SPEAKER RECOGNITION WITH REGION-CONSTRAINED MLLR TRANSFORMS

Andreas Stolcke[1] *Arindam Mandal*[2] *Elizabeth Shriberg*[1]

[1]Microsoft Speech Labs, Mountain View, CA, U.S.A.
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, U.S.A.
anstolck@microsoft.com arindam@speech.sri.com elshribe@microsoft.com

ABSTRACT

It has been shown that standard cepstral speaker recognition models can be enhanced by *region-constrained* models, where features are extracted only from certain speech regions defined by linguistic or prosodic criteria. Such region-constrained models can capture features that are more stable, highly idiosyncratic, or simply complementary to the baseline system. In this paper we ask if another major class of speaker recognition models, those based on MLLR speaker adaptation transforms, can also benefit from region-constrained feature extraction. In our approach, we define regions based on phonetic and prosodic criteria, based on automatic speech recognition output, and perform MLLR estimation using only frames selected by these criteria. The resulting transform features are appended to those of a state-of-the-art MLLR speaker recognition system and jointly modeled by SVMs. Multiple regions can be added in this fashion. We find consistent gains over the baseline system in the SRE2010 speaker verification task.

Index Terms— Speaker recognition, MLLR-SVM, region-constrained speaker modeling.

1. INTRODUCTION

One of the recurring ideas in speaker recognition is the specialization of feature extraction or modeling to specific speech units or regions that can be consistently defined. The rationale for this approach is that the resulting models can be more stable relative to nuisance variation (less intra-speaker variability), more focused on speaker-specific properties (more inter-speaker variability), or simply sufficiently uncorrelated with the baseline model so as to give valuable complementary information about speaker identity. Early experiments along these lines generally used cepstral features and standard speech units, such as phone classes [1] and words [2]. Word constraints have also been employed for phone N-gram modeling [3] and prosodic features [4].

A recent generalization of this approach uses region constraints that can be based on complex combinations of phonetic, lexical, prosodic and other criteria, such as "syllables containing nasals", or "regions of falling pitch" in work on *region-constrained cepstral models* [5, 6, 7]. These studies have shown that state-of-the-art cepstral models can yield complementary combination when constrained by suitable region constraints. The improvement is realized by combining the baseline and constrained models at the score level.

This prior work naturally leads to the question of whether other types of speaker models can also benefit from generalized, linguistically motivated region constraints applied to feature extraction. Another commonly used speaker modeling paradigm is that based on maximum likelihood linear regression (MLLR) speaker adaptation transforms [8]. This paper investigates the use of region-constrained MLLR transforms, i.e., transforms estimated from a subset of speech frames selected by phonetic and prosodic criteria, as depicted in Figure 1. Defining these regions requires access to automatic speech recognition output, but such output is already being computed for our state-of-the-art baseline MLLR system. We evaluate various region-constrained MLLR systems on the most recent NIST Speaker Recognition Evaluation (SRE) dataset, SRE2010.

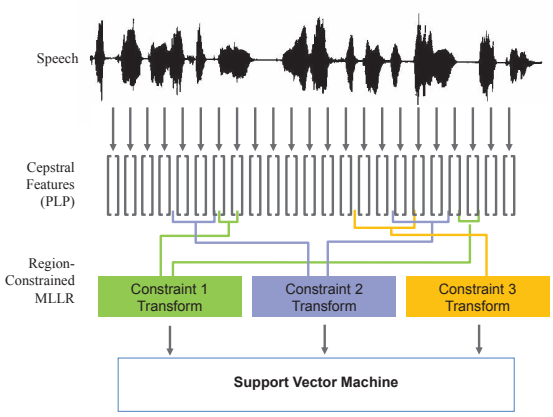

Fig. 1. Region-constrained MLLR speaker modeling

Note that the notion of "constraint" used here has no relation to the concept of "constrained MLLR", whereby the same transform is applied to both means and variances [9]. The MLLR transforms used here apply to model means only, and we use the term "region-constrained MLLR" to avoid confusion.

2. METHOD

2.1. Data and error metrics

Our data set is the NIST SRE 2010 extended core evaluation set. For phonecalls-over-microphone and interview sessions we use the recently released wide-band (sampled at 16 kHz) version of the data, as it avoids some significant issues with lossy waveform encoding that affect speaker recognition systems based on automatic speech recognition (ASR) in particular [10]. Table 1 summarizes the different evaluation conditions.

We report results according to three metrics: the traditional equal error rate (EER), which constrains false alarm and miss er-

Research done in part while first and third authors were with SRI International.

Table 1. SRE2010 evaluation set statistics, numbered according to NIST conditions. phn = phonecall, int = interview, mic = phonecall-over-microphone, nve = normal vocal effort, lve = low voc. eff., hve = high voc. eff.

Train-test condition	Target trials	Impostor trials
01.int-int.same-mic	4,304	795,995
02.int-int.diff-mic	15,084	2,789,534
03.int-nve.mic-phn	3,989	637,850
04.int-nve.mic-mic	3,637	756,775
05.nve-nve.phn-phn	7,169	408,950
06.nve-hve.phn-phn	4,137	461,438
07.nve-hve.mic-mic	359	82,551
08.nve-lve.phn-phn	3,821	404,848
09.nve-lve.mic-mic	290	70,500

ror rates to be the same, the old (pre-2010) detection cost function (oDCF), which weighs false alarm errors as ten times as costly as miss errors, and the new (2010) detection cost function (nDCF), which weighs false alarm errors as 1000 times more costly than miss errors. Old and new DCF values are scaled to make chance error rate equal to 1.

2.2. Baseline MLLR-SVM system

Our baseline MLLR-SVM system is identical to the one fielded as part of the 2010 SRI SRE system [11], modulo the use of wide-band microphone recordings and ASR (as described below). We also dropped the ZT score normalization step to expedite experimentation and since it was not adding significantly to performance.

An MLLR-SVM system uses speaker adaptation transforms, such as used by ASR systems, as features for speaker verification [8]. A total of 16 affine 39×40 transforms is used to map the Gaussian mean vectors from speaker-independent to speaker-dependent speech models; 8 phone-class specific transforms are estimated relative to male-only recognition models, another 8 transforms are computed based on female-only models (regardless of speaker sex). The transforms are estimated using MLLR [12], and can be viewed as a text-independent encapsulation of the speaker's acoustic properties. Speech features are 39-dimensional perceptual linear prediction (PLP) cepstra. The transform coefficients form a $39 \times 40 \times 8 \times 2 = 24,960$-dimensional feature space. Each feature dimension is rank-normalized, replacing the value with its rank in the background data, and scaling ranks to lie in the interval [0, 1]. Finally, nuisance attribute projection (NAP) [13] is applied to remove intra-speaker variability. The within-speaker variance was estimated on SRE04 telephone data, SRE05 microphone data, SRE08 and SRE10 sample data, and an SRE08 subset designated for training. The resulting normalized feature vectors are then modeled by SVMs using a linear kernel. The impostor (background) data for SVM training comes from SRE06 telephone and microphone sessions, as well as SRE08.

2.3. Region-constrained MLLR

To extend the notion of region-constrained feature extraction to MLLR speaker models, we first define regions in terms of phonetic, syllabic, and prosodic constraints, based on alignments of ASR output to the waveform, as well as energy and pitch tracks. SRI's *Algemy* prosodic engine then computes a set of start/end frame indices for each waveform, defining the regions where the constraints obtain. The MLLR estimation algorithm was modified to collect statistics only from the selected frames.

As an expedient, we chose regions based largely on prior experiments with constrained cepstral systems [5, 6]. Three constraints that had given among the best gains when combined with a baseline cepstral system were

- **Nasal syllables**—syllables containing one of the phones [m,n,ng]
- **[a] syllables**—syllables containing the phone [a]
- **Syllable nuclei**—the nuclear phone within a syllable
- **Falling energy**—syllables over which the smoothed energy contour had negative slope

Two gender-dependent transforms were estimated for each of these constraint regions (as for the baseline phone class transforms). The resulting transform coefficients were then appended to the baseline MLLR feature vectors. Note that more than one constraint-specific transform can be concatenated in order to combine information from multiple constraints. This feature-level combination method is different from the score-level combination explored so far for region-constrained cepstral systems. In preliminary experiments we found that score-level combination of multiple constrained MLLR-SVM systems, or score-level combination with the baseline, was not effective.

2.4. Speech recognition systems

The MLLR estimates used in our system rely on phone alignments generated by a word-ASR system. We used two systems, for telephone and microphone audio recording, respectively. Telephone sessions were transcribed by the ASR system used in SRE2008 and SRE2010 and described in [11]. This system uses acoustic models trained exclusively on telephone speech, and runs in two recognition passes, for purposes of unsupervised adaptation. We measured word error rate (WER) on transcribed portions of the Mixer corpus, giving 23.0% for native speakers and 36.1% for nonnatives (all SRE2010 data is English).

For microphone (including interview) sessions, we utilized a stripped-down version of the SRI/ICSI NIST RT-07 meeting recognition system [14]. It is similar to the telephone system in structure and modeling algorithms employed, but uses a combination of 8 kHz and 16 kHz acoustic models, trained on both near-field and distant-microphone meeting recordings, with telephone and broadcast news data used as background training, respectively. This system has a WER of 36.1% on single-distant-microphone test data from the 2007 Rich Transcription evaluation.

In developing the MLLR-SVM system on telephone data, we found the hypotheses from the first recognition pass most effective. For consistency with the telephone system we adopted the same strategy for the wide-band ASR system. However, we always use the final recognition output (the one with lowest WER) for computing the constraint regions.

3. EXPERIMENTS AND RESULTS

We start by evaluating individual constraint regions. Table 2 shows results for the baseline and MLLR-SVM systems based on each of the four regions defined earlier, as well as the percentage of speech frames selected by the respective constraints (as determined by alignment of ASR hypotheses). To reduce the number of results we focus on the telephone-telephone train-test condition of the SRE2010 data.

Table 2. Results for SRE2010 condition 5 (telephone-telephone, normal vocal effort), for baseline and each of the region-constrained MLLR systems. Systems are ordered by their share of speech frames used, shown in the second column.

System	share of speech frames	nDCF	oDCF	EER (%)
Baseline	100%	.4750	.1804	4.66
[a]-syl	49.1%	.7830	.3702	9.12
Eg-slope-neg	48.7%	.7910	.3890	9.35
Syl-nuclei	44.6%	.6980	.3080	7.20
Nasal-syl	25.4%	.8800	.5003	13.54

The individual regions yield EERs between 50% and 190% worse than the baseline, which is expected since fewer frames are used and only two gender-specific transforms are employed per constraint (compared to 16 transforms in the baseline system). The regions corresponding to nasal syllables perform the worst while also capturing only about half the amount of speech as the other three constraints. Still, performance is not simply a function of the amount of speech used, as demonstrated by the fact that the syllable nucleus constraint ranks third in amount of speech frames, but performs better than all other constraints.

In any case, performance of the constraints by themselves is not what matters for our purposes, since the intent is to extract speaker information that complements the baseline. Therefore, we evaluate each constraint in combination with the baseline MLLR features, by feature vector concatenation.

Table 3 compares Condition 5 results for the baseline with augmented systems, where one, two, three, or four region-specific transforms have been added. The improvements in nDCF go up to 8% relative. Old DCF gains are larger, up to 18%, as are EER reductions, up to 22% relative. Generally speaking, each added region improves the result, although for nDCF there is degradation when the third region is added.

The relative gains from adding one region-specific transform to the baseline (Table 3) do not reflect the performance of the regions by themselves (Table 2). In the latter case, syllables containing [a] give the second-best result; when added to the baseline, they actually produce a slight degradation (however, they do give a nice improvement in combination with other region-specific constraints). What these results highlight is that the selection of regions needs to be jointly optimized, something we have not attempted yet and which will be computationally expensive.

Table 4 presents results for all evaluation conditions, using the baseline augmented by all four region-specific transforms. The improvements for oDCF and EER are relatively consistent across conditions, on the order of 13% relative for oDCF and around 16% relative for EER. New DCF improves around 7.5%, but the gains vary greatly; presumably, many of the conditions lack sufficient trials to estimate error rates at very low false alarm rates. (This is a problem mainly for the conditions involving high/low vocal effort). As in the earlier results, we find that relative gains are larger toward the EER operating point, and decrease toward the very low false alarm region (nDCF).

4. CONCLUSIONS AND FURTHER WORK

We have investigated a further development of MLLR-SVM-based speaker modeling, incorporating the concept of region-constrained feature extraction, analogous to prior work on constrained cepstral speaker models. Results on the SRE2010 extended core data show gains for all performance metrics, across all evaluation conditions. Gains are generally highest for EER, then DCF, and smallest (and most variable) for nDCF.

Still, the results show clearly that regions defined in terms of syllable structure and low-level prosodic features can improve performance of a state-of-the-art MLLR system, combining all transforms at the feature level. Unlike for constrained cepstral models, the baseline MLLR system already incorporates phonetic information, through its use of phone-class specific transforms. In prior work [8] we had found that the eight baseline phone classes were highly optimized; in particular, splitting them further did not result in better performance. It is therefore highly significant, and encouraging for future work, that the addition of regions that are not defined purely in phonetic terms shows substantial gains.

We now have a matrix of acoustic speaker modeling techniques that comprises both cepstral and MLLR-based approaches, as well as unconstrained and region-constrained versions of these approaches. The work so far has shown that combining just two modeling approaches from this 2-by-2 matrix leverages complementary information and leads to improvements over the baseline. A full comparison of all these systems was beyond the scope of this work, but would clearly be of interest. Also, it remains to be seen if a 3-way or 4-way combination of modeling approaches would give still further improvements, and whether the gains are commensurate with the computational effort.

Furthermore, a number of other recent developments could be applied to the region-constrained MLLR system. As in [6], we must ask if language-independent constraints (e.g., based on multi-lingual phone recognition) perform similarly to MLLR based on English word recognition. Also, a new MLLR feature back end based on factor analysis and iVector fusion [15] looks very promising in conjunction with region-constrained MLLR.

5. ACKNOWLEDGMENTS

We thank Luciana Ferrer for creating much of the data infrastructure for the SRE2010 system, as well as Nicolas Scheffer and Martin Graciarena for fruitful discussions.

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government.

6. REFERENCES

[1] 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.

[2] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained Gaussian mixture

Table 3. Results for SRE2010 condition 5 (telephone-telephone, normal vocal effort), for baseline and region-constraint-augmented systems. %change refers to relative error reduction against baseline.

System	nDCF	%change	oDCF	%change	EER (%)	%change
Baseline	.4750	-	.1804	-	4.66	-
+ Syl-nuclei	.4560	4.00	.1637	9.26	4.20	9.88
+ Nasal-syl	.4550	4.21	.1723	4.49	4.41	5.39
+ Eg-slope-neg	.4520	4.84	.1686	6.54	4.20	9.88
+ [a]-syl	.4900	-3.16	.1804	-2.44	4.55	2.39
+ Syl-nuclei + Nasal-syl	.4370	8.00	.1598	11.42	3.99	14.37
+ Syl-nuclei + Nasal-syl + Eg-slope-neg	.4450	6.32	.1531	15.13	3.82	17.96
+ Syl-nuclei + Nasal-syl + Eg-slope-neg + [a]-syl	.4380	7.79	.1479	18.02	3.63	22.16

Table 4. Comparison of baseline and baseline augmented with 4 constraint regions, for all SRE2010 conditions.

Condition	nDCF			oDCF			EER (%)		
	Baseline	+ 4 regions	%change	Baseline	+ 4 regions	%change	Baseline	+ 4 regions	%change
01.int-int.same-mic	.4650	.4230	9.03	.1895	.1623	14.35	5.34	4.39	17.83
02.int-int.diff-mic	.5620	.5290	5.87	.2687	.2319	13.70	7.84	6.80	13.36
03.int-nve.mic-phn	.5340	.4660	12.73	.2306	.1995	13.49	6.47	5.41	16.28
04.int-nve.mic-mic	.4620	.4180	9.52	.1840	.1569	14.73	5.31	4.23	20.21
05.nve-nve.phn-phn	.4750	.4380	7.79	.1804	.1479	18.02	4.66	3.63	22.16
06.nve-hve.phn-phn	.9800	.9720	0.82	.3274	.2831	13.53	7.25	6.12	15.67
07.nve-hve.mic-mic	.8440	.8250	2.25	.3180	.3005	5.50	8.08	6.69	17.24
08.nve-lve.phn-phn	.6510	.6140	5.68	.1706	.1413	17.17	3.77	2.91	22.92
09.nve-lve.mic-mic	.2600	.2230	14.23	.0750	.0721	3.87	3.10	3.10	0.00
Condition-averaged			7.55			12.71			16.19

models", *in Proc. ICASSP*, vol. 1, pp. 677–680, Orlando, FL, May 2002.

[3] H. Lei and N. Mirghafori, "Word-conditioned phone N-grams for speaker recognition", *in Proc. ICASSP*, vol. 4, pp. 253–256, Honolulu, Apr. 2007.

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

[5] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection", *in Proc. ICASSP*, pp. 4525–4528, Taipei, Apr. 2009.

[6] E. Shriberg and A. Stolcke, "Language-independent constrained cepstral features for speaker recognition", *in Proc. ICASSP*, pp. 5296–5299, Prague, May 2011.

[7] M. H. Sanchez, L. Ferrer, E. Shriberg, and A. Stolcke, "Constrained cepstral speaker recognition using matched UBM and JFA training", *in Proc. Interspeech*, pp. 141–144, Florence, Italy, Aug. 2011.

[8] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987–1998, Sep. 2007.

[9] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 357–366, Sep. 1995.

[10] A. Stolcke, M. Graciarena, and L. Ferrer, "Effects of audio and ASR quality on speaker verification systems", Submitted to Odyssey 2012 Speaker and Language Recognition Workshop, 2012.

[11] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system", *in Proc. ICASSP*, pp. 5292–5295, Prague, May 2011.

[12] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of HMMs", *Comp. Speech Lang.*, vol. 9, pp. 171–186, 1995.

[13] A. Solomonoff, C. Quillen, and I. Boardman, "Channel compensation for SVM speaker recognition", *in Proceedings Odyssey-04 Speaker and Language Recognition Workshop*, pp. 57–62, Toledo, Spain, May 2004.

[14] A. Stolcke, K. Boakye, Özgür Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system", in R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*, vol. 4625 of *Lecture Notes in Computer Science*, pp. 450–463, Berlin, 2008. Springer.

[15] N. Scheffer, Y. Lei, and L. Ferrer, "Factor analysis back ends for MLLR transforms in speaker recognition", *in Proc. Interspeech*, Florence, Italy, Aug. 2011.