MLP TRAINED TO SEPARATE PROBLEM SPEAKERS PROVIDES IMPROVED FEATURES FOR SPEAKER IDENTIFICATION

Andrew C. Morris , Dalei Wu & Jacques Koreman

{amorris, daleiwu, jkoreman}@coli.uni-saarland.de
Saarland University, Institute of Phonetics, Saarbrücken, Germany

ABSTRACT

In automatic speech recognition (ASR) the non-linear data projection provided by a one hidden layer multi-layer perceptron (MLP), trained to recognise phonemes, has previously been shown to provide feature enhancement which can substantially increase ASR performance, especially in noise. Previous attempts to apply an analogous approach to speaker identification have not succeeded in improving performance, except by combining MLP processed features with other features. We present test results for the TIMIT database which show that the advantage of MLP preprocessing for open set speaker identification increases with the number of speakers used to train the MLP and that improved identification is obtained as this number increases beyond sixty. We also present a method for selecting the speakers used for MLP training which further improves identification performance.

Keywords: Speaker recognition, data enhancement, MLP

1. INTRODUCTION

It has previously been shown that the projection provided by the pre-squashed outputs from a one hidden layer MLP [1], pre-trained to output a probability for each phoneme, can significantly increase automatic speech recognition (ASR) performance [9]. In attempting to apply the same technique to speaker (rather than speech) recognition, a number of questions arise. What target classes should the MLP be trained to recognise if we want the features it generates to provide enhanced discrimination between speakers? If the MLP is trained to recognise some closed subset of speakers, would the mapping learnt also provide discriminative features for speakers not seen during training? The number of classes which an MLP can successfully learn to separate with a manageable amount of training data is quite limited. If speech data is available for a large number of speakers, which subset of these speakers would be most effective for MLP training? In Section 2 we briefly outline the models most commonly used for speaker recognition. In Section 3 we discuss related research in the area of MLP enhancement for speaker recognition, introducing the MLP architecture which was found to be most effective for this purpose [6]. In Section 4 we present the method we have developed for selecting a suitable speaker subset whose speech data is to be used for MLP training. In Section 5 we present the model details and experimental design which we use to test the effect on identification performance of different MLP architectures, different numbers of training speakers and random versus systematic training speaker selection. Section 6 presents the test results. This is followed by a discussion and conclusion in Sections 7 and 8.

2. SPEAKER RECOGNITION BASELINE

The speaker recognition problem may be one of identification or verification. Given a certain amount of preprocessed speech data X, in the case of identification the problem is to identify the speaker from some given set of speakers, while with verification it is to decide whether or not the speaker is who they claim to be.

2.1. Speaker identification

A Gaussian mixture model (GMM), with some fixed number of Gaussians and diagonal covariance, is trained to model the speech frame pdf for each speaker [7]. When training data is very limited it can be advantageous to train each GMM by MAP mean adaptation from a universal background model (UBM). Speaker identification is then performed by selecting the speaker S_j with the largest posterior probability, $P(S_j|X)$ (which corresponds to the largest data likelihood $p(X|\ S_j)$ if all speaker priors $P(S_j)$ are equal).

For identification the ideal speech features must be independent of the true speaker identity, which is not known. Speaker independent feature enhancement is therefore well suited to speaker identification.

2.2. Speaker verification

For speaker verification a GMM is trained for each speaker as with identification, but the claimant is accepted

iff the likelihood ratio of $p(X|S_j)/p(X|U)$ exceeds some fixed threshold, where $p(X|U)$ is the UBM which models the likelihood that X is from any speaker but S_j.

For verification the problem is to distinguish a given speaker from all other speakers, so the optimal feature enhancement may be speaker dependent and therefore not so well suited to the approach used here, although this is the application to which it was first applied.

3. RELATED RESEARCH

The method of training an MLP to classify a small subset of speakers was first investigated in [5, 6] in the context of speaker verification with the NIST 1998 SRE database. It was found that preprocessing by an MLP with just one hidden layer gave no advantage. This could be because speaker data, being clustered around every phoneme, requires more highly non-linear separation than speech data. However, the separating power of an MLP can be increased by using more hidden layers. In [6] an MLP with three hidden layers was trained to recognise 35 speakers and discriminative features were taken from the pre-squashed outputs from the central bottle-neck hidden layer (see Fig.1). In this case it was found that these features still did not provide performance enhancement except when concatenated with the original MFCC features.

In initial tests with TIMIT in [10] and noisy TIMIT in [11] we showed that the performance of the features provided by this MLP architecture increases with the number of speakers which the MLP is trained to separate. Due to various factors, test results are subject to significant random fluctuation. The tests we report here show that open set identification improves over the MFCC baseline when using data from upwards of 128 speakers.

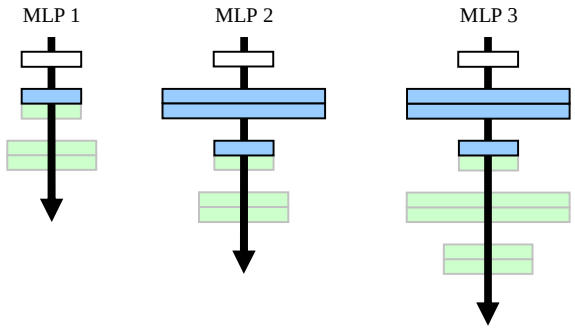

Fig. 1.Three MLP types tested for data enhancement. Each active layer is shown as a (net-input function / non-linear activation function) sandwich. Light sections are used only in MLP training, not for enhancement.

In [10] MLPs with from 1 to 3 hidden layers were tested (see Figure 1) and best identification performance was obtained taking enhanced features as the net-input values to the second hidden layer from an MLP which was trained with 3 hidden layers. Single frames were used for MLP training instead of the input vector of 9 concatenated frames (100 ms) which works best in ASR. These are the three MLPs we test in Section 6.

4. SPEAKER BASIS SELECTION

A random selection of the speaker subset which the MLP is trained to separate (which we call the speaker basis) would be expected to represent the open set speaker population. However, classifier training can be more effective when training data is selected from class boundaries, while many errors in speaker identification are often traceable to a small number of problem speakers [3].

We have tested several strategies for speaker basis selection based on a matrix of the distances between each speaker GMM pdf. We show here that this distance matrix can be estimated using only the speaker posterior probabilities $P_{ji} = P(S_j|X_i)$ for a set of development test data. P_{ji} are obtained by dividing the development data log likelihood for each speaker by their sum over all speakers for one utterance.

As a distance measure between speaker pdfs we use the symmetric Kullback-Leibler distance $KL(S_j, S_k)$ (1). This cannot be evaluated in closed form when $p(X|S_j)$ is modelled by a GMM. However, provided $P(S_j)=P(S_k)$,

$$KL(S_j,S_k)=\int\left(p(X\mid S_j)-p(X\mid S_k)\right)\log\frac{p(X\mid S_j)}{p(X\mid S_k)}dX \quad (1)$$

$$\propto \int p(X)\left(P(S_j\mid X)-P(S_k\mid X)\right)\log\frac{P(S_j\mid X)}{P(S_k\mid X)}dX \quad (2)$$

$$=\int p(X)K(S_j,S_k,X)dX=E\left[K(S_j,S_k,X)\right] \quad (3)$$

$KL(S_j, S_k)$ can therefore be estimated by averaging $K(S_j, S_k, X)$ over the development test data.

$$KL(S_j,S_k)\cong\sum_{X_i\in DevTestSet}K(S_j,S_k,X_i) \quad (4)$$

$$=\sum_i\left(P_{ji}-P_{ki}\right)\left(\log P_{ji}-\log P_{ki}\right) \quad (5)$$

The resulting speaker-distance matrix KL_{jk} can then be used in various ways to select a subset of speakers for MLP training. Of the methods we have tested, that which has given the best results is to choose speakers in order of decreasing average distance from every other speaker. We refer to this as the Maximum Average Distance (MaxAD) method for speaker basis selection.

5. MLP AND GMM TRAINING

The TIMIT speech database [4] was selected because, although it is only read speech, it is well suited for proof of concept tests and it is well known. As in [8], we first down-sampled TIMIT from 16 kHz to 8 kHz. At 16 kHz our baseline system (as in [8]) obtains 100% correct

speaker identification. However, it is of interest here to work with speech data which is close to telephone quality.

5.1. Baseline feature processing

As in [7,8] we used 20 ms frames and 20 Mel scaled filterbank log power features were extracted every 10 ms, using a Hamming window and a pre-emphasis factor of 0.97. A DCT was then applied to obtain MFCC features, from which the c0 energy coefficient was dropped. Neither silence removal, dynamic features nor cepstral mean subtraction were used, since none of these improved performance with TIMIT.

5.2. Train and test set divisions

The experiments we make are intended to test the use of MLP data enhancement for identification systems which are both speaker and text independent. The standard TIMIT division into training and test data is not suitable for this purpose so we defined our own gender and dialect region balanced division into speaker-disjoint training, development and evaluation sets, comprising 300, 168 and 162 speakers, respectively, which we denote SpkTr, SpkDv and SpkEv. Each of the 630 speakers in TIMIT has 10 utterances which are labelled as belonging to three sentence types: 6 type "X", covering a wide range of acoustic contexts; 3 type "I", being acoustically diverse, and 2 type "A" sentences which were the same for each speaker. We also divided these 10 sentences into disjoint training, development and evaluation sets: SenTr ($SA_{1\text{-}2}$, $SI_{1\text{-}2}$, $SX_{1\text{-}2}$); SenDv (SX_3, SI_3) and SenEv (SX_4, SX_5).

5.3. GMM training

All GMM and MLP training and testing was performed by the Torch machine learning API [2]. GMMs used 32 Gaussians, a variance threshold factor of 0.01 and minimum Gaussian weight of 0.05. With TIMIT MAP adaptation did not help and was not used.

5.4. MLP training

Hidden layer 1 in MLP2 and also layer 3 in MLP3 used 100 sigmoid units. The bottle-neck hidden layer, the net-input values from which comprise the MLP transformed features, always had the same number of units as the input layer (19). The output layer always had N log-softmax units, where N is the number of speakers which the MLP is being trained to recognise. MLP training was on-line, with an initial learning rate of 0.01 with a learning rate decay factor of 0.1. The data in each utterance was first normalised to have zero mean and unit variance. The training objective was maximum cross entropy. Initial MLP tests looked at making use of the learning curve for an MLP development set (SpkBs^SenDv) to decide when to stop iterative MLP training. However, for all MLPs tested the development set error continued to decrease even after several hundred training epochs, while the identification performance of GMMs trained on the resulting MLP preprocessed features always stopped increasing after about 30 training epochs. In all of the tests here, MLP training was stopped after 30 training epochs.

5.5. Feature transformation (MLP and PCA)

MFCC data is first normalised in the same way as in MLP training. For MLP 3 the 19 coefficient MFCC data is then passed through the net-input function and sigmoid functions in hidden layer 1 (100 units) and through the net-input function to hidden layer 2 (19 units). As in [9], this MLPC data is then orthogonalised by PCA projection (onto the unit eigenvectors of the covariance matrix for the MLPC features for the MLP training set SpkBs^SenTr).

5.6. Train and test set procedures

For the purpose of speaker basis selection, a GMM is trained for each speaker S_j in SpkTr on MFCCs for S_j ^SenTr (^ denotes set intersection). Each of these GMMs is then tested on MFCCs for SpkTr^SenDv.

Making use of these test likelihoods, a speaker basis for MLP training, comprising a given number of speakers, N, is selected either at random or by MaxAD (see Section 4) from SpkTr. Denote this SpkBs.

The MLP is trained on MFCCs for SpkBs^SenTr. The trained MLP and PCA matrix is then used to transform the data to be used for GMM training and testing first from MFCCs to MLPCs and then (by PCA) to MLPAs.

A GMM is then trained for each speaker in S_j in SpkEv on MLPA data for S_j^SenTr. Each GMM is then tested using MLPA data for every sentence in SpkEv^SenEv.

6. IDENTIFICATION TESTS

In order to confirm the MLP architecture previously proposed in [6], tests were made with MLPs having 1, 2 and 3 hidden layers. In each case the number of speakers whose data was used for MLP training (the basis size) was varied from 2 to 256, using random selection (see Fig.2). Each test was repeated 10 times because of the randomness introduced either by the random basis selection or, when MaxAD basis selection was used, by the random MLP weights initialisation, or by the random GMM weights initialisation for the GMMs used to set up the inter speaker distance matrix. The baseline GMM score was also subject to this random factor, so this test was also repeated 10 times (and had a % correct variance of 0.48). Fig.2 shows percent correct identification for enhancement using MLPs 1, 2 or 3, against $\log_2$ speaker basis size. Basis selection is random.

Further tests were made to compare the performance of random and MaxAD speaker basis selection, again varying the basis size from 2 to 256 and repeating each test 10 times (see Fig.3).

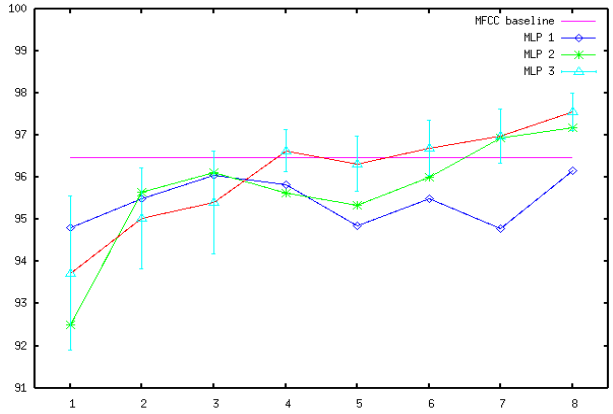

Fig.2. Identification performance for enhancement using MLPs 1, 2 or 3, against $\log_2$ speaker basis size, using random basis selection. Error bars shown only for MLP3.

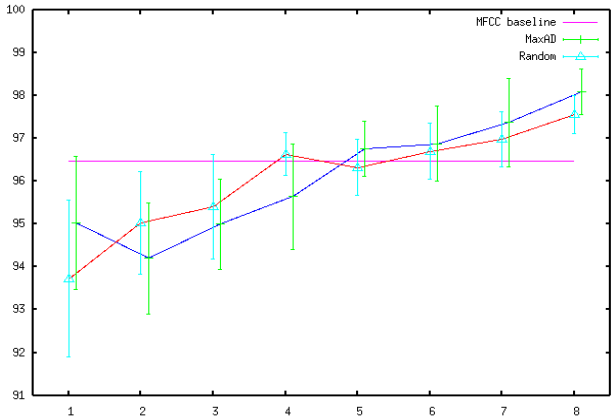

Fig.3. Identification performance for enhancement by MLP 3 using basis selected at random or by MaxAD.

7. DISCUSSION

Speaker identification using MLPAs from all three MLPs improves with the number of speakers used in MLP training, though MLPs with more hidden layers improve more consistently. No significant improvement over the MFCC baseline occurs until the basis size is at least 2^6. It looks as if performance would continue to increase with the basis size going well beyond 2^{10}. MaxAD basis selection significantly outperforms random selection when the basis size is above 2^5. That MaxAD gives better results than random selection even when at 256 selected out of 300 speakers most of the speakers selected must be the same, suggests that it is good at avoiding problem speakers rather than selecting useful speakers.

8. CONCLUSION

Our test results with TIMIT show that MLP based feature enhancement can be used to advantage in speaker identification providing that the data used to train the MLP comes from a large enough number of speakers. Performance cannot continue forever to increase with the number of speakers used for MLP training, but it would be interesting to explore how much further this approach can be taken. Further analysis is needed of the identification errors which occur and of the reasons why some speakers are more useful or harmful for MLP training than others. It remains to be tested whether or not this approach can be successfully applied to more challenging databases.

AKNOWLEDGMENTS

This work was supported by the EC SecurePhone project IST-2002-506883.

REFERENCES

[1] Bishop, C.M., Neural networks for pattern recognition, Oxford University Press, 1995.

[2] Collobert, R., Bengio, S. & Mariéthoz, J., "Torch: a modular machine learning software library", Technical Report IDIAP-RR 02-46, 2002.

[3] Doddington, G., Liggett, W., Martin, A., Przybocki, M. & Reynolds, D., "Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation", Proc. ICSLP, 1998.

[4] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L. and Zue, V., TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

[5] Heck, L., Konig, Y., Kemal Sönmez, M. & Weintraub, M., "Robustness to telephone handset distortion in speaker recognition by discriminative feature design", Speech Communication 31, pp.181-192, 2000.

[6] Konig, Y., Heck, L., Weintraub, M. & Sonmez, K., "Nonlinear discriminant feature extraction for robust text-independent speaker recognition", Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications, pp.72-75, 1998.

[7] Reynolds, D.A., "Speaker identification and verification using Gaussian mixture speaker models", Speech Commun., 17 (1995), pp.91-108, 1995.

[8] Reynolds, D.A., Zissman, M.A., Quatieri, T.F., O'Leary, G.C. & Carlson, B.A. "The effect of telephone transmission degradations on speaker recognition performance", Proc. ICASSP'95, pp.329-332, 1995.

[9] Sharma, S., Ellis, D., Kajarekar, S., Jain, P. & Hermansky, H., "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database", ICASSP 2000, 2000.

[10] Wu, D., Morris, A.C. & Koreman, J., "MLP internal representation as discriminant features for improved speaker recognition", Proc. NOLISP2005, 25-33, 2005.

[11] Wu, D., Morris, A.C. & Koreman, J., "Discriminative features by MLP preprocessing for robust speaker recognition in noise", Proc ESSV 2005, 2005.