The Case for Automatic Higher-Level Features in Forensic Speaker Recognition

Elizabeth Shriberg *Andreas Stolcke*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, U.S.A
{ees,stolcke}@speech.sri.com

Abstract

Approaches from standard automatic speaker recognition, which rely on cepstral features, suffer the problem of lack of interpretability for forensic applications. But the growing practice of using "higher-level" features in automatic systems offers promise in this regard. We provide an overview of automatic higher-level systems and discuss potential advantages, as well as issues, for their use in the forensic context.

Index Terms: speaker recognition, higher-level features, forensics

1. Introduction

Recent overview papers [1, 2] compare and contrast traditional forensic speaker recognition with automatic speaker verification systems. In doing so, the automatic systems are exemplified by approaches that are based on cepstral modeling, i.e., on low-level, short-term acoustic features of the speaker's speech. While such approaches are certainly still the standard, they alone by no means represent the state of the art in automatic speaker recognition. Work by several teams [3, 4] has shown that higher-level features based on long-term, often linguistically motivated units can significantly improve speaker recognition performance. Examples of higher-level features include phonetic, prosodic, and lexical observations, as distilled from automatic recognition output and other measurements (such as pitch tracks). Improved accuracy alone should motivate their consideration for forensic speaker recognition.

Furthermore, we would argue that higher-level features have properties that make them especially attractive in forensic applications. Higher-level features are often easier to interpret than low-level ones, and are in some cases directly related to features used by traditional forensic analysis. Higher-level features can also benefit from intrinsic robustness to acoustic mismatch between speech samples, a major problem for accurately estimating the likelihood ratios required for forensic use [2], and have other desirable properties.

To provide an idea of the range of available techniques, we start with an overview and categorization of higher-level modeling techniques for speaker recognition, based on [5]. We then make the case for developing these or similar techniques for the forensic setting, by discussing specific advantages they afford over standard, low-level speaker modeling methods.

2. Higher-Level Features in Automatic Systems

A recent survey of research on higher-level features used in automatic text-independent speaker recognition [5] considered as "higher level" any feature that involves either *linguistic information* (defined as requiring phone- or word-level automatic speech recognition [ASR],) or longer-range information (longer than the frame-level information used in cepstral-based systems). To clarify how the different approaches use higher-level information, [5] specified not only the type of feature used, but also three other factors:

1. Temporal span of the feature
2. Level of ASR used for *feature extraction*
3. Level of ASR used for *region conditioning*

as indicated in Table 1. 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* refers 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. A step further in the direction of linguistic constraints involves imposing phonotactic constraints obtained from a phone N-gram language model. This approach favors 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* refers 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.

Table 1: Multidimensional classification of higher-level features in automatic speaker recognition, adapted from [5]. *DTW* = dynamic time warping, *unc.* = unconstrained, *rec.* = recognition, *artic.* = articulatory, *freq.* = frequencies, *POS* = part of speech.

Feature Type	Feature Description	Time Span	ASR Used for		Selected References
			Feature Extraction	Region Conditioning	
Cepstral	phone-conditioned cepstral models	frame	none	phones, classes	[6]
	text-conditioned GMMs	frame	none	words, syllables	[7, 8]
	phone HMMs	frame	phone, word	phone	[9, 10]
	whole-word models	longer	none	frequent word N-grams	[11]
	DTW word models	longer	none	frequent word N-grams	[12, 13]
Cepstral-derived	MLLR transforms	frame	word, unc. phone	phone	[14]
Acoustic tokenization ("phonetic")	phone N-gram freq.	longer	unc. phone	none	[15, 16]
	word-conditioned phone N-gram freq.	longer	unc. phone	frequent word N-grams	[17]
	conditioned pronunciation model	longer	unc. phone + word	phones from word rec.	[18]
	conditioned pronunciation model	longer	unc. phone + artic.	phones from unc. phone rec.	[19]
Prosodic	prosody dynamics	longer	none	none / phone	[20, 21] / [22]
	DTW word-pitch models	longer	none	word	[22]
	interpause / conversation-level statistics	longer	word	none	[23] / [24]
	word-constrained phone duration	longer	word	word	[25]
	phone-constrained state duration	longer	word	phone	[25]
	syllable-based prosody sequence	longer	word	none / words, POS	[26] / [27]
Lexical	word N-grams	longer	word	none	[28, 29]
Lexico-prosodic	duration-conditioned word N-grams	longer	word	none	[30]

2.1. Cepstral and cepstral-derived features

Approaches based on cepstral features 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 [6]. 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 [7] conditions a cepstral Gaussian mixture model (GMM) on the identities of frequent words, based on recognizer word alignments. A variant conditions on syllables rather than words [8]. A more recent variant [11] 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 [12], but in a nonparametric fashion.

The maximum likelihood linear regression (MLLR) approach [14] 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. 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.

2.2. Acoustic tokenization 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. 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 [15]. Results can be improved by running several language-dependent or gender-dependent phone recognizers. Note that, due to lack of phonotactic constraints, these recognizers produce phone sequences that will not match dictionary pronunciations. But, it is precisely through these mismatches that the system can learn that particular acoustic tendencies are correlated with particular speakers.

An important advance was the use of SVMs instead of likelihood models to model phone N-gram frequencies [16]. In [17], 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. A unique combination of phone- and word-based modeling is described in [18, 3]. 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 averaged phone-specific pronunciation realizations

2.3. Prosodic features

Prosodic features attempt to capture speaker-specific variation in pitch, duration, and energy patterns. Early work on using prosody for text-independent speaker recognition is described in [20]. 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 [22], which looks at rises and falls of the fitted pitch and energy values as well. More recent work [21] employs polynomial fits and factor analysis to characterize a speaker's prosodic dynamics.

Several studies have looked at linguistically conditioned duration, pitch, and energy statistics in longer spans of speech. In [23], 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 missing features (such as pitch, which is missing during unvoiced regions). In [24, 3], 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.

Two prosodic approaches that use ASR for *conditioning* (as opposed to merely for extraction) are described in [25]. 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 of the three states in phone HMMs as features, and phones are used for conditioning.

A recent method models syllable-based prosodic feature sequences [26, 31]. 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 per syllable. In an unconstrained version, features are extracted for all syllable N-grams in a conversation side. In a word-constrained version [27], lexical, part-of-speech, and pause information is used to condition feature extraction s to specific locations believed to behave similarly prosodically.

2.4. Lexical and lexico-prosodic features

A speaker's distribution of word sequences is historically one of the earliest types of higher-level features explored for automatic speaker recognition. Such work uses lexical N-gram statistics to discriminate speakers, modeled with likelihood ratios or 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 [30]. 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.

3. The Case for Automatic Higher-Level Features in Forensics

Having provided an overview of higher-level features from work in automatic speaker verification, we now discuss potential benefits and issues for forensic applications.

3.1. Discriminative power

One obvious reason to use more, and specifically higher-level, features and models for forensic applications is that they improve recognition accuracy. Several research teams have shown in their NIST evaluation systems and elsewhere that combining multiple systems reduces detection costs. Our own analysis of the contributions of the various components of such a complex system has shown that higher-level features, while not giving the best recognition accuracies by themselves, *add* the most information to a traditional low-level speaker recognition system. For example, a syllable-based prosodic system was the best choice for a two-way combined system in which the other system was a high-performance acoustic model [4].

Therefore, given standard techniques for calibration and mapping of system scores to likelihood ratios [32], combined low- and high-level speaker models should reduce the expected probabilities of error for a range of priors. This, in itself, should constitute the strongest argument for higher-level techniques among forensic practitioners.

3.2. Interpretability and acceptance

For use in court, it is desirable that measurements and models used for technical speaker verification be interpretable to a non-expert (e.g., so that it can be explained to a jury) [1]. Many of the higher-level features described earlier are more interpretable than cepstral features, either because they are inherently more accessible to perception (lexical features, pitch, durational features) or more easily visualized (cepstral features are hard to visualize given their high dimensionality). High-level features might also be more acceptable to legal and forensic practitioners because of their overlap with traditional methods. For example, formant measurements, pronunciation characterization, and durational features used in traditional linguistic forensic analysis all have correlates in various methods described in Section 2.

It is important to convey that (1) automatic recognition systems exhibit less-than-perfect feature extraction (e.g., phone and word recognition errors), but that (2) their performance does not require perfect recognition of such features. In fact, (3) the errorful behavior can be speaker-dependent and may then be exploited for speaker discrimination. To illustrate the last point: phone-recognition-based speaker models are more powerful when phonotactic constraints are excluded, even though the exclusion of constraints leads to higher error rates in phone recognition itself.

3.3. Data acquisition

Automatic methods enable the processing of larger amounts of data than methods based on human annotation ever could. This enables the construction of large speech databases that can yield detailed population statistics that in turn are needed for the accurate estimation of the typicality of a feature, as needed for likelihood ration computation. We should note that combined human/automatic annotation methods [33] could also have a place in forensics. For example, to enhance accuracy of automatic feature extraction, one could produce human word transcripts (which are relatively inexpensive) for the speech samples in question, thereby enabling more accurate automatic alignments than those obtained by purely automatic word recognition.

3.4. Robustness to acoustic variability

A major problem for accurate likelihood ratio estimation is acoustic mismatch between background and test data, or between known and questioned recordings. Such mismatch could be due to extrinsic factors (background noise, different recording channels) or intrinsic factors (e.g., speaking context, emotional state). Many higher-level features are inherently more robust to extrinsic features (e.g., pronunciation and durational properties), but could also be more variable than low-level features to intrinsic factors. Access to the full range of speaker features gives the forensic speaker recognition expert the choice to select the method most appropriate to a particular case.

3.5. Computational versus human effort

A counter-argument to higher-level modeling in speaker recognition that is often cited is the added computational overhead. For example, many high-level features require some form of speech recognition. However, this argument is less valid in forensics than in, say, commercial applications. The legal system can afford techniques that require detailed processing of speech data (possibly taking a few times real time), given that the alternative is even slower and more expensive human labeling and expertise.

4. Acknowledgments

We are grateful to Geoff Morrison and Joe Campbell for helpful comments. This work was supported by NSF IIS-0544682. The views are those of the authors and do not reflect those of the funding agency.

5. References

[1] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence", *Comp. Speech Lang.*, vol. 20, pp. 159–191, April-July 2006.

[2] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 2104–2115, Sep. 2007.

[3] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition", *in Proc. ICASSP*, vol. 4, pp. 784–787, Hong Kong, Apr. 2003.

[4] L. Ferrer, E. Shriberg, S. S. Kajarekar, A. Stolcke, K. Sönmez, A. Venkataraman, and H. Bratt, "The contribution of cepstral and stylistic features to SRI's 2005 NIST speaker recognition evaluation system", *in Proc. ICASSP*, vol. 1, pp. 101–104, Toulouse, May 2006.

[5] E. Shriberg, "Higher-level features in speaker recognition", in C. Müller, editor, *Speaker Classification I*, vol. 4343 of *Lecture Notes in Artificial Intelligence*, pp. 241–259. Springer, Heidelberg, 2007.

[6] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification", in J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, pp. 1337–1340, Denver, Sep. 2002.

[7] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained Gaussian mixture models", *in Proc. ICASSP*, vol. 1, pp. 677–680, Orlando, FL, May 2002.

[8] B. Baker, R. Vogt, and S. Sridharan, "Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification", *in Proc. Interspeech*, pp. 2429–2432, Lisbon, Sep. 2005.

[9] J. L. Gauvain, L. F. Lamel, and B. Prouts, "Experiments with speaker verification over the telephone", in J. M. Pardo, E. Enríquez, J. Ortega, J. Ferreiros, J. Macías, and F. J. Valverde, editors, *Proc. EUROSPEECH*, Madrid, Sep. 1995.

[10] M. Newman, L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, "Speaker verification through large vocabulary continuous speech recognition", in H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 4, pp. 2419–2422, Philadelphia, Oct. 1996.

[11] K. Boakye and B. Peskin, "Text-constrained speaker recognition on a text-independent task", *in Proceedings Odyssey-04 Speaker and Language Recognition Workshop*, pp. 129–134, Toledo, Spain, May 2004.

[12] D. Gillick, S. Stafford, and B. Peskin, "Speaker detection without models", *in Proc. ICASSP*, vol. 1, pp. 757–760, Philadelphia, Mar. 2005.

[13] H. Aronowitz, D. Burshtein, and A. Amir, "Text independent speaker recognition using speaker dependent word spotting", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, pp. 1789–1792, Jeju, Korea, Oct. 2004.

[14] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition", *in Proc. Interspeech*, pp. 2425–2428, Lisbon, Sep. 2005.

[15] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition", *in Proc. ICASSP*, vol. 1, pp. 149–152, Orlando, FL, May 2002.

[16] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines", in S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pp. 1377–1384, Cambridge, MA, 2004. MIT Press.

[17] H. Lei and N. Mirghafori, "Word-conditioned phone N-grams for speaker recognition", *in Proc. ICASSP*, Honolulu, Apr. 2007.

[18] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection", *in Proc. ICASSP*, vol. 4, pp. 804–807, Hong Kong, Apr. 2003.

[19] Y. Ka-Leung, W. Man-Mak, and S.-Y. K. Kung, "Articulatory feature-based conditional pronunciation modeling for speaker verification", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, pp. 2597–2600, Jeju, Korea, Oct. 2004.

[20] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification", in R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 7, pp. 3189–3192, Sydney, Dec. 1998. Australian Speech Science and Technology Association.

[21] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 2095–2103, Sep. 2007.

[22] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition", *in Proc. ICASSP*, vol. 4, pp. 788–791, Hong Kong, Apr. 2003.

[23] S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for speaker recognition", *in Proceedings Odyssey-04 Speaker and Language Recognition Workshop*, pp. 51–56, Toledo, Spain, May 2004.

[24] B. Peskin, J. Navrátil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high performance speaker recognition: Report from jhu ws'02", *in Proc. ICASSP*, vol. 4, pp. 792–795, Hong Kong, Apr. 2003.

[25] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition", *in Proc. EUROSPEECH*, pp. 2017–2020, Geneva, Sep. 2003.

[26] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition", *Speech Communication*, vol. 46, pp. 455–472, 2005, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation.

[27] E. Shriberg and L. Ferrer, "A text-constrained prosodic system for speaker verification", *in Proc. Interspeech*, pp. 1226–1229, Antwerp, Aug. 2007.

[28] G. Doddington, "Speaker recognition based on idiolectal differences between speakers", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, pp. 2521–2524, Aalborg, Denmark, Sep. 2001.

[29] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST speaker recognition evaluation system", *in Proc. ICASSP*, vol. 1, pp. 173–176, Philadelphia, Mar. 2005.

[30] G. Tur, E. Shriberg, A. Stolcke, and S. Kajarekar, "Duration and pronunciation conditioned lexical modeling for speaker verification", *in Proc. Interspeech*, pp. 2049–2052, Antwerp, Aug. 2007.

[31] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonmez, "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition", *in Proc. ICASSP*, vol. 4, pp. 233–236, Honolulu, Apr. 2007.

[32] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection", *Comp. Speech Lang.*, vol. 20, pp. 230–275, April-July 2006.

[33] R. Schwartz, W. Shen, J. Campbell, S. Paget, J. V. D. Estival, and C. Cieri, "Construction of a phonotactic dialect corpus using semiautomatic annotation", *in Proc. Interspeech*, pp. 942–945, Antwerp, Aug. 2007.