Discriminative Features by MLP Preprocessing for Robust Speaker Recognition in Noise

Dalei Wu, Andrew Morris and Jacques Koreman

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

Abstract: Speaker identification performance in noise is compared with that for clean speech. A multi-layer perceptron (MLP) is used to project standard MFCCs onto an internal representation which enhances speaker discrimination. The MLP-enhanced features thus obtained have previously been shown to increase speaker discrimination in clean speech, are now applied to speaker identification in noisy speech conditions. Three different noise types are used at different signal-to-noise ratios. Although a mismatch between the training and test conditions is detrimental to speaker identification performance in noise, the effect can easily be counteracted by training the speaker models on clean speech data as well as data with artificially added noise (multi-condition training). It is shown that MLP-enhanced features outperform standard MFCCs in matched noisy conditions as well as in multi-condition training. For some of the noisy test conditions, the Gaussian mixture speaker models trained on multiple noise conditions even outperform the matched training-test conditions.

1. Introduction

This paper addresses the issue of robust speaker recognition in noise, and in particular investigates the possibility of using a multi-layer perceptron (MLP) to enhance discrimination between speakers. Three different, realistic types of additive noise were selected and added to the 8kHz downsampled TIMIT (TIMIT-8k) clean-speech database [1] at different signal-to-noise ratios (SNRs).

State-of-the-art Gaussian mixture models (GMMs) for speaker recognition, like hidden Markov models (HMMs) in automatic speech recognition (ASR), can achieve very good performance in clean speech, but performance degrades strongly in the presence of noise [2]. ASR performance in noise can be increased significantly by using a feature projection provided by the pre-squashed outputs from a one hidden layer MLP, pre-trained to output a posterior probability for each phoneme [3]. It is not possible to apply an MLP in the same way to speaker recognition. The reason is that in speaker recognition there are no fixed target classes like phonemes in ASR. For the purpose of speaker recognition, the MLP is trained with a representative subset of speakers *(speaker basis)* as its target classes, comparable to phonemes as target classes for ASR. The transformation which the MLP learns for the speaker basis has been shown beneficial for any other speakers from the same population for clean speech [4].

Unlike for ASR, a one-layer MLP applied to speaker recognition in clean speech does not lead to an increase in the percentage correct speaker identification. This may be because speaker data, being clustered around every phoneme, is less easy to partition than speech data. But the separating power of an MLP can be increased by using more hidden layers. In [5,6] an MLP with three hidden layers was trained to recognise 31 speakers, and discriminative features were taken as the outputs from the central, linear bottleneck hidden layer. The 31 speakers were selected because they had been recorded over multiple handsets. It was found that the features obtained from the MLP provide a performance enhancement, although not consistently across all training and test conditions [6] and sometimes only when the feature vector was concatenated with the original MFCC features which were used as input to the MLP [5]. The good results may be due to a better compensation for the different handsets that were used, or to a better separation of the speakers in the acoustic space, even if the speakers were not selected with this aim -- or by a combination of the two.

In tests with TIMIT-8k in [4,7] the automatic selection of the speaker basis to train the MLP with was investigated with the specific aim of making an optimal selection of the target speakers for training the MLP. It was shown that the performance of the features provided by the MLP leads to good speaker identification even for a small number of target speakers used train the MLP. Clearly, the good results must be ascribed to the ability of the MLP to enhance discrimination between speakers, since no variable noise or channel conditions are present. When only a small set of speakers (speaker basis) is used to train the MLP, it is especially important that they are selected so as to represent the whole population. It was shown in [4] that even a speaker basis of 50 speakers automatically selected on the basis of the GMM confusion matrix for 300 training speakers can lead to improved identification of 162 “unseen” test speakers for clean speech.

Here the effect of different types of additive noise, and the ability of the previously applied MLP to enhance speaker discrimination, in matched training and test conditions is investigated and compared with the system performance in clean speech (training and test) conditions. But in many practical applications, there is a mismatch between training and test conditions. Enrollment may take place in fairly clean speech conditions, e.g. when a new user has to go to a registration/certification authority so that his identity can be confirmed on the basis of official documents he possesses (passport, identity card, social security card, etc.). In this case, there will be a mismatch between the training data obtained during enrollment and the test data when the user requests verification in a noisy environment. We not only investigate the effect of a mismatch, where the training data is clean and the test data contain noise, we also evaluate the effect of simply adding several additive noise types to the training data, attempting to deal with the presence of noise in the test data.

In this paper the effectiveness of data enhancement is tested, using the MLP from [4,7] for speaker identification in different kinds of additive noise at different SNRs. Since MLPs have been shown to be able to deal well with noisy speech in ASR [3], we expect they may also enhance speaker recognition when noise is present. In Section 2 we describe the data used in the experiments, the MLP used for feature enhancement and the baseline GMM speaker recognition system. Section 3 then describes the results of the experiments for speaker recognition in clean speech and noise. This is followed by a discussion of the results in Section 4 and conclusions in Section 5.

2. Method

2.1. Data

The TIMIT-8k (clean) speech database is used in all experiments [1]. The reason for choosing this database is that we want to focus our investigations on the separation of speakers in the acoustic space first, and then add noise to evaluate its effect on MLP feature enhancement. Since the standard TIMIT-8k division does not include a development set, we created our own division into speaker-disjoint training, development and evaluation data, with 300, 168 and 162 speakers, respectively. The three sets are selected such that gender and dialect region have an equal proportional representation in the three sets. To make the speaker identification system text-independent, we used all sentences of type SA_{1-2}, SI_{1-2} and SX_{1-2} for training, sentences of type SX_3 and SI_3 for development and sentences of type SX_4 and SX_5 for evaluation. Whereas SA_1 and SA_2 sentences are always the same for different speakers, SI_n and SX_n sentences can be different ones and the index *n* only indicates the order as indicated by the numbers in the TIMIT database. The strict division optimises text-independence of the speaker recognition system.

2.2. Added noise

To evaluate the robustness of discriminative features for speaker identification in various kinds of noise, (stationary) car as well as (non-stationary) factory-1 noise and babble from the NOISEX-92 database [8] were added to the TIMIT-8k database at SNRs of 20, 10 and 0 dB, using the ITU software [9] to determine SNRs. These noises are also used in the Aurora evaluations for speech recognition in noise.

2.3. Feature extraction

The TIMIT speech data was first downsampled from 16 kHz to 8 kHz. At 16 kHz our baseline system (as in [10]) obtains 100% correct speaker identification. However, the interest here is to work with speech data which is close to telephone quality. Using 20ms frames and a 10ms step size, 20 Mel-scaled filterbank log power features were extracted, using a Hamming window and a pre-emphasis factor of 0.97. A DCT was then applied to these to obtain MFCC features, from which the c0 energy coefficient was dropped. Time difference features were not appended, because these did not improve performance with TIMIT-8k. Neither silence removal, dynamic features or cepstral mean subtraction were used, since none of these led to any performance improvement with TIMIT-8k.

2.4. MLP feature enhancement

When an MLP is trained to map speech feature frames onto their phone class probabilities in ASR, not only are the MLP output values useful for deciding which class the speech frame belongs to, but the outputs from its hidden layers within the MLP are also discriminative for the phone classes which were the targets during training. Each unit in a standard MLP has a two stage function. The first stage, the net-input function, is a many-to-one linear combination of the neuron’s inputs. The second stage is a one-to-one non-linear sigmoid function which squashes the net-input to a value between zero and one. From the point of view of using the MLP internal feature representation to provide discriminative features, the squashed outputs are not very suitable because they tend to be close to zero or one, thereby not complying with the GMM assumption that all features have an approximately Gaussian distribution.

Using Torch [11], discriminative preprocessing is carried out for the different noise conditions with the aim of feature enhancement. Each single frame of the standard MFCC features are preprocessed by a 5-layer MLP, as in [5,6]. This MLP was found to outperform MLPs consisting of fewer layers [4]. Training the MLP with single frames instead of the usual input vector of 9 concatenated frames gives the best results for this particular database. The MLP is trained, by gradient descent, to maximise the cross entropy objective (i.e. the mutual information between the actual and target outputs). We trained in batch mode, with a fixed learning rate of 0.01. The data in each utterance was first normalised to have zero mean and unit variance. Of the 3 hidden layers, the first and last hidden layer, which are both non-linear, have 100 units and the middle, linear hidden layer has 19 (compression or bottleneck layer). The features obtained from the compression layer were used as input to a GMM system, as in [4]. The assumption behind this is that this simple representation, which consists of vectors of the same size as the original MFCC vectors, is an internal representation of the acoustic signal which enhances discrimination between the target speakers and is generalisable to the speakers in the entire population. We used the net input to the second hidden layer as input to GMM modelling. The MLP and its application in GMM modelling are represented in Fig. 1. The MLP was trained with a fixed number of iterations (35), after which the error reduction on the training and development data in the MLP frame-based recognition was very small.

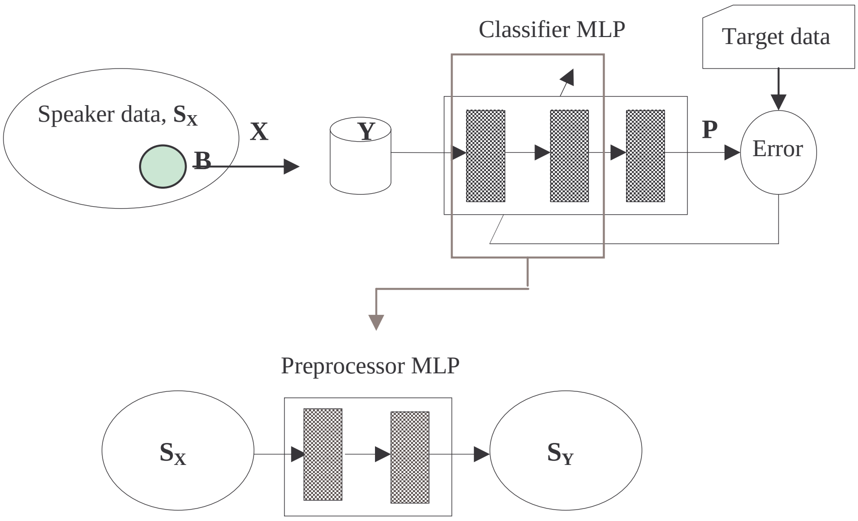

Figure 1: Feature enhancement procedure

Instead of using the 6 training sentences of *all* speakers to train the MLP, it was shown in [4,7] that a smaller selection of speakers, the speaker basis, is sufficient to obtain a good speaker discrimination. Here, a speaker basis consisting of 150 speakers was automatically selected from the training speakers. The automatic selection is made on the basis of the confusion matrix obtained from a GMM experiment. The confusion matrix is produced by classifying the MFCC features from the 2 evaluation sentences of the 300 training speakers with the speaker models for these speakers trained on the 6 training utterances (cf. next section). The log likelihoods in the confusion matrix are converted into likelihoods and then into posterior probabilities, by dividing each value by the row sum. The resulting table of posterior probabilities is then used to select the speaker basis.

The speaker basis which we use to train the MLP is selected using the speaker posterior probabilities $P_{ji} = P(S_j|X_i)$ for a set of development test data. These probabilities are obtained from the test data log likelihoods by dividing the log likelihood for each speaker by their sum over all speakers.

As a distance measure between speaker pdfs we use the symmetric Kullback-Leibler distance $KL(S_j, S_k)$ [12]. This cannot be evaluated in closed form when pdfs $p(X|S_j)$ are modelled by GMMs, but in [7] it is shown that $KL(S_j, S_k)$ is the expected value of K(Sj, Sk, X), where

$$K(S_j, S_k, X) = \left(P(S_j \mid X) - P(S_k \mid X)\right)\left(\log P(S_j \mid X) - \log P(S_k \mid X)\right) \qquad (1)$$

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

$$KL(S_j, S_k) \cong \sum_{X_i \in testSet} K(S_j, S_k, X_i) \ = \sum_i \left(P_{ji} - P_{ki}\right)\left(\log P_{ji} - \log P_{ki}\right) \qquad (2)$$

The resulting speaker-distance matrix can be used in various ways to select a subset of speakers for MLP training. The method which gave the best results for clean data in [4,7] was to choose speakers in order of decreasing average distance from every other speaker.

2.5. GMM modelling

The MFCCs or, alternatively, the enhanced features obtained from the compression layer of the MLP (as explained in the previous section) are used as input to GMM modelling of the diagonal covariances using 32 Gaussians, as in [10,13]. The baseline model trained with MFCCs of the 6 training utterances gives state-of-the-art speaker recognition performance. With TIMIT-8k (though not with other databases, such as the CSLU speaker recognition database) no gain is found in training speaker models by adaptation from a global model for all 300 training speakers, so that each speaker model was trained from scratch with data for that speaker only.

As in [13], GMMs are trained by *k*-means clustering, followed by EM iteration. This is performed by the Torch machine learning API [11], using a variance threshold factor of 0.01 and minimum Gaussian weight of 0.05 (performance falling sharply if either was halved or doubled), determined on the basis of the development sentences of the development speakers.

Test results are obtained for 162 test speakers (for two test sentences per speaker, cf. section 2.1). Speaker identification for utterance feature data X is performed by selecting the speaker S_j with the largest posterior probability, P(S_j|X) (which corresponds here to the largest data likelihood p(X| S_j), as all speaker priors P(S_j) ar equal).

3. Results

In this section, the speaker identification results are compared for the different noises and at different SNRs. The MLP-enhanced features are compared with the baseline system, in which the MFCCs are not preprocessed by the MLP and used as input to GMM directly. Table 1 shows the results for clean data, and for car, factory and babble noise at SNRs of 20, 10 and 0 dB. The conditions presented in Table 1 are all *"matched"* conditions, in which the test data were used with a system trained on data of the same noise type and at the same SNR.

	clean	car			factory			babble			mean
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	
MFCC baseline	3.40	11.42	18.21	38.89	18.83	43.83	87.35	17.90	36.11	87.65	36.36
MLP-enhanced	1.85	6.48	12.96	26.85	15.74	40.43	82.72	10.80	25.00	81.17	30.40

Table 1: Speaker identification error for training on matched noise type and level

The results show a strong increase in speaker recognition error with decreasing SNR. Although still well above chance level, speaker identification is particularly poor for the non-stationary noise types (babble and factory noise) at a SNR of 0 dB. Notice that no cepstral mean subtraction (CMS) was performed, even for the noisy data, which may result in better speaker identification performance. In the clean speech condition CMS leads to poorer performance, probably because subtraction of the spectral mean across an utterance filters out part of the speaker characteristics. As expected, the best results are found for clean speech.

Preprocessing of the MFCCs by an MLP to enhance speaker discrimination *always* reduces the speaker identification error. The absolute reduction is greatest for the stationary car noise, particularly at the lowest SNR. The positive effect, though present in all conditions, is greatly reduced for non-stationary noise types, with the exception of babble at an SNR of 10 dB.

In many applications, the user may want to be recognised in widely varying conditions, but the condition in which he must be recognised is not known beforehand. Two scenarios are possible. In the first scenario, the speaker enrolled in the system in a quiet environment, so that the speaker model (and the MLP) is trained on clean speech. But the actual conditions in which the system is subsequently used may vary from one occasion to the next. In order to evaluate the performance of our system under these *mismatched* conditions, the test data from all the noise conditions in Table 1 were scored with the GMM speaker models (and MLP) trained with clean speech only. (The data for clean speech are the same as in Table 1 and does not represent a mismatch. It is only included in Table 2, because it is used to compute the mean percentage error for correct speaker identification across all possible test conditions in the right-hand column.)

	clean	car			factory			babble			mean
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	
MFCC baseline	3.40	51.23	72.22	93.52	83.95	98.15	99.07	74.38	88.89	97.53	76.23
MLP-enhanced	1.85	79.63	88.27	96.60	95.37	98.46	99.38	91.05	94.44	97.84	84.29

Table 2: Speaker identification error for training on clean speech and testing in various conditions

As the results in Table 2 show, the mismatch between the noisy test data and the clean training data causes a severe deterioration of the performance of the speaker recognition system. For some of the conditions, recognition is only just above chance level (p=100/162=0.62%, i.e. error=99.38%). The effect, which is present for the MFCC features (comparison of first data rows in Tables 1 and 2), is even greater after MLP enhancement, with only chance level speaker identification for factory noise at 0 dB SNR.

In the case of (known) additive noise, the noises can easily be used to create “virtual” data containing this noise before the speaker model is trained. By catering for a variety of testing conditions in the training phase, the system is expected to better cope with the variability in real test conditions. As the results in Table 3 show, the performance in all noisy conditions is substantially better than when the GMM speaker models (and the MLP) are trained on clean speech only.

	clean	car			factory			babble			mean
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	20 dB	10 dB	0 dB	
MFCC baseline	14.20	12.96	19.75	56.48	19.75	36.11	85.19	17.90	25.62	74.69	36.27
MLP-enhanced	14.51	12.35	11.11	33.02	12.65	33.02	83.02	14.51	18.52	70.37	30.31

Table 3: Speaker identification error for training across all noise conditions

In some cases, the speaker identification error is even lower than in the matched noise condition in Table 1, e.g. for factory noise at 10 dB SNR and for babble at 10 and 0 dB SNR. In all conditions except for clean speech is the speaker identification error lower when MLP-enhanced features are used compared to the baseline system using MFCCs.

4. Discussion

In comparison to the stationary car noise, the speaker identification error is very high for non-stationary noises (factory noise and babble) with low SNRs.

As expected, a mismatch between (clean) training and (noisy) test data always leads to a strong deterioriation in performance. This deterioration is even greater when an MLP-enhancement based on clean speech is performed on the MFCC features, since it is not appropriate for the noisy acoustic space.

Both for matched training and test conditions and for GMM speaker models trained across noise conditions (i.e. using both clean speech and the same signals with added noise to train the speaker models), the MLP-enhanced features always outperform standard MFCC features.

The GMM speaker models trained across noise conditions even outperform GMMs trained on the same noise type as used for testing (matched condition). This is particularly the case for test signals with added factory noise at 10 dB SNR as well as for added babble at 10 and 0 dB SNR. The differences are not just caused by the different speaker basis selections for matched versus (Table 1) versus multi-condition models (Table 3), because the effect is found both for MLP-enhanced and for MFCC features. The effect is most likely due to an under-representation of the intra-speaker variance in the limited available training data in the matched tests, which is compensated for by the (artificial) addition of noisy data in the multi-condition tests. It is well-known in speaker recognition that the amount and above all the variety of training data from each speaker is critical. But the results do not show a systematic pattern across noise type and SNR, so that further investigations are needed to fully understand the observed effect.

No CMS was performed in any of the conditions. It is therefore necessary to carry out the experiments with CMS and compare the results with those presented here. Particularly in stationary noise, this should be expected to enhance speaker identification performance.

Of course, the MLP preprocessing described here must also be used for other, more realistic data, e.g. NIST. This was done by [5,6], but there the speaker basis was selected to contain speakers who used all different handset types, so that it is not clear whether the results are due to the MLP performing speaker discrimination or compensation for different handset types. The results shown here lead us to believe that the MLP can discriminate between speakers. The MLP feature enhancement may therefore also be helpful for databases where there is no subset of speakers who used all different handset types, but this remains to be verified.

5. Conclusions

In this paper, text-independent speaker identification was performed in clean and noisy conditions. The performance of GMM speaker modelling using standard MFCC features was compared with MLP-enhanced features, where the MLP was trained on a subset of the speakers which was not used in testing (hence the feature enhancement is speaker-independent). MLP-enhanced features strongly improve speaker identification performance, except when the noise condition of the test data is not represented in the training data. It was shown that, as in ASR, speaker identification in matched and multi-condition training is considerably better than when there is a mismatch between training and test data, both with MFCC and MLP-enhanced features. In some cases, GMM speaker models trained across noise conditions, either with MFCCs or with MLP-enhanced features, perform better than speaker models trained in matched training and test conditions. An explanation is offered in terms of the amount and variability in the training data, but this cannot fully explain the observed results.

References

[1] 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.

[2] Reynolds, D.A.: Experimental evaluation of features for robust speaker identification. *IEEE Transactions on SAP*, Vol. 2, 1994, 639-643.

[3] Sharma, S., Ellis, D., Kajarekar, S., Jain, P. & Hermansky, H.: Feature extraction using non-linear transformation for robust speech recognition on the Aurora database. *Proc. ICASSP2000*, 2000.

[4] Wu, D., Morris, A.C. & Koreman, J.: MLP Internal Representation as Disciminant Features for Improved Speaker Recognition. *Proc. NOLISP2005*, Barcelona, Spain, 2005, 25-33.

[5] 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*, 1998, 72-75.

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

[7] Morris, A.C., Wu, D. & Koreman, J.: MLP trained to separate problem speakers provides improved features for speaker identification. *Proc. ICCST 2005* (in press), Las Palmas, 2005.

[8] Varga, A. and Steeneken, H.J.M.: Assesment for automatic speech recognition:II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication* 12(3), 1993, 247-252.

[9] ITU recommendation P.56, Objective measurement of active speech level, March 1993.

[10] Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. *Speech Comm.* 17, 1995, pp.91-108.

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

[12] Theodoridis, S. & Koutroumbas, K.: *Pattern Recognition* (2nd Ed.). Elsevier Academic Press, 2003.

[13] 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*, 1995, 329-332.