SPEAKER RECOGNITION USING SYLLABLE-BASED CONSTRAINTS FOR CEPSTRAL FRAME SELECTION

Tobias Bocklet[1,2]*, Elizabeth Shriberg*[2]

[1] University of Erlangen-Nuremberg, Germany
[2] SRI International, Menlo Park, CA, USA

ABSTRACT

We describe a new GMM-UBM speaker recognition system that uses standard cepstral features, but selects different frames of speech for different subsystems. Subsystems, or "constraints", are based on syllable-level information and combined at the score level. Results on both the NIST 2006 and 2008 test data sets for the English telephone train and test condition reveal that a set of eight constraints performs extremely well, resulting in better performance than other commonly-used cepstral models. Given the still largely-unexplored world of possible constraints and combinations, it is likely that the approach can be even further improved.

Index Terms— Speaker recognition, higher-level features, GMMs, cepstral features, MFCCs, syllables

1. INTRODUCTION

A standard approach in automatic speaker identification models a speaker by spectral short time information using a Gaussian mixture model (GMM) [1] and Mel frequency cepstral coefficients (MFCCs) as features. The framework uses a universal background model (UBM) that is adapted by maximum a posteriori (MAP) adaptation to speaker-specific spectral features [2]. In this approach, all frames of speech (above some energy threshold) are considered together.

Several previous studies have investigated the use of word or phone information to condition the extraction of cepstral features, thereby reducing variability associated with phonetic content. For example, the approach in [3] conditions a cepstral GMM on the identities of frequent words. A variant conditions on syllables rather than words [4]. Reviews of a range of other studies that condition cepstral feature extraction regions on linguistic information are provided in [5] and [6]. In general such approaches can combine well with a standard cepstral system (i.e., one that uses all frames of speech), but have not outperformed standard systems on their own.

In the current paper we describe a new system based on various syllable-level constraints that to our knowledge have not been explored in previous work. Another novel aspect of the system is simply that it performs extremely well. The system was included as part of SRI's submission to the NIST speaker recognition evaluation (SRE) 2008. It has been run only for English data so far, and consists of eight GMM-UBM systems, each of which includes only frames from regions in the speech that match a particular "constraint". Note that these are not separate subsystems in the terms of cepstral features. Rather, all subsystems share exactly the same set of MFCC features, as shown in Figure 1. Our main intention was not to find the constraint system with the best performance but to investigate if reusing frames in different subsystems (i.e., in different contexts) adds new or complementary information to a combined system. The eight subsystems are combined by linear logistic regression (LLR) [7]. The resulting system outperforms SRI's otherwise top current cepstral-based systems on English telephone data, for both the NIST SRE 2006 and NIST SRE 2008 test data sets.

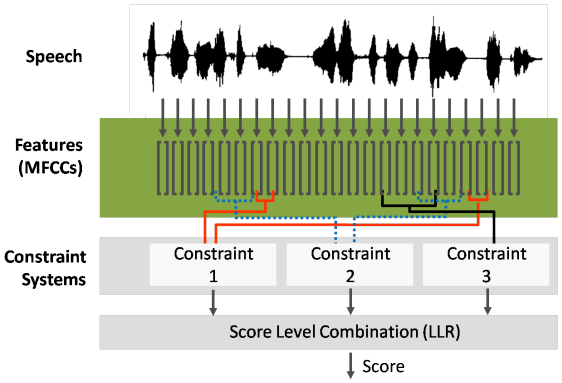

Fig. 1. Schematic depiction of the constrained GMM-UBM approach. Subsystems ("constraint systems") share the same cepstral features, but contain speech frames selected from different regions of the speech. Regions are based on syllable-level information.

The remainder of the paper is organized as follows. Section 2 describes the datasets for development and evaluation. Section 3 discusses the eight selected constraints, including their motivation and extraction. Section 4 describes the evaluation systems and their combination. In Section 5 we show the results of each subsystem on two standard speaker evaluation datasets (SRE 2006 and SRE 2008) and their combination. The paper concludes with a summary and proposed future research.

2. DATASETS

All datasets were provided by NIST and collected as part of the Mixer effort. Nontelephone data was preprocessed with the ICSI/Qualcomm Aurora Wiener Filter implementation [8]. Because the constraints used rely on word recognition (currently), and are thus language dependent, we used only the English subsets of the described datasets.

2.1. Development Sets

We used different datasets for system development. For background model training we used a subset (229 speakers) of the SRE 2004 English telephone data. For T-norm we used the same dataset. For intersession variability (ISV) compensation training, speakers from SRE 2004 were used. We used only those speakers for whom at

least eight different recordings were available. In addition, we used alternate microphone (altmic) data from SRE 2005. For training the logistic regression combiner we used SRE 2005 1-conversational (1conv) telephone-telephone (tel-tel) trials. The development was performed on SRE 2006 1conv tel-tel data [9]. The dataset contains recordings of one conversation per speaker. The speech duration is about 2.5 minutes per speaker for training and testing.

2.2. Evaluation Set

The system was evaluated on the test data for the 2008 NIST evaluation [10]. The SRE 2008 test data comprises a number of different conditions. In this paper we focus on the "short2-short3" and "8conv-short3" conditions. Short2-short3, i.e., training on short2 and testing on short3, was the required condition in the evaluation; we optimized our system for this condition. We also include results for the 8conv-short3 condition, which was optional but is useful for assessing the effect of increased training data on system performance. The training set (short2) and the test set (short3) contain one two-channel telephone recording per speaker. The 8conv training condition contains eight two-channel telephone conversations and no interview data.

3. CONSTRAINTS

A large set of candidate constraints was generated using information such as phone or phone class identity, syllable position (onset, nucleus, coda), combinations of these factors (e.g., voiced stops in onset position only; consonant clusters in coda position; presence of phone or class within a syllable), syllable structure (e.g., open syllables), adjacent pauses, number of syllables in each word and stress pattern in context of syllable. The number of Gaussians to use for each constraint had to be explored empirically, since constraints vary in number of selected frames as well as in the resulting homogeneity of selected frames. Well-performing constraints within a class of constraint types were retained. An ad hoc (but ultimately successful) approach was taken to choose a smaller set out of this set of about 100 candidates. The approach was a quasi-forward search, starting with a "syllable-nucleus" constraint. Note that no intersession-variability compensation (see below) was performed at this point on the individual constraints, for practical reasons, so the forward search performed is potentially suboptimal.

This approach does not select the best individual constraints. For example, a highly successful individual constraint selects frames only in syllables preceding a pause. Interestingly, this constraint does about as well as a full baseline system, while using only about one-sixth of the number of frames. However, once a "syllable-nucleus" constraint is present, a "*post*-pausal syllable" constraint is more useful to include in combination.

Our final set of eight constraints included the following, organized by type of constraint:

- (1) syllable onsets (31 %)
- (2) syllable nucleii (43 %)
- (3) syllable codas (22 %)
- (4) syllables following pauses (19 %)
- (5) one-syllable words (64 %)
- (6) syllables containing [N] (19 %)
- (7) syllables containing [T] (19 %)
- (8) syllables containing [B], [P], [V], or [F] (20 %)

The percentage value of frames in respect to an "all frames" system is given in brackets. An "all frames" system uses 3.1 Mio frames for UBM training.

We intentionally did not include a baseline or "all frames" system in the combination, as our end goal was a combination with SRI's set of other systems that includes such baseline systems. It is interesting to note that post-evaluation experiments showed we could reduce the set of constraints to five, without a significant loss in performance. Clearly, however, more research is needed in order to better understand the behavior of the constraints both alone and in combination.

4. SYSTEM DESCRIPTION

4.1. ASR System

The (ASR) is a fast and simplified version of SRI's conversational telephone speech recognition system, limited to two decoding and various rescoring passes [11, 12].

The word error rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus was 23.0 % and 36.1 %, respectively. On SRE 2006 altmic we measured a WER of 28.8 %.

4.2. Acoustic Modeling

As acoustic features we use MFCCs consisting of a 300-3300 Hz front end with 24 Mel filters. Thirteen coefficients were computed (C0-C12) with cepstral mean subtraction. Based on these 13 values the first, second and third order derivatives are calculated, resulting in a 52-dimensional feature vector. This feature vector is mean and variance normalized over the conversation side.

The features are used to model the speakers/impostors by GMMs. The dimension is different for each constraint (see Section 4.3). We used the eigenchannel MAP framework [13] for ISV compensation. The idea is to model a speaker by three different components: a speaker- and session-independent component, a speaker-dependent and a session-dependent component. A concatenation of the mean vectors of one GMM defines a so-called GMM supervector.

The rank of the eigenchannel matrix was set to 50; the number of iterations was 5. For matrix training a dataset with different sessions for each speaker is needed. For each constraint system we performed a ZT-NORM, i.e., a Z-NORM followed by a T-NORM [14]. Z-NORM was trained on 904 SRE 2004 speakers and T-NORM was trained on 229 speakers.

4.3. Constrained GMM Systems

For each of the eight constraints defined in Section 3 we created a different GMM-UBM subsystem based on the relevant frames. First we created a UBM for each constrained subsystem. This UBM was used to train the eigenchannel matrix. The target GMMs were then created by eigenchannel MAP adaptation to the specific constrained feature frames. The number of Gaussians was estimated heuristically on the SRE 2006 dataset. For the constraints that use syllables after pauses and syllables containing the phone [T], we used 1024 Gaussian densities. 512 were used for all the other constraints. The scores of each subsystem were combined by linear logistic regression (see Section 4.4).

4.4. Combination

We combined constraints at the score level using linear logistic regression [7]. While there are other options for combination, an ad-

vantage of the score-level combination at this early phase in understanding combinations of constraints is rapid turnaround when adding or removing constraints. The constrained system was used in combination with the other SRI systems for the NIST SRE 2008 submission. To perform a fair combination of these systems, we did not use auxiliary information as described in [15] for the combination of the constrained subsystems.

5. RESULTS

The short2-short3 condition of SRE2008 can be split into seven different English-only subconditions [10]. Because constraints were developed on 1-side telephone recordings, we describe the results on these conditions in detail (Section 5.2). In this section we also present the results of the 8conv-short3 condition and the combination results for both conditions. The combination results are compared to our baseline system, which is a single GMM-UBM that uses all frames. The features and training data are identical to the constrained system. Here, we compare the results of the constrained system and of a GMM baseline on the other conditions.

5.1. Tuning

We conducted tuning experiments using SRE 2006 1conv telephone-telephone condition data. The parameters we adjusted were the number of Gaussians for each subsystem, the number of expectation-maximization (EM) iterations and the method and data for training the eigenchannel matrix $\boldsymbol{U}$. The number of EM iterations was of minor impact for the results, so we set that to 5.

The training of the $\boldsymbol{U}$ matrix was performed with different data and different ways of training. We also altered the rank of the matrix, i.e., the number of eigenchannels. The first set of experiments addressed the rank of the matrix. We obtained best results with a rank of 50. We tried different datasets: SRE 2004 with telephone training data with and without adding SRE 2005 alternate microphone data. A training of two matrices (one for telephone and one for alternate microphone) with a rank of 25 followed by a concatenation of these matrices gave slightly better results than training one matrix with a rank of 50 and pooling all the data together. Changing the weights of these two matrices by using different ranks did not result in better performance. The combination of the two matrices degraded the performance of the telephone-telephone results but gave a large improvement for alternate microphone and mismatched data, i.e., training and testing data from a different domain. For the training of the $\boldsymbol{U}$ matrix the number of EM iterations was also of minor impact. We decided to use five iterations. Table 1 shows the combination results on the development set in the first column. The combiner was trained on SRE2005. For comparison to state-of-the-art systems that also use cepstral features, the table also shows results for SRI's other top systems in the 2008 NIST SRE submission.

system	SRE 2006		SRE 2008	
	EER	DCF	EER	DCF
Constrained Cep	1.30	0.075	2.77	0.134
GMM Cep	1.90	0.095	2.91	0.140
SV-PLP	1.79	0.074	3.42	0.142
SV-MFCC	1.84	0.089	3.68	0.143
SV-MLLR	2.38	0.108	4.15	0.189

Table 1. EER and DCF results for the five best-performing systems in SRI's NIST SRE 2008 submission

5.2. Telephone Results for Individual Constraints and Combination

Telephone condition results for the individual systems are shown in Table 2. The table shows the results on SRE 2008 for both English and native English speakers. The cuts to determine the native talkers were provided by NIST. Using only one-syllable words achieved the best stand-alone performance: An equal error rate (EER) of 4.40 % for Tel-Tel and 4.48 % for the native English condition. These two systems use far fewer frames than a baseline GMM with the same parameter setting, but achieve comparable results. They are followed by two constraints that use syllable subregions, i.e., nucleus and onset. As shown, the combination of all eight constraints achieves a large win for both conditions. A baseline system (i.e., a GMM-UBM with exactly the same parameter settings) the constraint systems achieves an EER of 3.91 % on native English data and 3.95 % on all English data. This is a relative improvement of the constraint system of 33 % on native English data and 30 % on all English data. Both values are significant with $p < 0.001$.

Table 1 shows results for the top SRI systems in the NIST SRE 2008 English telephone-telephone conditions for SRE 2006 and 2008. The first row shows the results of the constrained system, GMM Cep denotes a baseline GMM-UBM system, SV-PLP and SV-MFCC denote suportvector systems with GMM supervectors and either PLP or MFCC features and SV-MLLR denotes a suportvector with MLLR transforms as input vectors. A detailed explanation of all systems is given in [12].

Because the constrained system was developed only a short time before the evaluation, it was not possible to submit results for the 8conv-short3 condition in time for the evaluation deadline. But we evaluated this condition after the evaluation, and found even more impressive performance, i.e., an EER of 0.66 % and a DCF of 0.04. The EER is even better than SRI's submission for this condition. The relative improvement is 24 % at a significance level of $p < 0.1$.

5.3. Combination Results for Other Conditions

Table 3 shows the results of the nontelephone and mismatched conditions. The results are worse than the telephone-telephone results. We believe that this is due to several reasons. The constrained system is based on an ASR system, which is trained on telephone data only. Because of that, the subsystem did not perform very well on this condition. The EERs of the single systems are very high in case of interview-interview with different microphones. The problem of lack of interview data in ASR training is exacerbated by the problem of not having interview data in UBM training for the subsystems. For the interview-interview condition with the same microphone we achieved an EER of 3.77 % and a DCF of 0.08. The interview-interview condition with altmic is much worse than the same microphone condition, most likely reflecting the lack of different microphone training data for the ISV compensation. The interview-telephone condition appears similar to that of the interview-interview with different microphones, but the results are slightly better in the former. This can likely be attributed to the use of telephone data in the test condition.

6. SUMMARY AND CONCLUSION

We have described an approach that uses syllable-level constraints to restrict the frames in an otherwise-standard cepstral GMM-UBM system. Although clearly there are many other constraints to explore, we found a set of eight simple constraints that when com-

Test	# of trials	syllable onset	syllable nucleus	syllable coda	post pause	1-syll words	syllables with [N]	syllables with [T]	syllables with [B,P,V,F]	Combination EER (%)	Combination DCF (x10)
short2-short3											
Tel-Tel	17761	5.70	4.48	8.07	8.80	4.40	10.99	9.53	12.05	2.77	0.13
Tel-Tel (nat)	8489	5.76	4.77	8.39	9.38	4.28	11.19	9.38	11.68	2.63	0.12
8conv-short3											
Tel-Tel	7408	1.97	1.09	1.97	4.17	1.32	3.95	5.26	2.63	0.66	0.04
Tel-Tel (nat)	3993	2.26	1.88	2.26	4.53	1.51	4.15	5.28	3.02	1.13	0.05

Table 2. EER results of individual constrained subsystems on SRE 2008 short2-short3 and 8conv-short3 telephone conditions

Evaluation Condition	# of Trials	Constraint System EER	Constraint System DCF	Baseline GMM-UBM EER	Baseline GMM-UBM DCF
Int-Int	34181	12.87	0.53	17.18	0.66
Int-Int (same mic)	1727	3.77	0.08	2.57	0.145
Int-Int (diff mic)	32454	13.18	0.55	17.91	0.68
Int-Tel	10719	9.58	0.36	15.80	0.71
Tel-Alt	8442	7.33	0.25	7.74	0.35

Table 3. EER and DCF results for constrained and baseline system on SRE 2008 short2-short3 non-telephone conditions

bined at the score level yield state-of-the-art performance on NIST SRE 2006 and SRE 2008 evaluation data. The constrained system furthermore provides significant complementary information when combined with a baseline or "all frames" system. In addition, the constrained approach appears to show strikingly good performance for the NIST 8side training condition. In future work, we plan to investigate further constraints and their combination, as well as the question of optimal combination. For example, one could combine feature vectors from different constraints into a single supervector system. A final goal is to port the currently language-dependent approach to a more language-independent paradigm based on phone recognition.

7. ACKNOWLEDGMENTS

We thank Andreas Stolcke, Sachin Kajarekar, Nicolas Scheffer, Luciana Ferrer, and Martin Graciarena for development of the employed algorithms and providing speech recognition output. This work would not be possible without their help and fruitful discussions. This work was funded through an SRI development contract with Sandia National Laboratories, and by SRI NSF IIS-0544682 which supported work by the first author while visiting SRI. The views herein are those of the authors and do not reflect the views of the funding agencies.

8. REFERENCES

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transaction on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

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

[3] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. 677–680.

[4] B. Baker, R. Vogt, and S. Sridharan, "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*, 2005, pp. 2429–2432.

[5] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 1337–1340.

[6] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I: Fundamentals, Features, and Methods*, Christian Müller, Ed., number 4343 in Lecture Notes in Artificial Intelligence, pp. 241–259. Springer, 2007.

[7] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. van Leeuwen, P. Matějka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[8] L. Burget, S. Dupont, H. Garudadri, F. Grézl, H. Heřmanský, P. Jain, S. Kajarekar, and N. Morgan, "QUALCOMM-ICSI-OGI features for ASR," in *Proc. 7th International Conference on Spoken Language Processing*. 2002, p. 4, International Speech Communication Association.

[9] "NIST 2006 Speaker Recognition Evaluation Plan," 2006, http://www.nist.gov/speech/tests/sre/2006/sre06_evalplanv9.pdf.

[10] "NIST 2008 Speaker Recognition Evaluation Plan," 2008, http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf.

[11] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, College Park, MD., 2008.

[12] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "2008 NIST Speaker Recognition Evaluation: SRI System Description," in *NIST SRE Workshop*, Montreal, Canada, 2008.

[13] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech*, 2007, pp. 1242–1245.

[14] 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.

[15] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," *ICASSP*, pp. 4853–4856, 2008.