SPEECH RECOGNITION AS FEATURE EXTRACTION FOR SPEAKER RECOGNITION

A. Stolcke, E.Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, G. Tur

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

ABSTRACT

Information from speech recognition can be used in various ways in state-of-the-art speaker recognition systems. This includes the obvious use of recognized words to enable the use of text-dependent speaker modeling techniques when the words spoken are not given. Furthermore, it has been shown that the choice of words and phones itself can be a useful indicator of speaker identity. Also, recognizer output enables higher-level features, in particular those related to prosodic properties of speech. Finally, we discuss the use of mere by-products of word recognition, such as subword unit alignments, pronunciations, and speaker adaptation transforms to derive powerful nonstandard features for speaker modeling. We present specific techniques and results from SRI's NIST speaker recognition evaluation system.

Index Terms— Speaker recognition, speech recognition, high-level features, speaker adaptation, prosody.

1. INTRODUCTION

In a fundamental aspect, speaker recognition and speech recognition are dual problems. In speaker recognition, the goal is to identify the speaker irrespective of what is being said; in speech recognition the goal is to recognize what is being said irrespective of who is speaking. Thus, in speaker recognition, one of the fundamental problems is to normalize for variability due to the speaker's choice of phones, words, and so on; conversely, in speech recognition, a basic challenge is to normalize out speaker differences. (Normalizing out other sources of variability, such as channel properties, is common to both tasks.)

Maybe as a result of this dichotomy, speech recognition and speaker recognition have traditionally been pursued by different (although sometimes overlapping) research communities. Recent years, however, have seen widespread use of speech recognition in advanced speaker recognition research systems, such as those fielded in the annual NIST Speaker Recognition Evaluation (SRE). In this paper we survey some of the main techniques used, and point out possible future directions and as-yet-unexplored opportunities.

It should be noted that cross-fertilization has also occurred in the other direction. For example, speaker recognition techniques play a major role in diarization, the separation of audio signals into different speakers and nonspeech classes, a vital preprocessing step for speech recognition [1, 2]. However, in this paper we will focus on speaker recognition as the target application and examine links to speech recognition from that perspective.

2. OVERVIEW

We can distinguish four broad classes of approaches by which speech recognition can be leveraged for speaker recognition.

Modeling text dependency: here the output of the speech recognizer is used to bridge the gap between text-dependent and text-independent speaker modeling. Once the spoken words or phones are identified, the speaker features can be conditioned on this information, thereby obtaining sharper and hence more accurate speaker models.

Modeling lexical/phonetic speaker idiolect: it has been shown that choice of words, as well as of pronunciations for a given word, is speaker specific and can therefore be used to define features that identify speakers.

Modeling higher-level features: A virtually open-ended set of speaker features deals with information beyond frame-level acoustics and require phone, word, and alignment information as basic inputs to compute higher-level features A prime example is features based on speech prosody (supra-segmental pitch, duration, and energy patterns).

Modeling internal recognizer features: a speech recognizer computes a large amount of ancillary information in order to arrive at its final word or phone recognition output. Some of this information is very valuable as it tends to be highly speaker specific, either by design (such as by-products of speaker adaptation) or empirically (such as the duration patterns of subword units).

These four classes of approaches are conceptually distinct, but often overlap in practice. For example, the same alignment information that yields speaker-specific subword duration and pronunciation information is also required for defining higher-level prosodic features. In the following sections we will summarize existing approaches in the first three categories, and then focus on several instances of the

last category, whereby internal computations of the recognizer can be leveraged for speaker modeling.

3. MODELING TEXT DEPENDENCY

In this approach, acoustic models (typically based on cepstral features) are conditioned on the word or phone identities obtained from the speech recognizer. This allows the acoustic models to have lower variance from nuisance factors, and hence better speaker discrimination. A second benefit is that regions of speech that exhibit large inter-speaker variability, and hence potential for speaker discrimination, can be selected. One example of this approach is MITLL's text-constrained cepstral support vector machine (SVM) system where cepstral polynomial features are extracted separately for each of a set of frequent words; this approach in turn is based on a similar one using cepstral Gaussian mixture models (GMMs) [3]. Alternatively, one can use phone-based speech recognition hidden Markov models (HMMs) as speaker models [4], or even whole-word HMMs for frequent word ngrams [5]. In both cases, the text-constrained cepstral system and the whole-word approach, features are extracted based on the alignments provided by a word recognizer.

4. PHONETIC AND LEXICAL IDIOLECT MODELING

Phone recognition without lexical or phonotactic constraints ("open-loop recognition") has been used successfully as a feature extractor for speaker recognition [6, 7], where phone sequences are modeled directly, either by language models or SVMs. Another approach is to compare open-loop phone sequences with those obtained from a word recognizer [8].

Doddington observed that certain word N-gram frequencies can express idiosyncratic language use and hence model speaker identity [9]; it was also shown that this approach can be even more effective when coupled with SVMs [10]. Recently, we obtained further improvements by extracting N-gram frequencies that record the duration (short vs. long) of frequent words, thus combining lexical and pronunciation information into a single SVM model [11].

5. HIGHER-LEVEL AND PROSODIC FEATURES

Once recognition output with detailed time alignments is available, one can start to model features beyond phones and words. A defining characteristic of such features is that they are extracted over regions that vary in extent, based on the underlying speech units modeled: syllables, words, pause-delimited phrases, and so on. We group these features under the term "nonuniform extraction region features" (NERFs) [12] and have recently focused on syllable-level NERFs (SNERFs) [13]. The bulk of these features capture speech prosody, that is, the patterns of variation in pitch, energy, and durations of speech units.

Modeling features at this level presents an interesting challenge because of their large number and nonuniform nature. An effective approach (for SNERFs) is to segment the speech signal into syllables based on recognizer alignments, extract a large number of features for each syllable, discretize the feature values, and then form N-grams of the binned values. These N-grams can then be counted and their frequencies modeled with SVMs similar to the phone and word models mentioned earlier. Results show that prosodic features defined in this way can be highly effective for speaker modeling, and can improve traditional low-level acoustic models substantially when combined with them [14].

6. INTERNAL SPEECH RECOGNIZER FEATURES

Some of the recognizer-based speaker models described above can be improved if they are generalized to leverage detailed information that is available in the recognizer, but which is usually discarded. One such case concerns a NERF-type prosodic model. An important aspect of prosodic variation is the duration of speech units. Specifically, inspired by work in word recognition, we built speaker models of phone durations within words [15]. Each word instance (from a list of frequent word types) gives rise to a feature vector consisting of the phone durations within it. These vectors are then modeled and scored in the familiar GMM-UBM (universal background model) framework. While such a phone-level duration model is effective, it is bested by a similar model that records the durations of HMM states *within a phone* [15].

A second example of adding recognizer-internal details to speaker modeling comes from phonetic speaker recognition. Traditionally, such systems have modeled the phone N-gram frequencies in the *most likely* phone string recognized. However, it turns out that a dramatic performance improvement can be achieved by modeling the frequencies of all (or all reasonably likely) phone N-grams according to their posterior probabilities of occurrence [16]. The space of all likely phone N-grams is represented in the recognizer search space and can be recorded as a phone lattice, from which the expected phone N-gram counts can be computed efficiently.

As a final example, we consider the acoustic models used by a speech recognizer. To achieve good performance, speaker-independent triphone models are adapted to obtain speaker-dependent models, using output from a first recognition pass as reference. The most common approach is to estimate an affine transform of the Gaussian means so as to maximize the likelihood of the adaptation data (maximum likelihood linear regression — MLLR). We may regard the transform parameters themselves as speaker-

dependent features and view them as a text-independent encapsulation of speaker-specific acoustic properties. Since the transforms are shared among all words, usually at the phone level, they are, in theory, a text-independent feature that does not suffer from the data fragmentation problem incurred by the word-conditioning approaches mentioned in Section 3. Indeed, we obtained excellent results with this approach, on a par or better than standard cepstral models [17].

7. RESULTS

To give a feel for the performance achievable using recognition-based speaker models, we summarize here some results from SRI's speaker verification system. This system represents an updated version of the system fielded by SRI in the 2006 NIST Speaker Recognition Evaluation (SRE) [11]. While not all the techniques mentioned above are used in this system, it incorporates a good cross-section of those techniques, chosen to give optimal performance in combination. The SRI system also contains several cepstrum-based models that represent the state of the art in speaker modeling without the benefit of speech recognition; we can therefore evaluate the incremental benefit from recognition-based models.

The following subsystems are used in the SRI system. The three recognizer-independent cepstrum-based models are a standard cepstral GMM-UBM system [18], a cepstral SVM system [19, 20], and a Gaussian supervector SVM system [21]. The remaining models are recognition based, and include an MLLR-SVM system (cf. Section 6) [17], state- and word-duration GMM systems (cf. Section 6) [15], a word/duration N-gram SVM (cf. Section 4) [9, 10], and an SVM model of prosodic feature sequences extracted over syllables and words (SNERFs and GNERFs, cf. Section 5) [13].

Table 1: Subsystems (models) used in SRI's SRE system

System , Model type	ASR-based	Normali-zation
Cepstral , GMM	No	ISV
Cepstral, SVM	No	T-norm
Gaussian supervector, SVM	No	ISV
MLLR, SVM	Yes	ISV
State duration, GMM	Yes	T-norm
Word duration, GMM	Yes	T-norm
Word/duration N-gram, SVM	Yes	T-norm
SNERFs+GNERFs, SVM	Yes	ISV

The subsystems are summarized in Table 1, which also indicates which normalization methods were applied to each system: either T-norm score normalization [22] or feature-level intersession variability normalization with nuisance attribute projection (NAP) for SVMs [23] or factor analysis for GMMs [24, 25]. The overall system score is obtained by an SVM-based combiner using all subsystem scores as inputs and trained to minimize the NIST decision cost function (DCF).

The test data used is the common condition (English language) subset of the NIST 2006 SRE. (In the interest of brevity, we only present results with one training conversation side per speaker, but the results with 8-side training are similar.) Table 2 shows performance metrics of the individual systems, as well as of the combined system. Results are presented in terms of both equal error rate (EER) and minimal achievable DCF.

Table 2: Individual and overall system performance.

System	%EER	DCF(x10)
Cepstral, GMM	4.75	0.216
Cepstral, SVM	5.07	0.242
Gaussian supervector, SVM	4.15	0.198
MLLR, SVM	4.00	0.197
State duration, GMM	16.03	0.705
Word duration, GMM	22.24	0.874
Word/duration N-gram, SVM	23.46	0.815
SNERFs+GNERFs, SVM	10.41	0.461
SVM combination	2.59	0.145

The results show that, generally speaking, the best performing subsystems are those based on acoustic cepstral features, either directly or indirectly (as in the case of mean supervectors or MLLR transforms). The MLLR-SVM, which incorporates recognizer information into acoustic speaker modeling, is currently our single best subsystem. The noncepstral systems perform the better the more low level and acoustically detailed their features are, with the best subsystem being the SNERF+GNERF (grammar-constrained nonuniform extraction region feature) SVM.

To assess the contributions of each model to the overall performance it is not enough to compare individual performance metrics. Rather, we should be asking how much each subsystem is improving the overall combination given that other subsystems are already present. In this context it also makes sense to ask which subsystems constitute the two, three, four, and so forth best in combination.

Figure 1 answers both questions, by listing the performance metrics for the N-best system combinations and the subsystems chosen for each N-best combination. With one exception, the N-best subsystems are chosen monotonically, i.e., the N-best combination is a proper superset of the (N-1)-best combination, allowing us to put a strict ranking on the importance of the subsystems. The interesting result concerning this ranking is that *stylistic* recognizer-based systems (such as the prosodic and lexical sequence models) are chosen before other subsystems that have considerably better performance in isolation. In particular, the most important noncepstral subsystem is the

SNERF+GNERF system, which relies on recognizer information. A likely explanation for this phenomenon is that several of the cepstral systems are highly correlated and therefore give only little incremental speaker information

	MLLR SVM (A)	S+G SVM (S)	Supvec SVM (A)	WDur GMM (S)	SDur GMM (S)	WDNG SVM (S)	Cep GMM (A)	Cep SVM (A)	%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

Figure 1: N-best results for SRE06-1side common condition. (A) refers to acoustic feature-based system and (S) refers to stylistic feature-based system

.

8. CONCLUSIONS

We have given a brief survey of speaker recognition techniques that make use of speech recognition. The role of speech recognition can range from simple transcription to enable text-dependent modeling to the extraction of novel speaker features that characterize the behavior of the speech recognizer (such as phone-state durations and MLLR transforms). Results from a selection of these techniques show that such features have much potential, and contribute greatly to a system combination that also includes traditional state-of-the-art cepstral systems.

ACKNOWLEDGMENTS

We gratefully acknowledge Robbie Vogt of Queensland University of Technology for his implementation of the Gaussian supervector system within the SRI system, as well as for contributions to intersession variability (ISV) compensation. The research summarized here was partly funded by DoD KDD award NSF IRI-9619921, with additional support from the Department of Homeland Security via NSF IIS-0325399. The views herein are those of the authors and do not reflect the views of the funding agencies.

REFERENCES

[1] S. Tranter and D. Reynolds, "Speaker Diarisation for Broadcast News," *Proc. Odyssey Speaker Recognition Workshop*, pp. 337-344, Toledo, Spain, 2004.

[2] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," *Proc. Eurospeech*, pp. 2441-2444, Lisbon, 2005.

[3] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri, "Speaker Verification Using Text-Constrained Gaussian Mixture Models," Proc. *IEEE ICASSP*, vol. 1, pp. 677-680, Orlando, FL, 2002.

[4] M. Newman, L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, "Speaker verification through Large Vocabulary Continuous Speech Recognition." *Proc. ICSLP*, vol. 4, 2419-2422, Philadelphia, 1996.

[5] K. Boakye and B. Peskin. "Text-constrained Speaker Recognition on a Text-independent Task," *Proc. Odyssey Speaker Recognition Workshop*, pp. 129-134, Toledo, Spain, 2004.

[6] W. D. Andrews, M. A. Kohler, and J. P. Campbell, "Phonetic Speaker Recognition," *Proc. Eurospeech*, pp. 149-153, Aalborg, 2001.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic Speaker Recognition with Support Vector Machines," in *Advances in Neural Information Processing Systems 16*, 2004.

[8] D. Klusacek, J. Navratil, D. A. Reynolds, and J. P. Campbell, "Conditional Pronunciation Modeling in Speaker Detection," *Proc. IEEE ICASSP*, vol. IV, pp. 804-807, Hong Kong, 2003.

[9] G. Doddington, "Speaker Recognition Based on Idiolectal Differences between Speakers," *Proc. Eurospeech*, pp. 2521-2524, Aalborg, 2001.

[10] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST Speaker Recognition Evaluation System," *Proc. IEEE ICASSP*, vol. 1, pp. 173-176, Philadelphia, 2005.

[11] S. Kajarekar et al., "2006 NIST Speaker Recognition Evaluation: SRI System Description," National Institute of Standards and Technology, 2006.

[12] S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for Speaker Recognition," *Proc. Odyssey 04 Speaker and Language Recognition Workshop*, pp. 51-56, Toledo, Spain, 2004.

[13] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication* 46(3-4), 455-472, 2005.

[14] L. Ferrer, E. Shriberg, S. S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman, and H. Bratt, "The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System," *Proc. IEEE ICASSP*, vol. 1, pp. 101-104, Toulouse, 2006.

[15] 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," *Proc. Eurospeech*, pp. 2017-2020, Geneva, 2003.

[16] A. O. Hatch, B. Peskin, and A. Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding," *Proc. IEEE ICASSP*, Philadelphia, vol. 1, pp. 169-172, 2005.

[17] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR Transforms as Features in Speaker Recognition," *Proc. Eurospeech*, Lisbon, pp. 2425-2428, 2005.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing* 10, 19-41, 2000.
[19] W. M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," *Proc. ICASSP*, vol. 1, pp. 161-164, Orlando, 2002.
[20] S. S. Kajarekar, “Four Weightings and a Fusion: A Cepstral-SVM system for speaker recognition,” *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 17-22, San Juan, PR, 2005.
[21] W. M. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308-311, 2006.
[22] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, “Score Normalization for Text-independent Speaker Verification Systems,” *Digital Signal Processing* 10, 42-54, 2000.
[23] A. Solomonoff, C. Quillen, and I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," *Proc. ICASSP*, vol. 1, pp. 629-632, Philadelphia, 2005.
[24] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," *Proc. ICASSP*, vol. 1, pp. 113-116, Philadelphia, 2005.
[25] R. Vogt, B. Baker, and S. Shridharan, “Modelling Session Variability in Text-independent Speaker Verification,” *Proc. Eurospeech*, pp. 3117-3120, Lisbon, 2005.