DURATION AND PRONUNCIATION CONDITIONED LEXICAL MODELING FOR SPEAKER VERIFICATION

Gokhan Tur Elizabeth Shriberg Andreas Stolcke Sachin Kajarekar

SRI International,
Menlo Park, CA, 94025, USA
{gokhan,ees,stolcke,sachin}@speech.sri.com

ABSTRACT

We propose a method to improve speaker recognition lexical model performance using acoustic-prosodic information. More specifically, the lexical model is trained using duration- and pronunciation-conditioned word N-grams, simultaneously modeling lexical information along with their acoustic and prosodic characteristics. Support vector machines are used for modeling and scoring, with N-gram frequency vectors serving as features. Experimental results using NIST Speaker Recognition Evaluation data sets show that this method outperforms the regular word N-gram-based lexical models. Furthermore, our approach gives additional information when combined with a high-accuracy acoustic speaker model. We believe that this is a promising step toward integrated speaker recognition models that combine multiple types of high-level features.

Index Terms— speaker verification, speaker recognition, lexical modeling, SVM.

1. INTRODUCTION

Speaker verification systems aim to automatically detect whether the person who is speaking matches the given name on the basis of individual information included in speech waveforms. Speaker verification is widely used for forensic purposes and to control access to services such as voice dialing [1]. Speaker recognition and verification systems have been traditionally based on acoustic features, such as cepstral features, typically modeled using Gaussian Mixture Models (GMMs) [2], and these systems have been evaluated using only very short segments of speech. While such features are proven to be extremely useful, acoustic models are known to be sensitive to channel mismatch and environmental noise.

Recently, higher-level stylistic features have become more popular as official evaluations have started to include longer test conversations and higher-level features have been shown to improve performance when combined with acoustic features [3]. Among the higher-level features investigated are prosodic features, such as pitch, duration, and energy characteristics [4], and lexical features, such as word and phrase (N-gram) frequencies [5]. These stylistic models are by definition more robust to channel mismatch and environmental noise, and, if based on sufficiently accurate speech recognition, can be expected to perform better under those conditions. Even under clean acoustic conditions, stylistic models can capture information that is complementary to short-term spectral features.

Previous work on higher-level features for speaker recognition typically focused on building separate models using different types of features, followed by score-level combination. For example, in our earlier work we employed different models for acoustic, lexical, and prosodic features and combined them using a neural network [4]. An alternative approach is to use complementary features in a single speaker model, assuming that classifier training can find the best way to combine them. In this paper, we focus on the lexical N-gram model and investigate ways to integrate it with certain kinds of acoustic and prosodic information. More specifically, the lexical model is applied to duration- and pronunciation-conditioned word N-grams. We see this as a first step toward building more integrated speaker recognition models.

Earlier work is summarized in Section 2. In sections 3 and 4, we describe duration- and pronunciation-conditioned N-gram models, respectively. Section 5 presents experiments and results.

2. PREVIOUS WORK

Early work on using lexical information in speaker recognition is described in [6], but did not produce significant gains presumably due to the short training and test durations used at the time. In 2001, Doddington proposed using a model with only word unigrams or bigrams [5] and showed it to give promising results when applied to full conversations. The model was based on a conventional log-likelihood test, in which the log of the ratio of speaker and background model likelihoods is averaged for all N-grams in an utterance, indexed by j:

$$\text{Score} = \frac{\sum_j \log \frac{\Lambda_{\text{Speaker}(j)}}{\Lambda_{\text{Background}(j)}}}{\sum_j 1}$$

It was shown that performance improved steadily as the amount of training data per speaker increases, and using only a small subset of N-grams resulted in performance similar to that of using all N-grams.

Following this study, other researchers focused on combining the lexical model with existing acoustic models, as well as improving the model. Andrews *et al*. applied the N-gram frequency modeling framework to phone N-grams obtained from a phone recognizer [7]. Since the phone recognizer is unconstrained such an approach captures discretized acoustic properties, as well as idiosyncratic pronunciations. A combination of both lexical and phonetic models with a conventional GMM-based cepstral system showed significant improvements.

Recently, Baker *et al*. showed that the lexical and phonetic N-gram frequency models can be improved by training the speaker model via maximum a posteriori (MAP) adaptation of a background model [8]. They also showed the effectiveness of this approach when smaller amounts of speaker-dependent data are available [9].

Meanwhile, Campbell *et al*. proposed a way to model phone N-gram frequencies in the support vector machine (SVM) framework [10]. A similar approach for word N-grams was shown to be superior to log-likelihood ratio modeling and employed in SRI's 2004 NIST evaluation system [4], giving improvements in combination with acoustic and prosodic speaker models. In the SVM formulation, speaker verification is treated as a binary classification task, and relative frequencies of word N-grams (possibly scaled or normalized) are used as features. All the N-grams appearing more than twice in the background training data were included as features, and no smoothing or boosting was employed. A lexical SVM model combined with combined with a cepstral GMM system reduced equal error rate (EER) on the NIST 2004 evaluation set by 11% over the cepstral system alone, in the 1-conversation-side training condition, and by 50% in the 8-side condition.

SRI also investigated the effect of using state, phone, and word durations for the speaker verification task, employing GMM log-likelihood ratio models [11]. Such models gave an additional 12% error reduction with combined with both cepstral GMM and the Doddington word N-gram model. On the NIST 2004 evaluation task, the word duration model was shown to be almost as accurate as the SVM N-gram model [4]. The cepstral, lexical, duration, and additional prosodic models together achieved more than 60% error reduction over a cepstral GMM by itself.

3. DURATION-CONDITIONED WORD N-GRAM SVM SYSTEM

The duration-conditioned word N-gram-based SVM system aims to model speaker-specific word usage patterns combined with differences in the durations of frequent words. Following earlier work, our approach is to treat the N-gram frequencies of each conversation side as a feature vector that is classified by a speaker-specific SVM. Word durations are binned and different bins are counted separately.

The duration-conditioned word N-gram SVM system is constructed as follows: All instances of the most frequent 5000 word types (as optimized on a development set) are binned into two categories, "slow" and "fast", with respect to their duration. Durations are measured according to the acoustic alignments of the speech recognizer (ASR) output, and are therefore subject to ASR errors, just like the word labels themselves. Then, each of word w is labeled as either w_{slow} or w_{fast} for the purpose of computing the N-gram frequencies. Word types outside of the top 5000 are not differentiated according to their duration. with more than these two bins

N-grams were chosen for inclusion in the model based on frequency in the background training data. The background set comprised 1971 conversation sides from the Fisher corpus, Switchboard-2 NIST SRE 2003 data, Switchboard-2 Phase 5 data. N-gram lengths up to 3 were considered. Based on results with Fisher and Switchboard-2 test data, we retained all N-grams occurring at least 5 times in the background set, for a total of about 600,000 N-gram types.

The relative frequencies of the N-grams in a conversation side form a (typically sparse) vector of feature values. The values are then rank-normalized to the range $[0, 1]$, using the background data as the reference distribution. The SVM was trained using a linear kernel, with a bias of 500 against misclassification of positive examples to compensate for the imbalance of positive (target speaker) and negative (background) samples. This weight is due to the big mismatch in the number of examples for each class. The signed distance from the SVM decision boundary was used as the speaker verification score, and was normalized using T-NORM [12]. Normalization statistics were obtained from 248 Fisher speaker models.[1] The same set of T-NORM speakers is for both 1-side and 8-sides training conditions.

4. PRONUNCIATION-CONDITIONED WORD N-GRAM SVM SYSTEM

The pronunciation-conditioned word N-gram SVM system aims to model speaker-specific word usage patterns, represented via pronunciations of the words instead of their surface forms. Similar to the duration-conditioned lexical model we treat the N-gram frequencies of each conversation side as a feature vector that is classified by a speaker-specific SVM.

The pronunciation-conditioned word N-gram SVM system is built in a very similar fashion to the duration-conditioned lexical model. The only difference is that word instances, and hence word N-grams, are differentiated by their

[1] Fisher test conversations were trimmed to 2.5 minutes to better match the average amount of data in NIST SRE data.

	Baseline		Duration-Conditioned	
	EER (%)	DCF (x10)	EER (%)	DCF (x10)
Fisher-1	23.14	0.817	19.49	0.743
Fisher-2	21.01	0.734	18.29	0.673
NIST 2004 1-side	23.19	0.787	20.52	0.779
NIST 2004 8-side	10.93	0.505	10.20	0.486
NIST 2005 1-side	24.58	0.860	21.51	0.785
NIST 2005 8-side	11.25	0.484	9.03	0.389
NIST 2006 1-side	25.63	0.842	23.46	0.815
NIST 2006 8-side	11.14	0.515	9.95	0.446

Table 1. Comparison of the baseline and duration-conditioned lexical models for various evaluation data sets.

	Baseline		Duration Conditioned	
	EER (%)	DCF (x10)	EER (%)	DCF (x10)
NIST 2004 1-side	23.19	0.787	21.29	0.802
NIST 2004 8-side	10.93	0.505	10.49	0.568

Table 2. Comparison of the baseline and pronunciation-conditioned lexical models.

pronunciations (phone strings) in the ASR output. In our dataset, on the average there are 1.4 pronunciation alternatives per word as determined by the ASR dictionary. Every N-gram that occurs at least five times in the same background set is included in the N-gram vocabulary of the system, yielding a total of 200,000 N-gram types. As before, the feature values are rank-normalized to the range [0,1], and used in a linear-kernel SVM.

5. EXPERIMENTS AND RESULTS

We performed experiments using the two Fisher test sets, as well as NIST 2004, 2005, and 2006 SRE data sets. All SVM training and scoring was based on a modified version of the SVM-Light toolkit [13]. Results are presented in terms of equal error rate (EER) and minimum detection cost function (DCF) metrics. DCF is defined as

$$\mathrm{DCF} = C_{\mathrm{MD}} \times P_{\mathrm{target}} \times P_{\mathrm{MD}} + C_{\mathrm{FA}} \times (1 - P_{\mathrm{target}}) \times P_{\mathrm{FA}}$$

where C_{MD}=10, $C_{\mathrm{FA}} = 1$, and $P_{\mathrm{target}} = 0.01$.

Table 1 compares the baseline lexical model with the duration-conditioned lexical model. Performance can be seen to improve for all cases. The relative error reductions are typically larger for the 8-side condition. For the most recent (2006) test set, the EER reduction is 8.5% for 1-side and 10.7% for 8-side training. The minimum DCF reduction is small, only 3.2% for 1-side, but 13.4% for 8-side training.

Table 2 compares the baseline lexical model with the pronunciation-conditioned lexical model for the NIST 2004 evaluation data set. We get mixed results when using this method. The EER reduction is 8.2% for 1-side and 4.0% for 8-side training. However, DCF increases 12.5% for the 8-side case. These results indicate that while the duration-conditioned model is better for 8-side training, the pronunciation-conditioned model is worth considering only for 1-side, and prone to more missed detections for 8-side training.

The different behavior of the two models may be due to the data fragmentation resulting from different pronunciations. Note that the fragmentation effect is limited in the duration-conditioned model for two reasons: the number of duration bins was set at two, and duration is modeled only for the most frequent words. For future work we are considering binning of pronunciations to a small number, and limiting the pronunciation-conditioned vocabulary.

To investigate how much our new approach can add to a state-of-the-art speaker verification system, we combined the duration-conditioned lexical model with a maximum-likelihood linear regression (MLLR) based speaker verification system [14]. The MLLR system uses the speaker adaptation transforms used in speech recognition as features for speaker verification. The transforms are estimated using MLLR, and can be viewed as a text-independent encapsulation of the speaker's acoustic properties. After rank-normalization the MLLR features are modeled by SVMs using a linear kernel. For combining the lexical model with the MLLR system, we employed an SVM-based combiner using the individual system scores as features.

Table 3 presents the results using the combination of the MLLR system with the baseline and duration-conditioned lexical models for the NIST 2006 evaluation data set. As seen,

	MLLR only		+ Baseline N-grams		+ Duration-Conditioned N-grams	
	EER (%)	DCF (x10)	EER (%)	DCF (x10)	EER (%)	DCF (x10)
NIST 2006 1-side	4.64	0.213	4.69	0.208	4.58	0.210
NIST 2006 8-side	2.29	0.085	2.19	0.081	2.13	0.080

Table 3. Comparison of the baseline and duration-conditioned lexical models when combined with a baseline acoustic system.

the proposed method gives slightly better equal error rates than the baseline when combined with the acoustic (MLLR) system. The DCF is largely unaffected by the choice of lexical models. The largest improvement over the acoustic-only system is seen for 8-side training, where the equal error rate is reduced by 7.0% relative using the duration-condition N-grams, compared to only 4.4% relative using the baseline N-gram model.

6. CONCLUSIONS

We have shown the effectiveness of simultaneously modeling lexical and acoustic-prosodic features for speaker modeling, in the form of duration- and pronunciation-condition word N-gram SVM systems. The experimental results using NIST SRE data sets shows that our approach improves up on standard lexical N-gram SVM model, and is effective when combined with a state-of-the-art acoustic speaker model. We hope this study will serve as motivation for an open range of possible ways to simultaneously model multiple feature types. In future work we are planning to investigate other types of high-level information for feature-level combination, as well as ways to mitigate the data fragmentation problem inherent in conditioning.

Acknowledgments: This work was supported by NSF IIS-0544682. The views herein are those of the authors and do not necessarily represent the views of the funding agency.

7. REFERENCES

[1] S. Furui, *Survey of the State of the Art in Human Language Technology*, chapter 1.7, pp. 36–41, Cambridge University Press, Australia, 1998.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[3] Elizabeth Shriberg, "Higher-Level Features in Speaker Recognition," in *Speaker Classification I*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, Heidelberg / Berlin / New York, 2007.

[4] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and Jing Zheng, "SRI's 2004 NIST speaker recognition evaluation system," in *Proceedings of the ICASSP*, Philadephia, PA, March 2005.

[5] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proceedings of the Eurospeech*, Aalborg, Denmark, September 2001.

[6] Larry Heck, "Integrating high-level information for robust speaker recognition," http://www.clsp.jhu.edu/ws2002/groups/supersid/071002-_High_Level_Info_for_Spkr_Rec.pdf, 2002.

[7] W. D. Andrews, M. A. Kohler, J. P. Campbell, and J. J. Godfrey, "Phonetic, idiolectal, and acoustic speaker recognition," in *Proceedings of the A Speaker Odyssey, The Speaker Recognition Workshop*, Crete, Greece, June 2001.

[8] B. Baker, R. Vogt, M. Mason, and S. Sridharan, "Improved phonetic and lexical speaker recognition through MAP adaptation," in *Proceedings of the Odyssey: The Speaker and Language Recognition Workshop*, 2004.

[9] B. Baker, R. Vogt, and S. Sridharan, "Phonetic and lexical speaker recognition in reduced training scenarios," in *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, Sydney, Australia, December 2004.

[10] William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, Douglas A. Jones, and Timothy R. Leek, "Phonetic speaker recognition with support vector machines," in *Advances in Neural Information Processing Systems 16*, 2004.

[11] L. Ferrer, H. Bratt, V. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proceedings of the EUROSPEECH*, Geneva, Switzerland, September 2003.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[13] "Svmlight support vector machine toolkit," http://svmlight.joachims.org.

[14] Andreas Stolcke, Luciana Ferrer, and Sachin Kajarekar, "Improvements in MLLR-transform-based speaker recognition," in *Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006, pp. 1–6.