Improvements in MLLR-Transform-based Speaker Recognition

Andreas Stolcke *Luciana Ferrer** *Sachin Kajarekar*

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
*Department of Electrical Engineering, Stanford University, Stanford, CA, USA
{stolcke,lferrer,sachin}@speech.sri.com

Abstract

We previously proposed the use of MLLR transforms derived from a speech recognition system as speaker features in a speaker verification system [1]. In this paper we report recent improvements to this approach. First, we noticed a fundamental problem in our previous implementation that stemmed from a mismatch between male and female recognition models, and the model transforms they produce. Although it affects only a small percentage of verification trials (those in which the gender detector commits errors), this mismatch has a large effect on average system performance. We solve this problem by consistently using only one recognition model (either male or female) in computing speaker adaptation transforms regardless of estimated speaker gender. A further accuracy boost is obtained by combining feature vectors derived from male and female vectors into one larger feature vector. Using 1-conversation-side training, the final system has about 27% lower decision cost than a state-of-the-art cepstral GMM speaker system, and 53% lower decision cost when trained on 8 conversation sides per speaker.

1. Introduction

A fundamental problem in speaker recognition is feature variability due to factors other than the speaker's identity. In the case of cepstral features, this includes the choice of spoken words, which is not modeled in standard bag-of-frame type speaker models. In an approach we proposed recently [1], text dependency is largely removed by using the adaptation transform employed by a large-vocabulary speech recognition system as the speaker feature. These transforms map the Gaussian means of speaker-independent recognition models to speaker-dependent values, estimated by maximum likelihood linear regression (MLLR) [2]. Since triphones are approximately invariant to the words in which they occur, text dependency in the cepstral domain is normalized out. Also, since many triphones share the same transform, the problem of data fragmentation (common to word- or phone-conditioned modeling approaches) is avoided. The model can be refined by using several phone regression classes, which are defined linguistically in our system.

A key element in using MLLR transform features successfully in speaker verification is appropriate feature normalization and modeling. We found rank normalization in combination with support vector machine models to give excellent results on standard NIST speaker recognition evaluation (SRE) test data.

After the last NIST SRE we found that our MLLR feature extraction procedure had a major flaw, due to a possible mismatch in recognition models used in training and test conversations. The present paper updates our earlier results after we fixed this flaw, and introduces an additional enhancement that suggested itself in the process. The new results are greatly improved; they substantially and consistently improve on those obtained with state-of-the-art cepstral models.

Section 2 reviews the MLLR feature computation and modeling. Section 3 describes how our previous implementation was improved. Experiments and results appear in Section 4, followed by conclusions.

2. Method

2.1. Speech recognition system

Our speech recognition system is a fast, two-stage version of SRI's conversational telephone speech (CTS) system, as originally developed for the 2003 DARPA Rich Transcription evaluation [3] and later modified for the NIST 2004 speaker recognition evaluation [4]. The system performs a first decoding using Mel frequency cepstral coefficient (MFCC) acoustic models and a bigram language model (LM), generating lattices that are then rescored with a higher-order LM. The resulting hypotheses are used to adapt a second set of models based on perceptual linear prediction (PLP) acoustic features. The adapted models are used in a second decoding pass that is constrained by trigram lattices, which generates N-best lists. These are then rescored by a 4-gram LM and prosodic models to arrive at the final word hypotheses. The whole system runs in about 3 times real time on a hyperthreaded 3.4 GHz Intel Xeon processor.

2.2. Speaker adaptation transforms

In maximum likelihood linear regression (MLLR) [2], an affine transform (A, b) is applied to the Gaussian mean vectors to map from speaker-independent (μ) to speaker-dependent (μ') means: $\mu' = A\mu + b$, where A is a full matrix and b a vector. In unsupervised adaptation mode, the transform parameters (coefficients) are estimated so as to maximize the likelihood of the recognized speech under a preliminary recognition hypothesis. For a more detailed adaptation, the set of phone models can be partitioned or clustered by similarity, and a separate transform is applied to each cluster.

In our system, MLLR is applied in both recognition passes. The first pass is based on a phone-loop model as reference, and uses three transforms, for nonspeech, obstruent, and nonobstruent phones, respectively. The second decoding pass uses a more detailed MLLR scheme, based on word references generated by the first pass, and nine different transforms corresponding to phone classes for nonspeech, voiced/unvoiced stops, voiced/unvoiced fricatives, high/low vowels, retroflex phones, and nasals. Also, in the second pass, a single feature-level transform is used to effect speaker-adaptive training [5]. In previous work [1] we found that these transforms are helpful in normalizing out corpus and channel differences, but should not be used for speaker modeling.

2.3. Feature extraction and SVM modeling

The coefficients from one or more adaptation transforms are concatenated into a single feature vector and modeled using support vector machines. The data used is from conversational telephone speech, and each conversation side is processed as a unit by the speech recognition system. Consequently, each conversation side produces a single set of adaptation transforms pertaining to the same speaker, and hence a single feature vector. Since our acoustic features (after dimensionality reduction) contain 39 components, the number of SVM feature components will equal the number of transforms $\times 39 \times 40$. The transform for nonspeech (pause) models is left out of the feature vector, since it cannot be expected to help in speaker recognition.

An SVM is trained for each target speaker using the feature vectors from a background training set as negative examples (of which there are many, typically in the thousands), and the target speaker training data as positive examples (of which there are few, typically 1 or 8). To compensate for the severe imbalance between the target and background data, we adopted a cost model [6] to weight the positive examples 500-fold with respect to the negative examples. Throughout, a linear inner-product kernel function was used for SVM training.

Prior to SVM training or testing, features need to be normalized to equate their dynamic ranges. To this end, we apply rank normalization, replacing each feature value by its rank among all the background data samples on a given dimension, and then scaling ranks to a value between 0 and 1. Rank normalization not only scales the feature distribution to a fixed interval, it also warps the distribution to be approximately uniform. This has the intuitive effect that the distance between two datapoints (along a single dimension) becomes proportional to the percentage of the population that lies between them.

An alternative to feature normalization is to optimize the kernel function explicitly for minimal classification error. This can be done in a number of ways, for example, by applying scaling factors to subfeature-vectors [7], or by introducing a scaling matrix derived from the within-speaker variances [8]. However, neither of these methods was used here.

2.4. Baseline systems

In evaluating MLLR-feature-based speaker recognition systems, we compared results to two state-of-the-art cepstral feature systems. The first baseline system is a Gaussian mixture model (GMM) with universal background model (UBM) [9], based on 13 MFCCs (without C0) and first-, second-, and third-order difference features. The features are mean-subtracted and modeled by 2048 mixture components. Gender-handset models are adapted from this model and used for feature transformation [10]. The final features are mean and variance normalized at the utterance level. The detection score is the target/UBM likelihood ratio after TNORM [11].

The second baseline system is also based on MFCCs (with first- and second-order differences), followed by the same feature transformation and normalization steps. The final features are then modeled with SVMs utilizing the polynomial sequence kernel proposed by [12], with some recently developed enhancements [13]. Principal component analysis is performed on the polynomial feature space, and the features are projected onto the subspace spanned by the background speaker set, as well as its orthogonal complement (there are more feature dimensions than background speakers). This process is then carried out twice, for two different feature normalization variants, and four separate SVM models are trained. The overall system score is the sum of the four SVM scores, after TNORM. This enhanced cepstral SVM system was the single best-performing component of SRI's NIST 2005 speaker recognition system [14].

3. Improvements

3.1. Gender issues

The adaptation transforms are dependent on the recognition models relative to which they are computed. Features derived from the transforms can be compared meaningfully only if they were computed relative to the same recognition models. This becomes an issue if a recognition system uses multiple models, for example, if recognition models are gender dependent, as is the case for our recognizer. Still, in principle gender dependency should not be a problem because of the special way in which trials in the NIST SRE data sets (as well as our own development data) are constructed. NIST SRE trials always have the same speaker gender (either all male or all female) in target training and test conversations. Presumably, gender mismatches between training and test data are considered uninteresting, since they would be too easy to classify as speaker mismatches.

Unfortunately, the automatic gender classification performed by our recognizer makes a nonnegligible number of errors. After the 2005 NIST SRE, we realized that this led to a significant number of trials (on the order of 20%) in which at least one of the conversation sides had incorrect gender, and hence mismatched MLLR transform features. In the following, we refer to such trials as "gender mismatched", even though this refers only to the automatically detected gender according to our speech recognition system.[1]

Table 1 shows the average target and impostor trial scores, as well as equal error rates (EERs) for trials in which a gender mismatch occurs, and compares them to the statistics for trials overall. These statistics are given for a standard cepstral GMM system as well as the MLLR SVM system described in [1]. The data is drawn from the Fisher collection, and has one training conversation side per target speaker. The table shows that the MLLR system scores drop dramatically for trials with gender mismatch, unlike for the cepstral GMM. As a result, the MLLR system performs almost at chance level (46% equal error) for gender-mismatched cases.

3.2. Fixing MLLR gender mismatch

One solution to the problem of gender-mismatched trials is to compute MLLR transforms with gender-independent models. However, in our case this would have meant retraining the recognition system from scratch, a rather involved process. Instead, we decided to simply use one of the gender-dependent models (male or female) for all speaker samples. This means that gender-dependent processing steps (vocal tract length normalization and adaptation) are rerun with the opposite gender for about half the conversation sides. The recognition hypotheses used to compute the second, more detailed MLLR transforms, were kept unchanged from the original system, so no additional decoding was necessary relative to the original system.

[1] It should be pointed out that the gender identification subsystem of our recognizer has been developed independently of the speaker verification system, and works reasonably well for its intended purpose. In particular, we found that reducing the gender detection error rate does not improve word recognition accuracy.

Table 1: System output (scores and EER) comparison depending on whether the target training speaker and the test speaker are mismatched in automatic gender detection.

	Target trial scores		Impostor trial scores		EER	
Gender outputs	all	mismatched	all	mismatched	all	mismatched
GMM system	7.00	7.62	0.133	0.168	3.99%	2.32%
MLLR system	6.35	0.11	0.065	0.089	5.19%	46.2%

Table 2: Data sets used in experiments

Test set	SWB-II		Fisher	SRE-04 English-only		SRE-05 Common Condition	
Training	1-side	8-side	1-side	1-side	8-side	1-side	8-side
Conv. sides	3642	3058	734	1384	2695		
Models	578	546	734	479	225	506	384
Trials	9765	4911	16578	15317	7336	20907	15947

As reported in the next section, we tried using both male and female recognition models for all speakers. Furthermore, we explored the possibility of combining the two sets of transforms thus obtained. After both "male" and "female" transforms are computed, the corresponding feature vectors can be concatenated and again modeled by SVMs. Since the gender-dependent recognition models are not just linear transforms of each other, we can expect the two sets of MLLR features to afford two different, not entirely redundant "views" of the observation space, and the resulting combined system to have higher accuracy.

4. Experiments and Results

4.1. Datasets

We tested our baseline and MLLR-based systems on four databases: a subset of the NIST SRE-03 (Switchboard-II phase 2 and 3) data set, a selection of Fisher collection conversations, the NIST SRE-04 database, and the NIST SRE-05 data set. For all but Fisher, two data sets were available, for training on 1 and 8 conversation sides, respectively. The NIST SRE-04 and SRE-05 data sets were drawn from the Mixer data collection [15], which included telephone conversations in English as well as other languages. Since our method relies on a speech recognizer for English, we report on trials that involve only English conversations. For SRE-05 we chose the primary evaluation (Common Condition) subset, which is English only. Table 2 summarizes the statistics of these data sets. Note that the Switchboard-II trials were a subset of those used in the NIST SRE-03 evaluation, but had difficulty comparable to the full evaluation set, as measured by the performance of our baseline system.

The background training set consisted of 1553 conversation sides from Switchboard-II and Fisher that did not occur in (and did not share speakers with) any of the test sets, and that had duplicate speakers removed.

All data was processed identically by SRI's speech recognition system as described in Section 2. None of the test or background data were used in training or tuning of the recognition system.

In addition to feature-level normalization, we performed TNORM score-level normalization [11] in all experiments and for all systems, using speaker models drawn from a separate Fisher data set.

Table 3: Speaker verification results using MLLR features from 2nd adaptation stage (8 transforms). The top number (in italics) in each table cell is the EER (%). The bottom number is the minimum DCF value.

	Fisher	SRE-04	
MLLR gender	1-side	1-side	8-side
Mixed	*5.57*	*9.49*	*4.96*
(old system)	.08281	.33182	.18249
Male	*2.92*	*6.25*	*3.21*
	.06095	.28812	.12053
Female	*2.98*	*6.54*	*3.21*
	.05362	.29092	.14568
Male + Female	*2.85*	*5.34*	*2.62*
	.05493	.25640	.11767

Systems were optimized using the Fisher and SRE-04 data sets, and we give results on these to illustrate certain contrasts that guided our development. Final systems are then tested on all data sets.

4.2. MLLR system results

MLLR systems based on the more detailed (8) transforms from the second adaptation step give the best results, based on our previous experiments. Table 3 summarizes results on the data sets used for development, in terms of both minimum detection cost function (DCF)[2] and EER. The four rows of results correspond to the original MLLR system that is affected by the gender mismatch problem, a system based on "male" transforms, a system based on "female" transforms, and one in which both kinds of features are concatenated.

It is clear that fixing the gender mismatch problem reduces overall equal error rates substantially, by 31% to 35% for the SRE-04 data sets. (The mismatch problem is more severe on Mixer data than on Fisher, due to higher gender classification error rates.) DCFs are reduced by 12% to 33% relative on SRE-04. Male and female transforms give approximately equal results. Combining the two sets of transforms yields an additional gain. EERs on SRE-04 are reduced by an additional 12% to

[2]DCF is the Bayes risk function defined by NIST with $P_{\text{target}} = 0.1$, $C_{\text{fa}} = 1$, and $C_{\text{miss}} = 10$.

Table 5: Speaker verification results using baseline, MLLR, and combined systems. The MLLR SVM systems uses 8+8 transforms (same as last row in Table 3) or 2+2 transforms (last row in Table 4). The last row represents a three-way combined system.

	SWB-II		Fisher	SRE-04		SRE-05	
System	1-side	8-side	1-side	1-side	8-side	1-side	8-side
MFCC GMM	*4.63* .17857	*1.92* .08353	*4.57* .10259	*7.77* .31126	*4.95* .21146	*7.17* .24781	*4.91* .16886
MFCC SVM	*4.38* .15610	*1.06* .04470	*4.31* .11051	*8.01* .31339	*3.33* .12629	*7.26* .26839	*3.05* .10333
MLLR (2+2) SVM	*4.72* .18130	*1.12* .0.6387	*2.95* .05756	*8.22* .33962	*4.37* .16283	*7.91* .29533	*4.49* .15774
MLLR (8+8) SVM	*3.00* .10759	*0.48* .02419	*2.85* .05493	*5.34* .25640	*2.62* .11767	*5.91* .17950	*2.45* .07918
MFCC GMM +MLLR (2+2) SVM						*6.43* .22931	*4.01* .13056
MFCC GMM + MFCC SVM						*5.73* .21485	*3.11* .10279
MFCC GMM +MLLR (8+8) SVM						*4.84* .15209	*2.45* .07095
MFCC SVM +MLLR (8+8) SVM						*4.47* .1578	*2.15* .06591
MFCC (GMM+SVM) +MLLR (8+8) SVM						*4.61* .15044	*2.21* .06332

Table 4: Speaker verification results using MLLR features from 1st adaptation stage (2 transforms). The top number (in italics) in each table cell is the EER (%). The bottom number is the minimum DCF value.

	Fisher	SRE-04	
MLLR gender	1-side	1-side	8-side
Mixed (old system)	*6.37* .09934	*12.38* .41594	*6.12* .19934
Male	*3.45* .06820	*9.70* .37794	*4.66* .18342
Female	*3.12* .06153	*9.07* .38666	*5.10* .17158
Male + Female	*2.98* .05756	*8.22* .33962	*4.37* .16283

21% relative, and DCFs by 2% to 20%.

We also ran a similar set of experiments using the two transforms used in the first adaptation step of the recognizer. Note that computing these transforms does not require a full recognition (decoding) step and is therefore computationally inexpensive. Results, shown in Table 4, are qualitatively similar to those found with the more detailed MLLR system. It is remarkable that on the Fisher set (which is better matched to the background set), the simpler MLLR system does almost as well. Under mismatched conditions (SRE-04), however, the simpler MLLR system shows about 50% higher EERs.

In [1] we had reported that additional gains could be obtained by combining the MLLR transforms from the first and second adaptation step into a single extended feature vector. This is no longer true in our revised system after the male and female transforms are combined. Consequently, the step-2 MLLR system with 8+8 transforms is the system of choice for now, assuming a full recognizer can be run. The step-1 MLLR system is still of interest, for example, when a full recognizer is too costly to run, or for mixed-language speaker verification.

4.3. Baseline system comparison and combination

The top part of Table 5 gives complete results for our two cepstral baseline systems, as well as the MLLR systems using 2+2 or 8+8 transforms. We observe that the results across all data sets are quite consistent, and, in particular, SRE-05 results are very similar to those on SRE-04. The cepstral GMM is competitive with the cepstral SVM in the 1-side training condition, but falls significantly behind the two SVM systems in the 8-side condition. Interestingly, the 2+2-transform MLLR system is competitive with the MFCC GMM system, and beats it in the 8-side condition.

The middle and bottom parts of Table 5 shows results with combinations of the two MFCC baseline systems with the MLLR systems, using a neural network for combining the system output scores. The combiner is trained to minimize DCF on the SRE-04 data sets. Figures 1 and 2 plot the detection error tradeoffs for the two baselines, the 8+8 MLLR system, as well as for the best combined system.

As shown in the bottom part of the table, combining the 8+8-transform MLLR system with one of the cepstral systems generally yields sizeable improvements over the MLLR system by itself. By contrast, a combination of the two baseline systems yields a much smaller error reduction over the individual baselines, showing that system combination *per se* is not sufficient to obtain good results, and that the MLLR system contributes information that complements the baselines. A three-way combination does, however, improve over the best two-way system, yielding 22% (for 1-side training) and 10% (for 8-side training) relative EER reduction over the MLLR system by itself.

The middle row in Table 5 shows that even the 2+2-transform MLLR systems can boost the accuracy of a GMM baseline system signficantly when combine with the latter. This might be of interest if full word recognition is not an option, as transforms here are computed using only a simple phone-loop

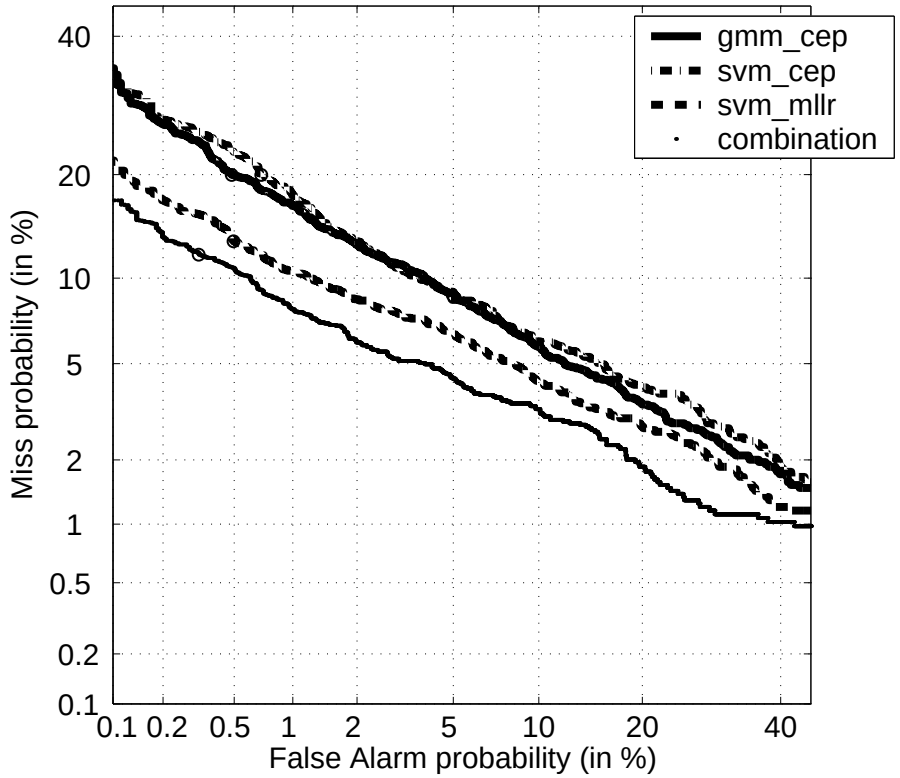

Figure 1: Detection error tradeoff (DET) curves for baseline, 8+8-transform MLLR, and combined systems, SRE-05 1-conversation-side condition.

decoding pass.

5. Conclusions and Future Work

We have discovered and corrected a major problem in our previous implementation of MLLR-transform-based speaker recognition, which was the result of gender classification errors in the speech recognition system. We fixed the resulting gender mismatch in speaker trials, and came upon a further improvement by combining transforms relative to multiple recognition models (male and female, in our case). Our improved MLLR-based SVM speaker verification system now gives markedly better results than our current cepstral GMM and SVM systems. Combining the MLLR systems with these baseline systems gives a significant further error reduction.

We plan to investigate MLLR for speech recognition in other languages, including the case where a trial involves a mix of several languages.

6. Acknowledgments

We thank the anonymous reviewers for valuable suggestions, Ramana Gadde for clarifications of the MLLR implementation, and our colleagues Liz Shriberg, Kemal Sonmez, and Andy Hatch for fruitful discussions. This work was funded by a DoD KDD award via NSF IRI-9619921. The views herein are those of the authors and do not reflect the views of the funding agencies.

7. References

[1] 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.

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

[3] A. Stolcke, H. Franco, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, Y. Huang, B. Peskin, I. Bulyko, M. Ostendorf, and K. Kirchhoff, "Speech-to-text research at SRI-ICSI-UW", *in DARPA RT-03 Workshop*, Boston, May 2003, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/sri+-rt03-stt.pdf.

[4] S. S. Kajarekar, L. Ferrer, E. S. 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.

[5] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubala, "Fast robust inverse transform SAT and multi-stage adaptation", *in Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 105–109, Lansdowne, VA, Feb. 1998. Morgan Kaufmann.

[6] K. Morik, P. Brockhausen, and T. Joachims, "Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring", in I. Bratko and S. Dzeroski, editors, *Proceedings of the 16th International*

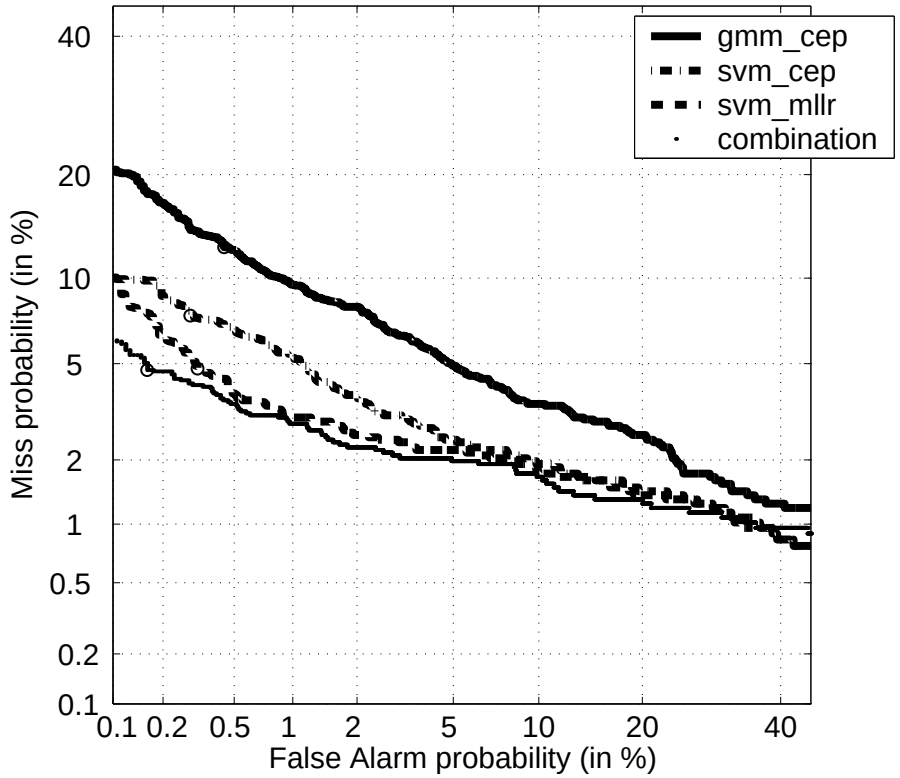

Figure 2: Detection error tradeoff (DET) curves for baseline, 8+8-transform MLLR, and combined systems, SRE-05 8-conversation side condition.

Conference on Machine Learning, pp. 268–277, San Francisco, CA, 1999. Morgan Kaufmann.

[7] A. O. Hatch, A. Stolcke, and B. Peskin, "Combining feature sets with support vector machines: Application to speaker recognition", *in Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 75–79, San Juan, Puerto Rico, Nov. 2005.

[8] A. O. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition", *in Proc. ICASSP*, Toulouse, May 2006.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[10] D. A. Reynolds, "Channel robust speaker verification via channel mapping", *in Proc. ICASSP*, vol. 2, pp. 53–56, Hong Kong, Apr. 2003.

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

[12] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition", *in Proc. ICASSP*, vol. 1, pp. 161–164, Orlando, FL, May 2002.

[13] S. S. Kajarekar, "Four weightings and a fusion: A cepstral-SVM system for speaker recognition", *in Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 17–22, San Juan, Puerto Rico, Nov. 2005.

[14] 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*, Toulouse, May 2006.

[15] A. Martin, D. Miller, M. Przybocki, J. Campbell, and H. Nakasone, "Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004", *in Proceedings 4th International Conference on Language Resources and Evaluation*, pp. 587–590, Lisbon, May 2004.