

Higher-Level Features in Speaker Recognition

Elizabeth Shriberg

SRI International, Menlo Park, CA
International Computer Science Institute, Berkeley, CA
ees@speech.sri.com

Abstract. Higher-level features based on linguistic or long-range information have attracted significant attention in automatic speaker recognition. This article briefly summarizes approaches to using higher-level features for text-independent speaker verification over the last decade. To clarify how each approach uses higher-level information, features are described in terms of their type, temporal span, and reliance on automatic speech recognition for both feature *extraction* and feature *conditioning*. A subsequent analysis of higher-level features in a state-of-the-art system illustrates that (1) a higher-level cepstral system outperforms standard systems, (2) a prosodic system shows excellent performance individually and in combination, (3) other higher-level systems provide further gains, and (4) higher-level systems provide increasing relative gains as training data increases. Implications for the general field of speaker classification are discussed.

Keywords: Speaker recognition, speaker verification, higher-level features, high-level features, long-range features, prosodic features, stylistic features, automatic speech recognition, prosody, phonetic speaker recognition, speaker idiosyncrasies.

1 Introduction

The broad field of speaker classification makes use of a wide range of properties of spoken language—from lower-level features reflecting voice parameters to higher-level features that capture phonetic, prosodic, and lexical information. In subfields such as emotion, language, and dialect classification, among others, higher-level features play an important role in both human-based and automatic classification. In forensic phonetics, for example, it is common practice for human experts to use not only voice characteristics but also speaker characteristics based on pronunciation, prosody, and lexical information to assess the match between a suspect's speech and speech in a recording of interest.

In contrast, in automatic speaker recognition, the dominant approach in both commercial and research systems has been the use of Gaussian mixture models (GMMs) to model distributions of spectral information from short time frames of speech [1,2,3]. This approach, which reflects information about a speaker's vocal physiology, is highly successful, is simple, and has the further advantage of applicability to text-independent recognition since it does not rely on phonetic content. Nevertheless, it fails to capture

a wealth of longer-range and linguistic information that also resides in the signal. As we will see, such higher-level information can significantly improve performance when combined with lower-level cepstral information. Higher-level information also offers the possibility of increased robustness to channel variation, since features such as lexical usage or temporal patterns do not change with changes in acoustic conditions. And finally, higher-level features can provide useful metadata about a speaker, such as what topic is being discussed, how a speaker is interacting with another talker, whether the speaker is emotional or disfluent, and so on.

The last decade has seen increased interest in exploring such higher-level features in automatic speaker recognition. One facilitating factor in this regard has been the greater availability of high-performance automatic speech recognition (ASR) systems. A second is the development of data resources and task definitions that encourage the study of higher-level features, which typically operate at longer ranges and thus require longer train and test samples. An influential task has been the “Extended Data” task in speaker recognition evaluations (SREs) conducted by the National Institute of Standards and Technology (NIST) [4]. Since its introduction in 2001, the task has provided speaker training and test data consisting of whole conversation sides, with multiple sides available in training.

The goals of this chapter are (1) to provide a brief overview of work on higher-level features, and (2) to demonstrate how higher-level features can contribute to performance in a state-of-the-art system. Since the term “higher-level” (as well as a host of similar terms) has had different meanings in the literature, a working definition is certainly in order. We will take a broad view and include as higher-level any features that involve either *linguistic information* or *information at longer time spans than used in frame-based systems*. As discussed in more detail in the section to follow, linguistic information will refer to information that requires an automatic speech recognition system. Linguistic information is further specified according to its use for either feature extraction or feature conditioning. Longer-time-span information refers to features that are either extracted over regions longer than a frame or to approaches that model frame sequence dynamics. As will be shown, approaches using linguistic information are typically longer-range, but this is not always the case. And conversely, approaches using longer-range information often, but not always, use linguistic information. An example of approaches that fall outside the scope of this review (i.e., that are not considered higher-level) are those based on distributions of frame-level pitch or energy [5,6,7,8], which, although often called “prosodic,” involve neither linguistic nor long-range information.

Section 2 provides a summary of approaches to using higher-level features for text-independent speaker verification over the last decade. To clarify how each approach uses higher-level information, features are described in terms of a classification framework that specifies their type, time span, and reliance on automatic speech recognition for feature extraction and feature conditioning. Section 3 presents an analysis of higher-level features in a state-of-the-art system that includes multiple low-level and higher-level subsystems. Section 4 concludes with general implications for speaker recognition and the wider field of speaker classification.

2 Overview and Classification of Approaches

Although not exhaustive, the overview in this section aims to illustrate the wide array of methods and techniques used in higher-level feature modeling. A summary is presented in Table 1. In the first column of the table, and in the text to follow, features are grouped according to their feature *type*—progressing from lower-level cepstral features and features that essentially tokenize the acoustic space, to prosodic and finally to word-based features. The *description* of the feature is meant to convey its nature and contrast it to other features in the table; it may not match the original term used by the researcher(s).

To further specify just how each approach uses high-level information, and to contrast it with other approaches, three additional dimensions are introduced:

1. The temporal span of the feature
2. The level of ASR used for *feature extraction*
3. The level of ASR used for *region conditioning*

A feature's *time span* refers to the temporal region over which the feature is defined. We make a contrast here between frame-level and longer regions; this, of course, is a continuum. Note that a longer time span can be the result either of using a longer feature extraction region (e.g., a region based on lexical information) or of modeling sequential information based on frame-level features (e.g., pitch or energy dynamics over a sequence of many frames).

ASR used for feature extraction will refer to the highest level of ASR information needed to define and extract the feature. Features that require the output of an automatic speech recognition system necessarily involve some amount of linguistic information, but ASR systems can utilize varying degrees of linguistic constraints. At one end of the continuum are “open loop” phone recognizers, which decode using acoustic phone models but no phonotactic, lexical, or syntactic constraints. These systems essentially provide a means of tokenizing the acoustic space according to recognizer phone models. They often produce unusual (even unpronounceable) phone sequences that differ from those associated with the possible dictionary pronunciations for the words a speaker actually says. It is precisely because of these mismatches that such systems are useful in speaker recognition: the mismatches can reveal acoustic tendencies correlated with particular speakers. A step further in the direction of linguistic constraints involves imposing phonotactic constraints obtained from an N-phone language model. This approach restricts output to phone sequences that are observed in the language. At the extreme, the recognizer uses pronunciation dictionaries and word-level N-gram language models to hypothesize phones and words that make sense as part of complete sentence hypotheses. Higher-level features based on such output aim to capture information associated with specific words or word sequences, including not only their frequency of occurrence but also their acoustic realization, pronunciation, and prosodic rendering.

Finally, *ASR used for region conditioning* will be used to refer to the highest level of ASR required for filtering the output stream of features. If chosen appropriately, conditioning can improve speaker recognition in two ways: by reducing variability or by shifting means. Conditioning can reduce the variance of feature distributions by collecting data over more constrained (and thus more homogeneous) regions. And it

can focus on regions that exhibit greater inherent between-speaker variation, i.e., that move the means of one speaker's feature distribution farther away from those of other speakers. Both effects result in improved speaker discrimination.

While in principle any features can restrict comparison regions to subspans of speech, region conditioning fragments the data. Thus, there is typically some linguistic motivation that the cost of data fragmentation will be more than offset by the reduction in variability and/or shift in means brought about by the conditioning. A simple example is that of text-constrained cepstral features. The features themselves are neither long-range nor ASR-dependent. The only sense in which they are higher-level is in their region conditioning, which compares cepstral features of specific words or subword units to reduce within-speaker spectral variability associated with phonetic content. A second example is the maximum likelihood linear regression (MLLR) approach (see below), which factors out phonetic content both at the level of features (by using transforms derived from ASR phone information) and at the level of conditioning (by comparing transforms for specific phones individually).

Region conditioning is not restricted to variability reduction for phonetic content. For example, as described further below, the constrained prosody sequence approach conditions on sets of words that behave similarly prosodically. Although in principle region conditioning includes any means of reducing within-speaker variability on the feature of interest, in practice it has typically involved information from ASR. An exception, however, is the conditioned syllable-based prosody sequence model (see below), which in addition to conditioning on words makes significant use of pause contexts (obtainable without ASR). One can imagine other contexts (e.g., regions of high energy) that could also prove useful in constraining regions of interest.

A description of the approaches in Table 1 follows. Because studies differ in terms of data sets, amount of training data, ASR systems, combination of the approach with other systems, and other factors, it is not possible to compare performance directly. Nevertheless, we can look at performance in two ways. First, we can look at the relative error reduction that higher-level features contribute when combined with a baseline cepstral GMM system. Such information is provided at the end of the discussion of each feature type. Second, a within-site study allowing for direct comparisons of the performance of high-level systems is provided in Section 3. The analysis includes only a subset of feature types but uses state-of-the-art systems and recent NIST evaluation data.

2.1 Cepstral and Cepstral-Derived Features

Several approaches use the output of a word or phone recognizer to condition the extraction of cepstral features, thereby reducing variability associated with phonetic content. A review of some of these approaches is provided in [9]. Note that constraining the features to specific words essentially confers on text-independent speaker models some of the advantages of text-dependent speaker verification. The approach in [10] conditions a cepstral GMM on the identities of frequent words, based on recognizer word alignments. A variant conditions on syllables rather than words [11]. Another approach is to use multi-state HMMs as used in ASR as speaker models, thereby conditioning at the phone level but also capitalizing on a more detailed model of the sequential aspects of

Table 1. A Multidimensional Classification of Higher-Level Features in Speaker Recognition. *Ref.* = selected reference(s), *GMM* = Gaussian mixture model, *SVM* = support vector machine, *HMM* = hidden Markov model, *DTW* = dynamic time warping, *MLLR* = maximum likelihood linear regression, *unc.* = unconstrained, *LR* = likelihood ratio, *rec.* = recognition, *artic.* = articulatory, *freq.* = frequencies, *POS* = part of speech.

Feature Type	Feature Description	Time Span	ASR Used for		Model	Ref.
			Feature Extraction	Region Conditioning		
Cepstral	phone-conditioned cepstral models	frame	none	phones, classes	GMM, SVM	[9]
	text-conditioned GMMs	frame	none	words, syllables	GMM	[10, 11]
	phone HMMs	frame	phone, word	phone	HMM	[12, 13]
	whole-word models	longer	none	frequent word N-grams	word HMM	[14]
	DTW word models	longer	none	frequent word N-grams	template	[15, 16]
Cepstral-derived	MLLR transforms	frame	word, unc. phone	phone	SVM	[17]
Acoustic tokenization (“phonetic”)	phone N-gram freq.	longer	unc. phone	none	LR, SVM	[18–22]
	word-conditioned phone N-gram freq.	longer	unc. phone	frequent word N-grams	SVM	[23]
	conditioned pronunciation model	longer	unc. phone + word	phones from word rec.	LR	[24]
	conditioned pronunciation model	longer	unc. phone + artic.	phones from unc. phone rec.	LR	[25]
Prosodic	prosody dynamics	longer	none	none / phone	LR	[26, 27] / [27]
	DTW word-pitch models	longer	none	word	template	[27]
	interpause / conversation-level statistics	longer	word	none	GMM / LR	[28] / [29]
	word-constrained phone duration	longer	word	word	GMM	[30]
	phone-constrained state duration	longer	word	word	GMM	[30]
	syllable-based prosody sequence	longer	word	word	SVM	[31, 32] / [33]
Lexical	word N-grams	longer	word	none	LR, SVM	[34, 35]
Lexico-prosodic	duration-conditioned word N-grams	longer	word	none	SVM	[36]

speech, in contrast to the bag-of-frames model used in GMM-based approaches. HMM speaker modeling can be based on phone recognition [12] or word recognition [13]. A more recent variant [14] uses whole-word HMMs, thereby enabling even more detailed modeling; the HMMs represent not only words but frequent bigrams and trigrams as well. Whole words and phrases are also modeled by [15], but in a nonparametric fashion. The cepstra for a given phrase are aligned by using dynamic time warping to fit a standard length, after which the stacked cepstral feature vectors can be compared directly and a match score computed. In [16], dynamic time warping is used both to find frequent words and to score them against the speaker model of the word.

The MLLR approach [17] uses speaker-specific model adaptation transforms from a speech recognizer (either phone or word level) as features, modeled by a support vector machine (SVM). Instead of cepstral features, it uses the *difference* between speaker-adapted Gaussian means and corresponding speaker-independent means as features. This difference is expressed as the coefficients of an affine transform that rotates and shifts the speaker-independent model to obtain a speaker-dependent model, computed with maximum likelihood linear regression. Furthermore, the Gaussian models used in this approach are not unstructured GMMs but the detailed context-dependent phone models used in a speech recognizer, making the resulting features text independent. This has the advantage that features are text independent while being shared among all instances of a given phone, thus avoiding the data fragmentation implied by the conditioning on words. Transforms specific to different phone classes are combined for greater representational detail.

Cepstral models are usually the most accurate speaker recognition models when used on their own. State-of-the-art cepstral systems give about 4% to 5% equal error rate (EER) on the most recent NIST SRE test set when trained on 1 conversation side per speaker, and roughly 2% to 3% EER with 8 sides of training data, after intersession variability compensation. Small gains (about 10% to 15%) can be achieved by combining more than one state-of-the-art cepstral system. Systems using phone- or word-conditioned cepstral models typically are not much better than standard (unconditioned) cepstral models when used on their own. But they can provide substantial gains when combined with the latter, with reported improvements of up to 50% for 8-side training [10,14]. It is not yet known how such systems combine when multiple cepstral systems are available.

2.2 Acoustic Tokenization (“Phonetic”) Features

A large body of work, often referred to as “phonetic” recognition or modeling, employs unconstrained phone recognition essentially as a means by which to discretize the acoustic space and enable acoustic sequence modeling. (An alternative acoustic tokenization approach using GMM-generated events is described in [37].) Unconstrained-phone-based speaker models capture an assortment of speaker-dependent factors—including spectral characteristics, pronunciation idiosyncrasies, and lexical preferences—and can therefore be difficult to interpret. The basic approach obtains the top phone decoding hypothesis and then evaluates likelihood ratios of speaker-specific and generic (background) phone N-gram models [18]. Results can be improved by running several language-dependent or gender-dependent phone recognizers. The

phone N-gram distributions are modeled by bigram or trigram language models; refinements of this approach include the use of decision tree models for better smoothing [21] and the modeling of phone N-grams representing simultaneous outputs from multiple phone recognizers [22].

An important advance was the use of SVMs instead of likelihood models to model phone N-gram frequencies [19]. Improvements can also be obtained by modeling not just the top hypothesized phone sequence from the recognizer, but rather the expected phone N-gram frequencies extracted from phone recognition lattices [20]. In [23], lattice-based phone N-gram frequency modeling is combined with word conditioning. This approach is thus analogous to that used for the word-conditioned cepstral models discussed earlier. The phone N-grams occurring in specific words and frequent phrases are tallied and assembled into a more detailed feature vector that is modeled by SVMs.

A unique combination of phone- and word-based modeling is described in [24,38]. The output of an unconstrained phone recognizer is time-aligned with the phone sequence from a word recognizer, and the conditional probabilities of the former given the latter are modeled. Thus, this model captures phone-specific pronunciation realizations, albeit averaged over all words. An interesting variant aligns hypothesized articulatory features with the unconstrained phone recognition sequence [25].

Approaches based on unconstrained phone recognition show about 2 to 3 times the EER of the best cepstral systems, but can provide substantial gains when combined with them. Reported results show EER reductions of about 25% for 1-side and 44% for 8-side training [20]. Recent experiments with word-constrained phone N-gram methods also give promising results [23] for 8-side training. How systems such as [20] combine when multiple cepstral systems are available is less clear, since preliminary work did not find large gains, but further research is warranted.

2.3 Prosodic Features

Prosodic approaches attempt to capture speaker-specific variation in intonation, timing, and loudness. Because such features are suprasegmental (are not properties of single speech segments but extend over syllables and longer regions), they can provide complementary information to systems based on frame-level or phonetic features. One of the most studied features is speech fundamental frequency (or as perceived, pitch), which reflects vocal fold vibration rate and is affected by various physical properties of the speaker's vocal folds, including their size, mass, and stiffness [39]. Distributions of frame-level pitch values have been used in a number of studies [5,6,7,8]. Although they convey useful information about a speaker's distribution of pitch values, such statistics do not capture dynamic information about pitch contours and are thus not viewed as high-level here.

Dynamic variation in pitch operates at longer temporal spans and is used to convey not only message content (e.g., syntactic units, semantic focus) but also paralinguistic information. Modeling of prosody dynamics (which captures longer-range information and is thus included as higher-level) was used in early work on text-dependent speaker recognition [40]. In [26], a method is described for contour modeling for text-independent recognition. The speaker's pitch movements are modeled by fitting a piecewise linear model to the pitch track to obtain a stylized pitch contour. Parameters of the

stylized model are then used as statistical features for speaker verification. Variants are described in [27], which looks at rises and falls of the fitted pitch and energy values based on [26] and models the symbol sequence as a simple bigram. Additional integrated information includes rise and fall durations and phone context. Related work looks at piecewise linear fitting of pitch to help recover from performance degradations in speaker recognition for low-bit-rate coded speech [41]. Two interesting alternatives to using piecewise linear approximations of prosodic contours are proposed in [42], which uses wavelet analysis, and [43], which models sequences of quantized prosody symbols using latent semantic indexing.

An approach to prosodic modeling that is loosely analogous to the whole-word and DTW word modeling methods described for cepstral features is also explored in [27]. In this case, frequent words are matched for F0 contour, rather than for cepstral features. Thus, like its cepstral counterpart, this approach uses no linguistic information for feature extraction but conditions on word-level information from ASR.

A small number of studies have looked at linguistically conditioned duration, pitch, and energy statistics in longer spans of speech. In [28], prosody statistics are computed for units between pauses. The interpause unit is but one example of a larger world of features that could be defined at different temporal spans; the focus is on modeling approaches and modifying GMMs to cope with undefined or inherently missing features (such as pitch, which is missing during unvoiced regions). In [29,38], statistics are computed over an entire conversation side, and distances of each conversation-level feature vector from vectors for target versus impostor speakers are compared using log likelihood ratios. Earlier work on conversation-level statistics [44] includes lexical features such as disfluency rates. Finally, [29,38] explore sequential modeling of “turn”-level prosodic feature statistics. Because turns were automatically inferred from pause and speaker change information, they bear some similarity to the interpause extraction units used in [28], although features and models differ.

Two approaches that use ASR for *conditioning* (as opposed to merely for extraction) are described in [30]. One method, the phone-in-word-duration GMM, models the durations of phones within specific words. Unlike the previous prosodic approaches, it employs ASR for conditioning because it compares durations on a per-word basis. A second method, the state-in-phone-duration GMM, uses the durations (numbers of frames) of the three states in phone HMMs as features, and phones are used for conditioning. In each case, the durations for different positions form a feature vector and are modeled in the adapted-GMM framework used for standard cepstral GMM systems.

A recent method models syllable-based prosodic feature sequences [31,32]. In contrast to interpause-based and conversation-level prosody statistics, this approach uses smaller time units (resulting in more features) and models sequential information. Syllables are automatically inferred from ASR output, and a variety of F0, duration, and energy values are extracted from each syllable. In the unconstrained model, features are extracted for all syllable N-grams in a conversation side. To turn the variable-length sequences of feature vectors into a single conversation-level vector, a set of GMM models is created for each feature sequence (sequence of syllables and pauses). Given a sample, the posterior probabilities of each Gaussian in each GMM are computed and concate-

nated into the final conversation-level feature vector. These features are provided to an SVM to perform regression on the class labels.

A further refinement is a conditioned version of the syllable-based prosodic feature sequence SVM just described. In this approach, detailed in [33], lexical, part-of-speech, and pause information is used to condition extraction of the same features to specific locations believed to behave similarly prosodically. The goal of the conditioning is thus conceptually similar to that for word-constrained cepstral features [10], but for prosodic rather than phonetic similarity. Note that in the case of prosodic features, phonetic content can be normalized out, allowing multiple words (such as lists of backchannels) per wordlist, increasing robustness. Interestingly, although the unconditioned and conditioned systems use the same features and differ only in conditioning, there is a considerable gain by combining them at the feature level in a single SVM.

Prosodic systems comprise a wide range of approaches and results, making it difficult to summarize performance. The best-performing individual system appears to be a feature-level combination of the unconditioned and conditioned syllable-based prosodic sequence model. Combination of this prosodic system with a cepstral system reduces EER by about 20% and 40% for 1- and 8-side training, respectively. An advantage of this system is that it offers significant complementary information when multiple cepstral systems are present (see Section 3).

2.4 Lexical Features

A speaker's distribution of word sequences is historically one of the earliest types of higher-level features explored for speaker recognition, with roots in the analogous task of author attribution in the text classification domain. Early work using lexical N-gram statistics to discriminate speakers is described in [45]. The approach did not produce a significant gain at the time, presumably because of the brief training and test samples used in task definitions at the time. With the advent of the extended data condition, however, it was found that rates of idiosyncratic word N-grams (for example, "how shall") could be used to help discriminate speakers [34]. The study in [34] used likelihood ratios; in [35], the relative frequencies of frequent word unigrams, bigrams, and trigrams are obtained and assembled into a feature vector that is modeled by SVMs.

More recently, the approach has been extended to encode the duration (slow/fast) of frequent word types as part of the N-gram frequencies [36]. This technique represents a true *hybrid* model of lexical and prosodic features, since it explicitly models both N-gram frequencies and word durations. It thereby simultaneously captures lexical, pronunciation, and prosodic characteristics of the speaker. An interesting further line of research in this area is to postprocess lexical features with latent semantic analysis, so that by grouping words similar in semantic space, one may increase the robustness of estimates for less-frequent words [46].

In terms of performance results, word N-gram modeling yields about 25% EER on 1-side and 10% EER on 8-side training for recent NIST SRE data. Despite the poor performance when used individually, combination with a state-of-the-art cepstral system on recent SRE data improves the overall system by about 15-20% for 8-side training.

Table 2. Data Sets Used in Experiments

Test set	SRE-06 Common Condition	
	1-side	8-sides
Training		
Conversation sides	3,209	6,556
Models	517	483
Trials	24,013	17,547

3 Performance in a Recent System

The preceding section provided an overview of higher-level features, with minimal discussion of performance because of difficulties in comparing across studies. It also did not address the important question of how systems combine with others, beyond simple comparisons with a baseline cepstral system. To this end, it is useful to look at performance on a recent corpus and task. A set of useful results is available from SRI International, which has in-house efforts to develop both systems based on frame-level cepstra and systems using high-level features.

3.1 Task and Data

We will look at the task of speaker verification on the 2006 NIST SRE evaluation data [4,47]. Results are for the primary subset (the “Common Condition”), which consists of English-only conversations. Test data consist of 1 conversation side. Because high-level features are defined at a longer time scale than are frame-level features, it is interesting to ask how high-level systems perform as a function of the amount of training data per speaker. We will thus look at two training conditions: one with 1 conversation side per speaker, the other with 8 conversation sides per speaker (each with a different conversational partner). Data set statistics are provided in Table 2.

Background training data consisted of 1,553 conversation sides from separate data collections (Switchboard-II and Fisher). Background data did not share any speakers with the data in the test set.

3.2 ASR system

All speech was processed by SRI’s speech recognition system. None of the test or background data were used in training or tuning of the recognition system. The system is a fast, two-stage version of SRI’s conversational telephone speech (CTS) system, as originally developed for the 2003 DARPA Rich Transcription evaluation [48] and later modified for the NIST 2004 speaker recognition evaluation [35]. It performs a first decoding using Mel frequency cepstral coefficient (MFCC) acoustic models and a bigram language model (LM), generating lattices that are then rescored with a higher-order LM. The resulting hypotheses are used to adapt a second set of models based on perceptual linear prediction (PLP) acoustic features. The adapted models are used in a second decoding pass that is constrained by trigram lattices, which generates N-best lists. These

are then rescored by a 4-gram LM and by prosodic models to arrive at the final word hypotheses.

3.3 Session Variability Compensation and TNORM

The SRI system employs techniques for reducing the effect of within-speaker variability associated with the speaking context or environment, rather than the speaker. In the speaker verification community, techniques are often referred to as “session variability” compensation techniques, because they were applied to handle the variability found when the same talker speaks in different conversations. As such, the techniques may capture differences in handsets, background noise, topic of conversation, emotion, speaker health, and so on. The idea is to estimate from data the feature space directions along which intersession variability lies, and then project the features onto the remaining directions. The techniques used are factor analysis for GMM-based models [49] and nuisance attribute projection (NAP) for SVM-based models [50]. An interesting aspect of these approaches is that although they were developed for systems based on cepstral features, they also significantly benefit the SRI prosodic SVM system, with error reductions of over 20% for the 8-side condition. Various systems also make use of TNORM [51], a score normalization technique.

3.4 Systems

While the set of SRI systems does not cover all system types reviewed in Section 2, it has the advantage of including five higher-level and three lower-level systems based on frame-level cepstral features. Where applied, systems used the same ASR output and similar methods for session variability compensation [49,50] and score normalization (TNORM [51]). Systems are roughly, albeit not directly, comparable.¹

Higher-level systems. The higher-level systems represent five approaches from Table 1: (1) the MLLR system based on word recognition (Section 2.1), (2) a combination of constrained and unconstrained syllable-based prosodic feature sequences in a single SVM (Section 2.3), (3) the word-constrained phone duration system (Section 2.3), (4) the phone-constrained state duration system (Section 2.3), and (5) the duration-conditioned word N-gram (Section 2.4). No phonetic system is represented, because earlier work showed little gain from combining such systems with multiple frame-level cepstral systems. This issue should certainly be revisited, however, given the many updates to various approaches since that time. Another missing feature type is text-conditioned cepstral systems, which is obviously important to explore as well.

Frame-level cepstral systems. In addition to the MLLR system, three other systems model frame-level cepstral features: a cepstral GMM, a cepstral SVM, and a Gaussian

¹ For practical reasons, TNORM was applied for the cepstral SVM, duration, and word N-gram systems, and session variability compensation was applied for the cepstral GMM, MLLR, Gaussian supervector, and prosodic sequence systems. Although the latter technique generally produces larger gains, direct comparisons of systems without normalizations indicate that the ordering of systems by individual performance does not depend on normalization.

supervector SVM. The cepstral GMM system is a generative model of the cepstral feature distribution in the form of a mixture of Gaussian densities [1]. It is trained on a large background set of speakers to cover the entire observed distribution of cepstral features. Frames are treated as unordered, independent samples, discarding longer-term sequence information. Given target speaker training data, the GMM is then adapted by reestimating the Gaussian means on the target speaker data (with a mixture of the background data for smoothing). This results in the target speaker GMM. The system computes the likelihood ratio that the test sample was generated by the target speaker model versus the background model, and accepts the sample if the score exceeds an empirically set threshold.

The cepstral SVM system computes polynomials of the cepstral features and averages them over the entire conversation [52]. For example, one feature might be the average product of the first cepstral coefficient times the square of the second. A feature vector consisting of a large number of these polynomial features characterizes the joint distribution of cepstral features. These feature vectors are then modeled by SVMs. SVMs are trained using a large population of diverse (background) speakers as negative samples and a small set of target speaker instances as positive samples. In testing, a feature vector extracted from the test data is classified by the SVM, and the signed distance from the decision hyperplane is used as a score to be thresholded.

The Gaussian supervector SVM is based on the adapted target speaker GMM mentioned above [53]. Instead of modeling the cepstral features directly, it uses the adapted Gaussian means as features, stringing them together into a long “supervector.” The supervector is then modeled as an SVM classifier input, similar to the cepstral SVM.

3.5 Results

Performance results for individual systems are summarized in Table 3. As expected from the review in the previous section, systems based on frame-level cepstral or cepstral-derived features show higher accuracy than longer-range systems. Within the set of cepstral-based systems, the MLLR system has best performance, presumably because it takes advantage of linguistic information from ASR. Of the longer-range systems, the conditioned syllable-based prosody sequence system is the most successful, with less than half the error rate of other longer-range systems for the 8-side condition.

As noted earlier, however, the importance of higher-level systems for such tasks is not individual performance but how well they complement standard systems. To answer that question, we examine results for various system combinations. Individual system scores are combined using an SVM with a linear inner product kernel; the combiner is trained using scores for separate data (from the NIST 2005 evaluation). We first look at how well each system combines with the cepstral GMM system. Results for 1- and 8-side training are shown in Figure 1, respectively. For reference, the cepstral GMM system alone and the MLLR system (best single system) alone are also indicated.

As shown, all combinations (triangle and square symbols) improve performance over the baseline cepstral GMM alone, in most cases by a significant degree. For both training conditions, combinations with other frame-level cepstral systems (squares) are better than combinations with some higher-level systems (triangles) and worse than others. The MLLR system alone performs better than the combination of the cepstral GMM

Table 3. Individual System Results for Eight Systems. “(H)” denotes higher-level systems. Performance is given as both equal error rate (EER) and the detection cost function (DCF) used by NIST [4].

System (Feature Type and Model)	Time Range	ASR	EER/DCFx10 1-side	EER/DCFx10 8-side
Cepstral GMM	frame	no	4.75 / 0.216	2.79 / 0.107
Cepstral SVM	frame	no	5.07 / 0.242	2.33 / 0.093
Gaussian supervector SVM	frame	no	4.15 / 0.198	3.24 / 0.164
(H) MLLR SVM	frame	yes	4.00 / 0.197	2.14 / 0.073
(H) State-in-phone-duration GMM	longer	yes	16.02 / 0.705	8.07 / 0.423
(H) Phone-in-word-duration GMM	longer	yes	22.22 / 0.874	9.30 / 0.420
(H) Syllable-based prosody sequence SVM	longer	yes	10.41 / 0.461	3.74 / 0.162
(H) Duration-conditioned word N-gram SVM	longer	yes	23.46 / 0.815	9.95 / 0.446

with either the cepstral SVM or the supervector system with 8 sides of training data, but this is not the case for the 1-side training condition—demonstrating that higher-level systems add more value as training data increases. For the 8-side condition, the best two combinations with the cepstral GMM are clearly the MLLR system and the prosodic feature sequence system, both higher-level systems. In this condition even the word N-gram system, which performs poorly on its own, combines about as well with the baseline as does the cepstral SVM system.

To understand how more than two systems combine, we can look to Figure 2. This figure shows which systems are selected when one optimizes an N-way system combination for best performance (in DCF terms). Since there are 8 SRI systems, N ranges from 1 to 8. If the selection pattern is monotonic—i.e., if the list of systems for N+1 includes all systems from the list for N—then the order in which systems are progressively added can be construed as reflecting system importance in the combination. As shown in Figure 2, with one exception, the selection order is monotonic, thus providing information about which systems make the most contribution to the overall result.

We can extract a number of useful observations by comparing Table 3 and Figure 2. As already mentioned, by themselves noncepstral systems perform less well than cepstral systems. Among the systems using higher-level information, the more acoustic information a system models, the better it tends to perform on its own (MLLR-SVM > prosody sequence model > duration > word-duration N-grams), which is not surprising. The MLLR-SVM system, which takes advantage of both high-level constraints and frame-level acoustic information, is also the best single system overall.

What is striking is the finding that of the four systems using frame-level cepstral features (MLLR, cepstral GMM, cepstral SVM, and supervector SVM), only two are actually useful for a particular training condition (1 or 8 sides). Within each condition, only two such systems appear at the left side of the figures; the other two appear at the

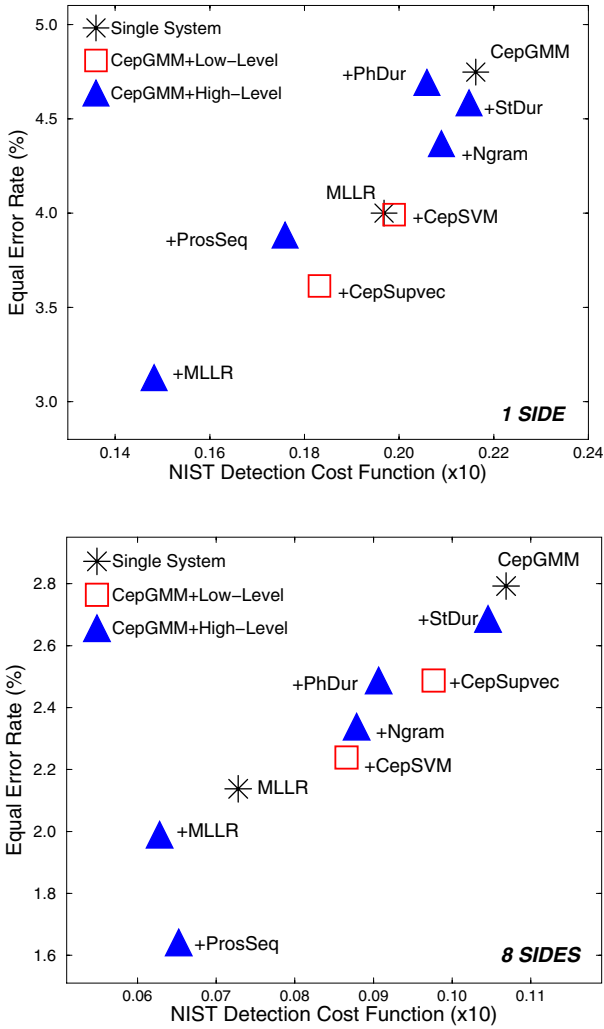


Fig. 1. Two-way combination results (system + cepstral GMM) by type of system. CepGMM = cepstral GMM, CepSVM = cepstral SVM, CepSupvec = Gaussian supervector SVM, MLLR = MLLR SVM, ProsSeq = syllable-based prosodic feature sequence SVM, PhDur = phone-in-word duration GMM, StDur = state-in-phone duration GMM, Ngram = duration-conditioned word N-gram SVM.

right and do not add any real performance improvements. In fact, they can even degrade performance (as can any system added late in the combination) because of overfitting in combiner training. Note that different cepstral systems are useful for different amounts of training data. Complementary information comes from higher-level

	MILLR SVM (H)	Supervector SVM	Prosody Sequence SVM (H)	Phone-in-Word Duration GMM (H)	State-in-Phone Duration GMM (H)	Word+Duration N-gram SVM (H)	Cepstral GMM	Cepstral SVM	%EER	DCF (x10)
1 Best									4.00	0.197
2 Best									3.13	0.148
3 Best									2.86	0.139
4 Best									2.86	0.137
5 Best									2.80	0.136
6 Best									2.86	0.140
7 Best									2.64	0.141
8 Best									2.59	0.144

	MILLR SVM (H)	Prosody Sequence SVM (H)	Cepstral SVM	Word+Duration N-gram SVM (H)	State-in-Phone Duration GMM (H)	Phone-in-Word Duration GMM (H)	Supervector SVM	Cepstral GMM	%EER	DCF (x10)
1 Best									2.13	0.0728
2 Best									1.74	0.0561
3 Best									1.59	0.0501
4 Best									1.59	0.0488
5 Best									1.59	0.0478
6 Best									1.54	0.0483
7 Best									1.64	0.0480
8 Best									1.64	0.0485

Fig. 2. Results for the best N-way combination of systems for the 1-side (top) and 8-side (bottom) training conditions. Filled boxes indicate which systems were selected; corresponding performance is given in both equal error rate (EER) and detection cost function (DCF).

systems, including systems that perform modestly when used alone. Such systems become increasingly useful as training data increases from 1 to 8 sides. It would thus be interesting to look at even larger amounts of speaker training data (more than 20 minutes), to see at which point the higher-level models begin to level off in their performance. Overall, these findings point to a nonobvious strategy for future overall system development. Because systems based on cepstral features tend to be highly correlated, the exploration of complementary systems based on higher-level features should become only more important as lower-level systems continue to improve.

4 Conclusions and Implications for Speaker Classification

Despite the dominance of GMM systems based on frame-level cepstral features, we have seen that higher-level features provide significant complementary information for speaker identification. Higher-level features are increasingly useful as training data increases, and we have not yet witnessed the point at which they level off in relative contribution to performance. Furthermore, because certain higher-level features are inherently more invariant to channel and noise characteristics than are spectral features, they offer the possibility of additional robustness for speaker recognition under degraded acoustic conditions.

For the wider area of speaker classification, higher-level features in speech provide far more information about a talker than only his or her identity. For example, research in [54] reveals that speaker age is reflected not only in acoustic features but also in temporal features such as phone durations. Features based on phone-level, lexical, or prosodic information are correlated with language and dialect classification [55], emotion classification [56], deception detection [57], and perceived charisma [58], as well as a host of other health-related, cognitive, and sociolinguistic factors. Given sufficient data labeled for such characteristics, one might apply some of the features and techniques described here, substituting the new class of interest for speaker identity. Since we know that higher-level features are quite successful at classifying individual speakers, an additional interesting research area in classifying speaker characteristics, rather than individual speakers, would be to apply nuisance attribute projection [59] to project out the variability that is speaker-related. In this way, one might achieve sharper models that can assist speaker classification in other domains.

Acknowledgments

The author thanks Christian Müller; SRI colleagues Andreas Stolcke, Sachin Kajarekar, Luciana Ferrer, Kemal Sönmez, Gokhan Tür, and Martin Graciarena; and SRI visitor Robbie Vogt. This work was supported by NSF Contract IIS-0544682. The views herein are those of the author and do not represent the views of the funding agency.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10, 181–202 (2000)
2. Sturim, D.E., Campbell, W.M., Reynolds, D.A.: Classification Methods for Speaker Recognition. In: Müller, C. (ed.) *Speaker Classification I. LNCS (LNAI)*, vol. 4343, Springer, Heidelberg (2007)
3. Markowitz, J.: The Many Roles of Speaker Classification in Speaker Verification and Identification. In: Müller, C. (ed.) *Speaker Classification I. LNCS(LNAI)*, vol. 4343, Springer, Heidelberg (2007)
4. Martin, A.F.: Evaluations of Automatic Speaker Classification Systems. In: Müller, C. (ed.) *Speaker Classification I. LNCS(LNAI)*, vol. 4343, Springer, Heidelberg (2007)
5. Carey, M., Parris, E., Lloyd-Thomas, H., Bennett, S.: Robust prosodic features for speaker identification. In: Bunnell, H.T., Idsardi, W. (eds.) *Proc. ICSLP. Philadelphia*, vol. 3, pp. 1800–1803 (1996)

6. Sönmez, M.K., Heck, L., Weintraub, M., Shriberg, E.: A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (eds.) Proc. EUROSPEECH, Rhodes, Greece, pp. 1391–1394 (1997)
7. Arcienega, M., Drygajlo, A.: Pitch-Dependent GMMs for Text-Independent Speaker Recognition Systems. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01 – Interspeech), Aalborg, Denmark, pp. 2821–2825 (2001)
8. Kinnunen, T., Gonzalez-Hautamaki, R.: Long-Term F0 Modeling for Text-Independent Speaker Recognition. In: Proceedings of the 10th International Conference Speech and Computer (SPECOM), Patras, Greece, pp. 567–570 (2005)
9. Park, A., Hazen, T.J.: ASR Dependent Techniques for Speaker Identification. In: Hansen, J.H.L., Pellom, B. (eds.) Proc. ICSLP, Denver, pp. 1337–1340 (2002)
10. Sturim, D.E., Reynolds, D.A., Dunn, R.B., Quatieri, T.F.: Speaker Verification Using Text-Constrained Gaussian Mixture Models. In: Proc. ICASSP. vol. 1., Orlando, pp. 677–680 (2002)
11. Baker, B., Vogt, R., Sridharan, S.: Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal, pp. 2429–2432 (2005)
12. Gauvain, J.L., Lamel, L.F., Prouts, B.: Experiments with Speaker Verification Over the Telephone. In: Pardo, J.M., Enríquez, E., Ortega, J., Ferreiros, J., Macías, J., Valverde, F.J. (eds.) Proc. EUROSPEECH, Madrid (1995)
13. Newman, M., Gillick, L., Ito, Y., McAllaster, D., Peskin, B.: Speaker Verification Through Large Vocabulary Continuous Speech Recognition. In: Bunnell, H.T., Idsardi, W. (eds.) Proc. ICSLP. vol. 4, Philadelphia, pp. 2419–2422 (1996)
14. Boakye, K., Peskin, B.: Text-Constrained Speaker Recognition on a Text-Independent Task. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain (2004)
15. Gillick, D., Stafford, S., Peskin, B.: Speaker Detection without Models. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 757–760 (2005)
16. Aronowitz, H., Burshtein, D., Amir, A.: Text Independent Speaker Recognition Using Speaker Dependent Word Spotting. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea, pp. 1789–1792 (2004)
17. Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A.: MLLR Transforms as Features in Speaker Recognition. In: Proc. Interspeech, Lisbon, pp. 2425–2428 (2005)
18. Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J., Hernandez-Cordero, J.: Gender-Dependent Phonetic Refraction for Speaker Recognition. In: Proc. ICASSP. Orlando, vol. 1, pp. 149–152 (2002)
19. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A., Leek, T.R.: Phonetic Speaker Recognition with Support Vector Machines. *Advances in Neural Information Processing Systems* 16, 1377–1384 (2004)
20. Hatch, A.O., Peskin, B., Stolcke, A.: Improved Phonetic Speaker Recognition Using Lattice Decoding. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 169–172 (2005)
21. Navrátil, J., Jin, Q., Andrews, W.D., Campbell, J.P.: Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 796–799 (2003)
22. Jin, Q., Navrátil, J., Reynolds, D.A., Campbell, J.P., Andrews, W.D., Abramson, J.S.: Combining Cross-Stream and Time Dimension in Phonetic Speaker Recognition. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 800–803 (2003)
23. Lei, H., Mirghafori, N.: Word-Conditioned Phone N-Grams for Speaker Recognition. In: Proc. ICASSP, Honolulu (2007)

24. Klusáček, D., Navrátil, J., Reynolds, D.A., Campbell, J.P.: Conditional Pronunciation Modeling in Speaker Detection. In: Proc. ICASSP Hong Kong, vol. 4, pp. 804–807 (2003)
25. Ka-Leung, Y., Man-Mak, W., Kung, S.Y.K.: Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea, pp. 2597–2600 (2004)
26. Sönmez, K., Shriberg, E., Heck, L., Weintraub, M.: Modeling Dynamic Prosodic Variation for Speaker Verification. In: Mannell, R.H., Robert-Ribes, J. (eds.) Proc. ICSLP. vol. 7, pp. 3189–3192, Australian Speech Science and Technology Association, Sydney (1998)
27. Adami, A.G., Mihaescu, R., Reynolds, D.A., Godfrey, J.J.: Modeling Prosodic Dynamics for Speaker Recognition. In: Proc. ICASSP. Hong Kong, vol. 4, pp. 788–791 (2003)
28. Kajarekar, S., Ferrer, L., Sönmez, K., Zheng, J., Shriberg, E., Stolcke, A.: Modeling NERFs for Speaker Recognition. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain, pp. 51–56 (2004)
29. Peskin, B., Navrátil, J., Abramson, J., Jones, D., Klusáček, D., Reynolds, D.A., Xiang, B.: Using Prosodic And Conversational Features for High Performance Speaker Recognition: Report From JHU WS'02. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong, pp. 792–795 (2003)
30. Ferrer, L., Bratt, H., Gadde, V.R.R., Kajarekar, S., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A.: Modeling Duration Patterns for Speaker Recognition. In: Proc. EUROSPEECH, Geneva, pp. 2017–2020 (2003)
31. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A.: Modeling prosodic feature sequences for speaker recognition. *Speech Communication*. (Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation) 46(3-4), 455–472 (2005)
32. Ferrer, L., Shriberg, E., Kajarekar, S., Sönmez, K.: Parameterization of Prosodic Feature Distributions for SVM Modeling in Speaker Recognition. In: Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07), Honolulu, Hawaii (2007)
33. Shriberg, E., Ferrer, L.: A Text-Constrained Prosodic System for Speaker Verification. In: Proceedings of Interspeech, Antwerp, Belgium (2007)
34. Doddington, G.: Speaker Recognition Based on Idiolectal Differences Between Speakers. In: Dalsgaard, P., Lindberg, B., Benner, H., Tan, Z. (eds.) Proc. EUROSPEECH, Aalborg, Denmark, pp. 2521–2524 (2001)
35. Kajarekar, S.S., Ferrer, L., Shriberg, E., Sonmez, K., Stolcke, A., Venkataraman, A., Zheng, J.: SRI's 2004, NIST Speaker Recognition Evaluation System. In: Proc. ICASSP. Philadelphia, vol. 1, pp. 173–176 (2005)
36. Tür, G., Shriberg, E., Stolcke, A., Kajarekar, S.: Duration and Pronunciation Conditioned Lexical Modeling for Speaker Verification. In: Proceedings of Interspeech, Antwerp, Belgium (2007)
37. Scheffer, N., Bonastre, J.F.: Speaker Detection using Acoustic Event Sequences. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal (2005)
38. Reynolds, D., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A., Jin, Q., Klusáček, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B.: The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong (2003)
39. Titze, I.: Principles of Voice Production. Prentice Hall, Englewood Cliffs (1994)
40. Atal, B.: Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America* 52(6), 1687–1697 (1972)

41. Chen, S.H., Wang, H.C.: Improvement of Speaker Recognition by Combining Residual and Prosodic Features with Acoustic Features. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada (2004)
42. Chen, J., Dai, B., Sun, J.: Prosodic Features Based on Wavelet Analysis for Speaker Verification. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal, pp. 3093–3096 (2005)
43. Chen, Z.H., Liao, Y.F.L., Juang, Y.T.: Eigen-Prosody Analysis for Robust Speaker Recognition under Mismatch Handset Environment. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea (2004)
44. Weber, F., Manganaro, L., Peskin, B., Shriberg, E.: Using Prosodic and Lexical Information for Speaker Identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), Orlando, Florida (2002)
45. Heck, L.: Integrating High-Level Information for Robust Speaker Recognition (2002), <http://www.cisp.jhu.edu/ws2002/groups/supersid/>
46. Nayeemulla Khan, A., Yegnanarayana, B.: Latent Semantic Analysis for Speaker Recognition. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP '04), Jeju Island, South Korea (2004)
47. Martin, A., Miller, D., Przybocki, M., Campbell, J., Nakasone, H.: Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004. In: Proceedings 4th International Conference on Language Resources and Evaluation, Lisbon, pp. 587–590 (2004)
48. Stolcke, A., Franco, H., Gadde, R., Graciarana, M., Precoda, K., Venkataraman, A., Vergyri, D., Wang, W., Zheng, J., Huang, Y., Peskin, B., Bulyko, I., Ostendorf, M., Kirchhoff, K.: Speech-to-text Research at SRI-ICSI-UW. In: DARPA RT-03 Workshop, Boston (2003)
49. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Factor Analysis Simplified. In: Proc. ICASSP. vol. 1, pp. 637–640 (2005)
50. Solomonoff, A., Campbell, W.M., Boardman, I.: Advances in Channel Compensation for SVM Speaker Recognition. In: Proc. ICASSP, Philadelphia, vol. 1, pp. 629–632 (2005)
51. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing* 10(1-3), 42–54 (2000)
52. Campbell, W.M.: Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. In: Proc. ICASSP, Orlando, vol. 1, pp. 161–164 (2002)
53. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support Vector Machines Using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters* 13(5), 308–311 (2006)
54. Schötz, S., Müller, C.: A Study of Acoustic Correlates of Speaker Age. In: Müller, C. (ed.) *Speaker Classification II. LNCS(LNAI)*, vol. 4441, Springer, Heidelberg (2007)
55. Schultz, T.: Speaker Characteristics. In: Müller, C. (ed.) *Speaker Classification I. LNCS(LNAI)*, vol. 4343, Springer, Heidelberg (2007)
56. Devillers, L., Vidrascu, L.: Real-life Emotion Recognition in Speech. In: Müller, C. (ed.) *Speaker Classification II. LNCS(LNAI)*, vol. 4441, Springer, Heidelberg (2007)
57. Graciarana, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., Kajarekar, S.: Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection. In: Proc. ICASSP, vol. 1, pp. 1033–1036 (2006)
58. Rosenberg, A., Hirschberg, J.: Acoustic/Prosodic Correlates of Charismatic Speech. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05 – Interspeech), Lisbon, Portugal (2005)
59. Solomonoff, A., Quillen, C., Boardman, I.: Channel Compensation for SVM Speaker Recognition. In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain (2004)