (2001–2003). His current research interests include automatic speech recognition, speaker and language recognition, and natural language processing. He has published over 200 technical papers in international journals and conferences.

Dr Li has served as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, *ACM Transactions on Speech and Language Processing*, and *Computer Speech and Language*, and a Guest Editor of the PROCEEDINGS OF THE IEEE. He is an elected Board Member of the International Speech Communication Association (ISCA, 2009–2013), President-Elect of Asia Pacific Signal and Information Processing Association (APSIPA, 2013–2014), and Executive Committee Member of the Asian Federation of Natural Language Processing (AFNLP, 2010–2012). He was a recipient of the National Infocomm Award of Singapore in 2001. He was named one of the two Nokia Visiting Professors 2009 by the Nokia Foundation in recognition of his contribution to speaker and language recognition technologies.

Bin Ma (Senior Member, IEEE) received the B.Sc. degree in computer science from Shandong University, Jinan, China, in 1990, the M.Sc. degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 1993, and the Ph.D. degree in computer engineering from The University of Hong Kong, Hong Kong, in 2000.

He was a Research Assistant from 1993 to 1996 at the National Laboratory of Pattern Recognition, IACAS. In 2000, he joined Lernout & Hauspie Asia Pacific as a Researcher working on the speech recognition of multiple Asian languages. From 2001 to 2004, he worked for InfoTalk Corp., Ltd, as a Senior Researcher and a Senior Technical Manager engaging in telephony speech recognition. Since 2004, he has been working for the Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore, and is currently a Senior Scientist and the Group Leader of Speech Processing Group. His current research interests include robust speech recognition, speaker and language recognition, spoken document retrieval, natural language processing, and machine learning.

Dr. Ma has served as a Subject Editor of *Speech Communication*.

Kong Aik Lee (Member, IEEE) received the B.Eng. (first class honors) degree from the University Technology Malaysia, Skudai, Malaysia, in 1999 and the Ph.D. degree from Nanyang Technological University, Singapore, in 2006.

He is currently a Scientist with the Human Language Technology Department, Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore. His research area covers speaker recognition, spoken language recognition, speech signal processing, and statistical pattern classification. He is the leading author of the book *Subband Adaptive Filtering: Theory and Implementation* (New York, NY, USA: Wiley, 2009). He has been participating in the NIST SRE and LRE evaluations since 2006.