Speaker Recognition Via Nonlinear Discriminant Features

Lara Stoll[1,2], *Joe Frankel*[1,3], *Nikki Mirghafori*[1]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]University of California at Berkeley, USA
[3]Centre for Speech Technology Research, Edinburgh, UK
{lstoll,nikki}@icsi.berkeley.edu, joe@cstr.ed.ac.uk

Abstract

We use a multi-layer perceptron (MLP) to transform cepstral features into features better suited for speaker recognition. Two types of MLP output targets are considered: phones (Tandem/HATS-MLP) and speakers (Speaker-MLP). In the former case, output activations are used as features in a GMM speaker recognition system, while for the latter, hidden activations are used as features in an SVM system. Using a smaller set of MLP training speakers, chosen through clustering, yields system performance similar to that of a Speaker-MLP trained with many more speakers. For the NIST Speaker Recognition Evaluation 2004, both Tandem/HATS-GMM and Speaker-SVM systems improve upon a basic GMM baseline, but are unable to contribute in a score-level combination with a state-of-the-art GMM system. It may be that the application of normalizations and channel compensation techniques to the current state-of-the-art GMM has reduced channel mismatch errors to the point that contributions of the MLP systems are no longer additive.

1. Introduction

The speaker recognition task is that of deciding whether or not a (previously unseen) test utterance belongs to a given target speaker, for whom there is only a limited amount of training data available. The traditionally successful approach to speaker recognition uses low-level cepstral features extracted from speech in a Gaussian mixture model (GMM) system. Although cepstral features have proven to be the most successful choice of low-level features for speech processing, discriminatively trained features may be better suited to the speaker recognition problem. We utilize multi-layer perceptrons (MLPs), which are trained to distinguish between either phones or speakers, as a means of performing a feature transformation of cepstral features.

There are two types of previous work that are directly related to our research, both involving the development of discriminative features. In the phonetically discriminative case, the use of features generated by one or more MLPs trained to distinguish between phones has been shown to improve performance for automatic speech recognition (ASR). At ICSI, Zhu and Chen, et al. developed what they termed Tandem/HATS-MLP features, which incorporate longer term temporal information through the use of MLPs whose outputs are phone posteriors [1, 2].

In the area of speaker recognition, Heck and Konig, et al. focused on extracting speaker discriminative features from MFCCs using an MLP [3, 4]. They used the outputs from the middle layer of a 5-layer MLP, which was trained to discriminate between speakers, as features in a GMM speaker recognition system. The MLP features, when combined on the score-level with a cepstral GMM system, yielded consistent improvement when the training data and testing data were collected from mismatched telephone handsets [3]. A similar approach was followed by Morris and Wu, et al.[5]. They found that speaker identification performance improved as more speakers were used to train the MLP, up to a certain limit [6].

In the phonetic space, we use the Tandem/HATS-MLP features in a GMM speaker recognition system. The idea is that we can use the phonetic information of a speaker in order to distinguish that speaker from others.

In the speaker space, we train 3-layer Speaker-MLPs of varying sizes to discriminate between a set of speakers, and then use the hidden activations as features for a support vector machine (SVM) speaker recognition system. The intuition behind this method is that the hidden activations from the Speaker-MLP represent a nonlinear mapping of the input cepstral features into a general set of speaker patterns. Our Speaker-MLPs are on a larger scale than any previous work: we use more training speakers, training data, and input frames of cepstral features, and larger networks.

To begin, Section 2 outlines the experimental setup. The results of our experiments are reported in Section 3. Finally, we end with discussion and conclusions in Section 4.

2. Experiments

2.1. Overall Setup

The basic setups of the Tandem/HATS-GMM and Speaker-SVM systems are shown in Figures 1 and 2, respectively. Frames of perceptual linear prediction (PLP) coefficients, as well as frames of critical band energies in the former case, are the inputs to the MLPs. A log is applied to either the output or hidden activations, and after either dimensionality reduction or calculation of mean, standard deviation, histograms, and percentiles, the final features are used in a speaker recognition system (GMM or SVM).

2.2. Baseline GMM Systems

We make use of two types of GMM baselines for purposes of comparison. The first is a state-of-the-art GMM system, which was developed by our colleagues at SRI, and which we will refer to as SRI-GMM [7]. It utilizes 2048 Gaussians, CMS, T-norm, H-norm, and channel mapping to improve its results. We use this system for score-level combinations, in which the scores from SRI's GMM system are combined with the scores from our MLP features systems. For more details, see Section 2.6.

The second system, on the other hand, is a very basic GMM system, with 256 Gaussians, and which includes only CMS, without any other normalizations. This system, which we will

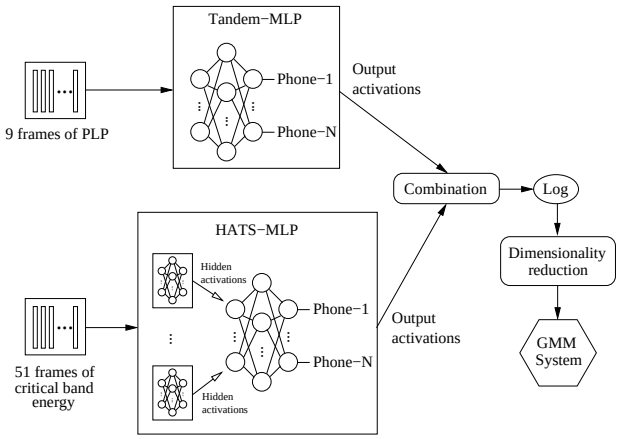

Figure 1: *Tandem/HATS-GMM System*

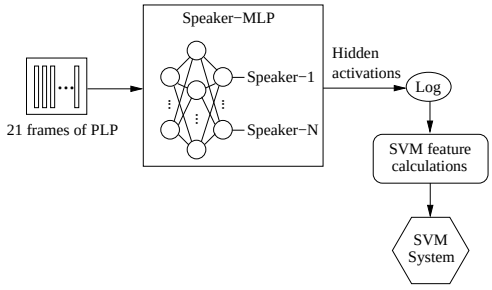

Figure 2: *Speaker-SVM System*

refer to as Basic-GMM, is useful for the purpose of feature-level combination (where we use MFCC features augmented with MLP features as features in the GMM system), as well as for score-level combination.

2.3. Tandem/HATS-MLP Features

There are two components to the Tandem/HATS-MLP features, namely the Tandem-MLP and the HATS-MLP. The Tandem-MLP is a single 3-layer MLP, which takes as input 9 frames of PLPs (12th order plus energy) with deltas and double-deltas, contains 20,800 units in its hidden layer, and has 46 outputs, corresponding to phone posteriors. The hidden layer applies the sigmoid function, while the output uses softmax.

The HATS-MLP is actually two stages of MLPs that perform phonetic classification with long-term (500-1000 ms) information. The first stage MLPs take as input 51 frames of log critical band energies (LCBE), with one MLP for each of the 15 critical bands; each MLP has 60 hidden units (with sigmoid applied), and the output layer has 46 units (with softmax) corresponding to phones. For the HATS (Hidden Activation TRAPS) features, the hidden layer outputs are taken from each first-stage critical band MLP, and then input to the second-stage merger MLP, which contains 750 hidden units, and 46 output units.

The Tandem-MLP and HATS-MLP features are then combined using a weighted sum, where the weights are a normalized version of inverse entropy. The log is applied to the output, and a Karhunen-Loeve Transform (KLT) dimensionality reduction is applied to reduce the output feature vector to an experimentally determined optimal length of 25. This process is illustrated in Figure 1.

The Tandem/HATS-MLP system is trained on roughly 1800 hours of conversational speech from the Fisher [8] and Switchboard [9] corpora.

2.4. Speaker-MLP Features

2.4.1. Speaker Target Selection Through Clustering

As a contrast to using all speakers with enough training data available (with the idea that including more training speakers will yield better results), we also implemented MLPs trained using only subsets of specifically chosen speakers. These speakers were chosen through clustering in the following way. First, a background GMM model was trained using 286 speakers from the Fisher corpus. Then, a GMM was adapted from the background model with the data from each MLP-training speaker. These GMMs used 32 Gaussians, with input features of 12th order MFCCs plus energy and their first order derivatives. The length-26 mean vectors of each Gaussian were concatenated to form a length-832 feature vector for each speaker. Principal component analysis was performed, keeping the top 16 dimensions of each feature vector (accounting for 68% of the total variance). In this reduced-dimensionality speaker space, k-means clustering was done, using the Euclidean distance between speakers, for $k = 64$ and $k = 128$. Finally, the sets of 64 and 128 speakers were chosen by selecting the speaker closest to each of the (64 or 128) cluster centroids.

2.4.2. MLP Training

A set of 64, 128, or 836 speakers was used to train each Speaker-MLP, with 6 conversation sides per speaker used for training, and 2 for cross-validation (CV). The training speaker data came from the Switchboard-2 corpus [9]. The set of 836 speakers included all speakers in the Switchboard2 corpus with at least 8 conversations available. The smaller sets of speakers, selected through clustering, used training and CV data that was balanced in terms of handsets.

ICSI's QuickNet MLP training tool [10] was used to train the Speaker-MLPs. The input to each Speaker-MLP is 21 frames of PLPs (12th order plus energy) with first and second order derivatives appended. The hidden layer applies a sigmoid, while the output uses softmax.

Table 2 shows the sizes of MLPs (varying in the number of hidden units) trained for each set of speakers.

2.5. SVM Speaker Recognition System

The GMM system is well suited to modeling features with fewer than 100 dimensions. However, problems of data sparsity and singular covariance matrices soon arise in trying to estimate high dimensional Gaussians. Previous work in speech recognition (HATS) has shown that there is a great deal of information in the hidden structure of the MLP. Preliminary experiments also showed that reducing the dimensionality of the hidden activations using principal component analysis (PCA) or linear discriminant analysis (LDA), so that the features could be used in a GMM system, yielded poor results. In order to take advantage of the speaker discriminative information in the hidden activations of the Speaker-MLPs, we use an SVM speaker recognition system, which is better suited to handle the high dimensional sparse features, is naturally discriminative in the way it is posed, and has proven useful in other approaches to speaker verification.

Since the SVM speaker recognition system requires the same length feature vector for each speaker (whether a target, an impostor, or a test speaker), we produce a set of statistics

to summarize the information along each dimension of the hidden activations. These statistics (mean, standard deviation, histograms of varying numbers of bins, and percentiles) are then used as the SVM features for each speaker. For our experiments, the set of impostor speakers used in the SVM system is a set of 286 speakers from the Fisher corpus designed to be balanced in terms of gender, channel, and other conditions.

2.6. System Combinations Using LNKnet

In order to improve upon the baseline of the SRI-GMM system, we choose to combine our various systems on the score-level with the SRI-GMM, using LNKnet software [11]. We use a neural network with no hidden layer and sigmoid output nonlinearity, which takes two or more sets of likelihood scores as input. We use a round-robin approach and divide our test data into two subsets for development and evaluation.

3. Results

3.1. Testing Database

In order to compare the performance of our systems, we use the database released by NIST for the 2004 Speaker Recognition Evaluation (SRE) [12]. This database consists of conversational speech collected in the Mixer project, and includes various languages and various channel types. We use only telephone data, containing a variety of handsets and microphones.

One conversation side (roughly 2.5 minutes) is used for both the training of each target speaker model and the testing of each test speaker. As performance measures, we use the detection cost function (DCF) of the NIST evaluation and the equal error rate (EER). The DCF is defined to be a weighted sum of the miss and false alarm error probabilities, while the EER is the rate at which these error probabilites are equal.

3.2. Tandem/HATS-GMM

For NIST's SRE2004, the DCF and EER results are given in Table 1 for the Basic-GMM system, the Tandem/HATS-GMM system, and their score- and feature-level combinations, as well as for the SRI-GMM system and its combination with the Tandem/HATS-GMM. Changes relative to each baseline (where a positive value indicates improvement) are shown in parentheses.

Alone, the Tandem/HATS-GMM system performs slightly better than the Basic-GMM system. Feature-level combination of MFCC and Tandem/HATS features in a GMM system, as well as score-level combination of the Tandem/HATS-GMM system with the Basic-GMM, both yield significant improvements. When the Tandem/HATS-GMM system is combined on the score-level with the SRI-GMM system, there is no gain in performance over the SRI-GMM alone.

	DCF×10	EER (%)
Basic-GMM	0.724	18.48
Tandem/HATS-GMM	0.713 (2%)	18.48 (0%)
Score-level fusion	0.618 (15%)	16.26 (12%)
Feature-level fusion	0.601 (17%)	16.35 (12%)
SRI-GMM	0.374	9.01
Tandem-GMM	0.713	18.48
Score-level fusion	0.378 (-1%)	9.09 (-1%)

Table 1: *Tandem/HATS-GMM system improves upon Basic-GMM system, especially in combination, but there is no improvement for SRI-GMM system.*

3.3. Speaker-SVM

Both the cross-validation and SRE2004 results for the Speaker-MLPs are shown in Table 2 for each size MLP. It is clear that the CV accuracy increases with respect to the number of hidden units, for each training speaker set. The accuracy increase on adding further hidden units does not appear to have reached a plateau at 2500 hidden units for the 836 speaker net, though for the purposes of the current study the training times became prohibitive. With the computation shared between 4 CPUs, it took over 4 weeks to train the MLP with 2500 hidden units.

Similar to the CV accuracy, the speaker recognition results improve with an increase in the size of the hidden layer when considering a given number of training speakers.

# spkrs	Hid. units	CV acc.	DCF×10	EER (%)
64	400	37.8%	0.753	21.04
64	1000	47.8%	0.715	20.41
128	1000	39.4%	0.702	20.45
128	2000	44.5%	0.691	19.70
836	400	20.5%	0.756	22.88
836	800	25.5%	0.734	21.37
836	1500	32.0%	0.711	20.45
836	2500	35.5%	0.689	19.91

Table 2: *Speaker-SVM results improve as the number of hidden units, as well as the CV accuracy, increase.*

In Table 3, the results are given for the score-level combination of the 64 speaker, 1000 hidden unit, Speaker-SVM system with the Basic-GMM and SRI-GMM systems. For the SRI-GMM, the best combination is yielded when the Speaker-MLP is trained with 64 speakers and 1000 hidden units (although the 128 speakers with 2000 hidden units does somewhat better in combination with the Basic-GMM). There is a reasonable gain made when combining the Speaker-SVM system with the Basic-GMM, but there is no significant improvement for the combination of the Speaker-SVM and SRI-GMM systems.

	DCF×10	EER (%)
Basic-GMM	0.724	18.48
Speaker-SVM	0.715	20.41
Score-level fusion	0.671 (7%)	17.52 (5%)
SRI-GMM	0.374	9.01
Speaker-SVM	0.715	20.41
Score-level fusion	0.373 (0%)	9.01 (0%)

Table 3: *System combination with 64 speaker, 1000 hidden unit, Speaker-SVM improves Basic-GMM results, but not the SRI-GMM.*

3.4. Mismatched Train and Test Conditions

We now consider matched (same gender and handset) and mismatched (different gender or handset) conditions between the training and test data. Such a breakdown is given in Table 4 for both the Tandem/HATS-GMM and Speaker-SVM systems and their score-level combinations with the Basic-GMM and SRI-GMM. For each combination, changes relative to the appropriate baseline system are given in parentheses.

When considering a score-level fusion with the Basic-GMM system, gains are made in the matched and especially the mismatched conditions for both the Tandem/HATS-GMM and Speaker-SVM. For the SRI-GMM baseline, combination with the Tandem/HATS-GMM and Speaker-SVM systems has marginal impact in either the matched or mismatched case.

	System Alone		Fusion with Basic-GMM		Fusion with SRI-GMM	
	Matched EER (%)	Mismatched EER (%)	Matched EER (%)	Mismatched EER (%)	Matched EER (%)	Mismatched EER (%)
Basic-GMM	9.13	22.65	–	–	–	–
SRI-GMM	5.74	10.71	–	–	–	–
Tandem/HATS-GMM	12.53	21.54	8.67 (5%)	19.84 (12%)	5.74 (0%)	10.77 (-1%)
Speaker-SVM (1000hu, 64ou)	13.93	23.56	8.78 (4%)	20.95 (7%)	5.74 (0%)	10.64 (1%)

Table 4: *Breakdown of results for matched and mismatched conditions for the MLP-based systems and their score-level fusions with the Basic-GMM and SRI-GMM.*

4. Discussion and conclusions

For the first time, phonetic Tandem/HATS-MLP features were tested in a speaker recognition application. Although developed for ASR, the Tandem/HATS-MLP features still yield good results for a speaker recognition task, and in fact perform better than a basic cepstral GMM system; even more improvement comes from score- and feature-level combinations of the two.

Prior related work used discriminative features from MLPs trained to distinguish between speakers. Motivated by having a well-established infrastructure for neural network training at ICSI, we felt that there was potential for making greater gains by using more speakers, more hidden units, and a larger contextual window of cepstral features at the input. Even though preliminary experiments confirmed this, ultimately, however, a smaller subset of speakers chosen through clustering proved similar in performance and could be trained in less time.

Although the MLP-based systems do not improve upon the SRI-GMM baseline in combination, this result could be explained by considering the difference in the performance between the two types of systems: standalone, each MLP-based system performs much more poorly than the SRI-GMM. The addition of channel compensating normalizations, like T-norm [13], to an MLP-based system should help reduce the performance gap between the MLP-based system and the SRI-GMM. It may then be possible for the MLP-based system to improve upon the state-of-the-art cepstral GMM system in combination, in the event that the performance gap is narrowed sufficiently.

Similar to results observed in prior work, the Speaker-SVM system improved speaker recognition performance for a cepstral GMM system lacking sophisticated normalizations (such as feature mapping [14], speaker model synthesis (SMS) [15], and T-norm); such a result was also true for the Tandem/HATS-GMM system. However, no gains were visible in addition with the SRI-GMM, which is significantly improved from the Basic-GMM (as well as the GMM systems of Wu and Morris, et al. and Heck and Konig, et al.) by the addition of feature mapping, T-norm, as well as increasing the number of Gaussians to 2048.

As shown in Table 4, combinations of the Basic-GMM with the phonetic- and speaker-discriminant MLP-based systems of this paper do yield larger improvements for the mismatched condition (which refers to the training data and test data being different genders or different handset types). However, such a result does not hold for combinations of the MLP-based systems with the SRI-GMM. The previous work of Heck and Konig, et al., completed prior to the year 2000, showed that the greatest strength of an MLP-based approach was for the case when there is a handset mismatch between the training and test data, however, the state-of-the-art has since advanced significantly in normalization and channel compensation techniques. As a result, the contributions of the MLP-based systems, without any normalizations applied, to a state-of-the-art cepstral GMM system are no longer significant for the mismatched condition.

5. Acknowledgements

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship and upon work supported by the National Science Foundation under grant number 0329258. This work was also made possible by funding from the EPSRC Grant GR/S21281/01 and the AMI Training Programme. We would also like to thank our colleagues at ICSI and SRI.

6. References

[1] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *ICSLP*, 2004.

[2] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," in *ICSLP*, 2004.

[3] L. P. Heck, Y. Konig, M. K. Sönmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communications*, vol. 31, no. 2-3, pp. 181–192, 2000.

[4] Y. Konig, L. Heck, M. Weintraub, and K. Sönmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," in *Proceedings of RLA2C - Speaker Recognition and Its Commercial and Forensic Applications*, Avignon, France, 1998.

[5] A. C. Morris, D. Wu, and J. Koreman, "MLP trained to separate problem speakers provides improved features for speaker identification," in *IEEE Int. Carnahan Conf. on Security Technology*, 2005.

[6] D. Wu, A. Morris, and J. Koreman, "MLP internal representation as discriminative features for improved speaker recognition," in *Proc. NOLISP*, 2005, pp. 25–33.

[7] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sönmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST speaker recognition evaluation system," in *ICASSP*, vol. 1, 2005, pp. 173–176.

[8] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech to text," in *LREC*, 2004, pp. 69–71.

[9] Linguistic Data Consortium, "Switchboard-2 corpora," http://www.ldc.upenn.edu.

[10] D. Johnson, "QuickNet3," http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[11] MIT Lincoln Labs, "LNKNet," http://www.ll.mit.edu/IST/lnknet, 2005.

[12] National Institute of Standards and Technology, "The NIST year 2004 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf, 2004.

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

[14] D. Reynolds, "Channel robust speaker verification via feature mapping," in *ICASSP*, 2003.

[15] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *ICSLP*, 2000.