TELEPHONE BASED SPEAKER RECOGNITION USING MULTIPLE BINARY CLASSIFIER AND GAUSSIAN MIXTURE MODELS

Pierre J. Castellano, Stefan Slomka and Sridha Sridharan

Signal Processing Research Centre
Queensland University of Technology
GPO Box 2434 Brisbane QLD 4001 Australia

ABSTRACT

The present study evaluates MBCM and GMM solutions for both ASV and ASI problems involving text-independent telephone speech from the King speech database. The MBCM's accuracy is enhanced by selectively removing those classifiers within the model which perform worst (pruning). An unpruned MBCM outperforms a GMM for ASV and speakers taken from within the same dialectic region (San Diego, CA). Once pruned, the MBCM is found to be 2.6 times more accurate than the GMM. For closed set ASI, based on the same data, the MBCM is roughly twice as accurate as the GMM but only after pruning.

1. INTRODUCTION

The Gaussian Mixture Model (GMM) is recognised as one of the most accurate models for Automatic Speaker Recognition (ASR), using telephone speech [1]. The Multiple Binary Classifier Model (MBCM) is more recent [2] and has been successfully evaluated under a range of experimental conditions characterising speech based Automatic Speaker Verification (ASV) and Automatic Gender Separation (AGS). These conditions include signals affected by band limitation, white noise, coding loss and reverberation [3][4][5]. GMM and MBCM accuracies are compared both for ASV and Automatic Speaker Identification (ASI) in the present text-independent work based on telephone speech.

2. SPEECH DATA AND ITS PRE-PROCESSING

Conversational speech was extracted from King (all 51 males) and Switchboard telephone speech databases [6]. King speakers are from the San Diego, CA and Nutley, NJ regions of the United States. Approximately one minute of speech was processed for each speaker, for all databases. Low energy segments were removed along with silent parts. The resulting signal was high frequency pre-emphasised with transfer function $1 - 0.96z^{-1}$. Other pre-processing specifications include a 256 point hamming window. Speech was parametrised using Mel-based cepstrum coefficients. Training data were extracted from the first three King recording sessions. Test data originated from the last two King recording sessions (9 and 10). In order to minimise channel variation effects, the speech data was subjected to mean normalisation as recommended by Reynolds and Rose [7].

3. ASR DISCRIMINANT MODELS

3.1. GMM

With a GMM, feature vector distribution, for a particular speaker, can be modelled using a GM density given by [7]:

$$p(\overline{x}|(p_j, \mu_j, \Sigma_j)) = \sum_{j=1}^{M} p_j b_j(\overline{x}) \tag{1}$$

where $b_j(\overline{x})$ are uni-modal Gaussian densities, each characterised by mean vector μ_j and covariance matrix Σ_j; p_j are corresponding mixture weights satisfying:

$$\sum_{j=1}^{M} p_j = 1. \tag{2}$$

The GMM marries the concept of uni-modal Gaussian classifiers to that of codebooks produced by vector quantisation. The more mixtures utilised in the model, the tighter the latter fits training data. The GMM implementation adopted in the present work is that of Reynolds and Rose which utilised 50 mixture components [7]. (Extensive testing by the authors of the present paper showed that a GMM based on that number of components was indeed optimum for the data considered.) One GMM is trained for each reference speaker considered in an ASR problem.

3.2. MBCM

In the MBCM model, K individual classifiers are allocated to each of N true speakers. Each set of K classifiers is trained with identical speech from the corresponding true speaker as well as with speech from a single alternative speaker who is different for each individual classifier within the set. The MBCM may be implemented with any statistical or connectionist classifier which has shown promise in ASV. When an identity is claimed, during an ASV application, all K classifiers, associated with that identity, are tested with an identical sample of the claimant's parametrised speech. The mean over the K classifier outputs (fusion) is compared against a pre-set threshold. For classifier J tested ($J <= K$), let $f_{ts,J}$ be the fraction of M test vectors which are more closely identified with the true speaker rather than the alternative. Let

$$F_{ts} = \frac{1}{K} \sum_{J=1}^{K} f_{ts,J}. \tag{3}$$

In the ASV context, the reference speaker class will be retained if F_{ts} exceeds the pre-set threshold. The MBCM is illustrated in Figure 1. Advantages of this discriminative approach include:

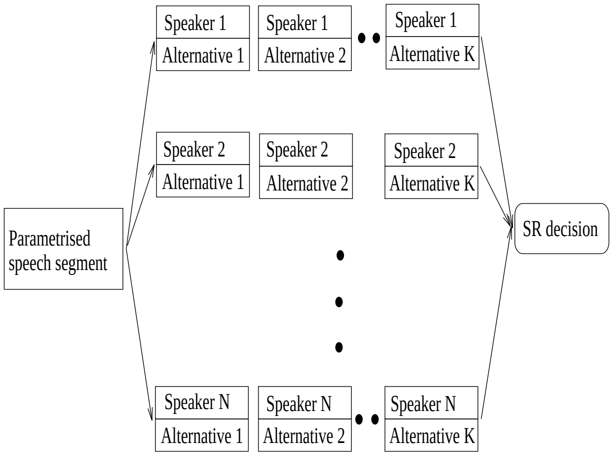

Figure 1. Multiple Binary Classifier Model

- a minimisation of speaker pattern overlap since only two speaker classes are considered per individual classifier, irrespective of true speaker population size and
- the possibility of correctly discriminating classes even if some individual classifiers, within a set of K, misclassify.

Furthermore, since the reference speaker data required to train an MBCM is the same as that needed to train any of the MBCM's individual classifiers:

- the number of classifiers incorporated within an MBCM is limited only by the number of alternative speakers utilised (and not the number of available classifiers or speech parametrisation types) and
- only the best performing individual classifiers within an MBCM can be retained following pruning.

Individual classifier performance within an MBCM is determined by testing the classifier with validation data from reference speakers (distinct from training and test data) and with more validation data from speakers for which the classifier has not been trained. Poor classifiers are pruned from MBCMs (in inverse order of performance). In practice, alternative speakers may be found outside the problem of interest. Thus an MBCM can, theoretically, be composed of a near infinite number of classifiers of the same type. If real time computerised ASR is required then MBCM classifiers could be implemented in parallel.

The individual classifier selected as building block for the MBCM, in the present study, was the the Moody-Darken Radial Basis Function Neural Network (MD-RBFN) which is one of the most robust classifiers for ASR [8] [9]. The MD-RBFN consists of a K-means clustering front end, delivering an initial solution and a weighted back end implementing gradient descent and refining the initial solution. The network is characterised by a transfer function akin to a Gaussian distribution. The MD-RBFN's parameters were adopted from a previous study [10] where they had been optimised for the wide band version of the King database and not the narrow band version considered in this study.

4. ASV OF SPEAKERS FROM THE SAME DIALECTIC REGION

Twenty true speakers from the San Diego region of the United States and the narrow band portion of the King database were considered. Each true speaker was allocated a single GMM trained with all speech available for that speaker, taken from the first three recording sessions. In parallel with the above and before pruning, each of the 20 true speakers was allocated an MBCM consisting of K = 45 MD-RBFNs. The true speaker training data for these MBCMs consisted of 300 parameterised vectors drawn from the same sessions as for the GMM. The 45 alternative speakers required for the MBCMs contributed 300 vectors each and were provided by 20 King speakers (distinct from the true speakers) and 25 male speakers selected at random from Switchboard. To implement MBCM pruning, the remaining King speakers (those not used as true or alternative speakers) and 24 additional Switchboard speakers (selected at random and distinct from those used as alternative speakers) were used to provide 150 vectors each for validation imposter data. True speaker validation data consisted of 150 vectors distinct from MBCM training data and drawn from the same recording sessions as the GMM training data.

Test data for the 20 true speakers were part of the the last two King recording sessions and consisted of 150 parameterised vectors per speaker. Test data were the same for GMMs and MBCMs so that a performance comparison of both models could be made. For each true speakers' GMM and MBCM, test data were used to provide one true speaker score (when the discriminant model was tested with test data from the same speaker with which it was trained) and 19 imposter scores (when the discriminant model was tested with test data from the other 19 speakers). A percentage of instances (scores) where parametrised speech vectors, belonging to an imposter, were classified more often as belonging to a particular true speaker than that true speaker's own test data was taken into account. The criterion provides an absolute means of assessing the MBCM's discrimination performance. This avoids the consideration of a threshold which is a less clear cut criterion since it is application dependent and may have to be set by trial and error [11].

Figure 2 illustrates GMMs and MBCMs (pruned and unpruned) mean speaker discrimination performances as a function of the number of individual classifiers included in the models. Results were averaged over the possible combinations of K ($\mathrm{C}^{19}_{\mathrm{K}}$). Means were computed over 380 imposter attempts (19 imposters each trying to beat 20 MBCMs, in turn). The MBCMs outperformed GMMs by 21 per cent on average, before pruning. The margin increased to 62 per cent following MBCM pruning. Pruned MBCMs consisted of 8 classifiers each as opposed to the initial 45 (unpruned model). The difference in performance is attributed to the GMMs' occasional mediocre discrimination of speakers. This is explained by the non-stationary nature of the speech signal causing differences between training and test data characteristics which are often significant. By contrast, the MBCMs' performances were more consistent due to their ability to average out poor individual MD-RBFN classifications.

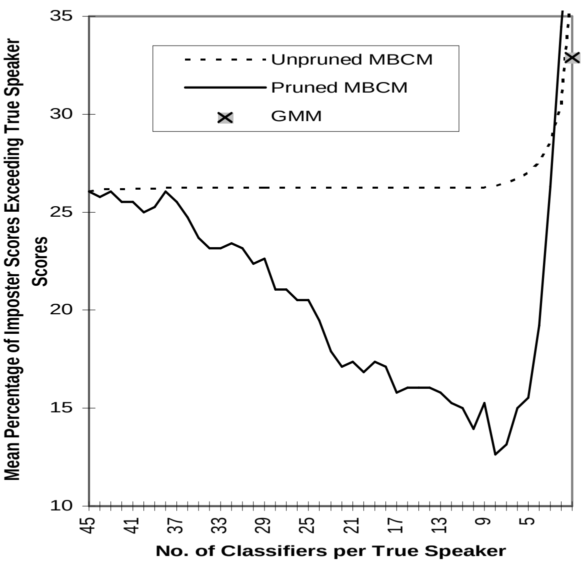

Figure 2. Mean percentage of imposter scores exceeding true speaker scores for GMMs, MBCMs and pruned MBCMs, when imposters and true speakers are from the same dialectic region, as a function of number of individual classifiers included in the models

5. ASI OF SPEAKERS FROM THE SAME DIALECTIC REGION

As in the previous experiment, twenty King reference speakers from the San Diego region were considered. Each speaker was allocated a single GMM trained with all speech available for that speaker, taken from the first three King recording sessions.

In parallel with the above, each of the twenty reference speakers was also allocated six MBCMs (A to E) based on MD-RBFNs. These discriminant models consisted of:

1. $MBCM_A$: 19 classifiers, each trained with speech from one of the remaining 19 speakers considered in the closed set problem
2. $MBCM_B$: 45 classifiers, each trained with speech either from one of the remaining 19 speakers considered in the closed set problem or from one of 26 speakers taken from outside that problem (from Switchboard)
3. $MBCM_C$: $MBCM_A$ pruned down to 5 classifiers using, as validation data, closed set speech (from the 20 King speakers of interest). This speech was distinct from that used to train and test the remaining 5 classifiers
4. $MBCM_D$: $MBCM_A$ pruned down to 5 classifiers using the closed set (from $MBCM_C$) as well as additional validation data taken from outside the closed set problem (from 25 Switchboard speakers)
5. $MBCM_E$: $MBCM_B$ pruned down to 8 classifiers using, as validation data, closed set speech (from the 20 King speakers of interest). This speech was distinct from that used to train and test the remaining 8 classifiers
6. $MBCM_F$: $MBCM_B$ pruned down to 8 classifiers using the closed set (from $MBCM_E$) as well as additional validation data taken from outside the closed set problem (from 25 Switchboard speakers)

		GMM	$MBCM_A$	$MBCM_B$	$MBCM_C$	$MBCM_D$	$MBCM_E$	$MBCM_F$
Validation data utilised		none	none	none	closed set	open set	closed set	open set
Number of classifiers pruned		0	0	0	14	14	37	37
Speaker	1	0	1	1	0	0	0	0
	2	0	1	0	0	0	0	0
	3	8	8	9	7	8	7	6
	4	0	3	5	3	1	4	3
	5	12	1	1	3	3	3	1
	6	12	9	7	5	6	8	2
	7	2	6	6	5	6	4	4
	8	15	12	11	3	8	3	2
	9	1	1	1	0	2	0	0
	10	4	1	2	3	1	4	1
	11	8	8	8	11	10	8	8
	12	13	1	1	0	2	8	0
	13	10	0	0	0	0	0	1
	14	0	0	0	0	1	0	0
	15	3	5	5	5	6	7	7
	16	6	4	4	4	5	3	4
	17	11	2	11	0	0	0	0
	18	3	2	1	4	4	1	5
	19	7	7	16	16	0	8	0
	20	10	12	7	7	18	6	5
Column mean		6.25	4.2	4.8	3.8	4.05	3.7	2.45

Table 1. Closed set ASI based on 20 San Diego speakers using GMMs, unpruned MBCMs and pruned MBCMs

In contrast to ASV, MBCMs implemented for closed set ASI may be trained exclusively with (closed set) speech provided by the speakers considered in the problem ($MBCM_A$, $MBCM_C$ and $MBCM_D$). As is the case for ASV, the dimension of those MBCMs need not be limited by the size of the reference speaker set ($MBCM_B$, $MBCM_E$ and $MBCM_F$). Results for all 20 reference speakers, the GMMs and MBCMs are listed in Table 1. In that table, all such speakers are identified using each of the GMM and 6 MBCMs in turn. Each individual outcome is expressed in terms of the number of reference speakers (out of a maximum number of 19) which are mistaken for the particular reference speaker targeted. Thus a '0' corresponds to a correct identification while 'Column mean' establishes the mean ability of each of the discriminant models to separate the 20 reference speakers. (The lower a mean the more accurate the model.) It is apparent from Table 1 that ASI accuracy was poor regardless of which model was used to discriminate speakers. This is in agreement with previous ASI work conducted using telephone speech [12]. Although the different MBCMs' column means were smaller than that of the GMM, the latter correctly classified more speakers (4 out of 20) than unpruned MBCMs: $MBCM_A$ (2 speakers) and $MBCM_B$ (3 speakers). Pruned MBCMs outperformed the GMMs and unpruned MBCMs with between 5 and 7 speakers correctly identified. However, fetching additional data from outside the closed set problem in order to train and/or more thoroughly validate higher dimensional MBCMs did not materialise in significant gains in ASI accuracy ($MBCM_E$ and $MBCM_F$) although column means improved (see table 1).

The option of augmenting closed set data with open set data

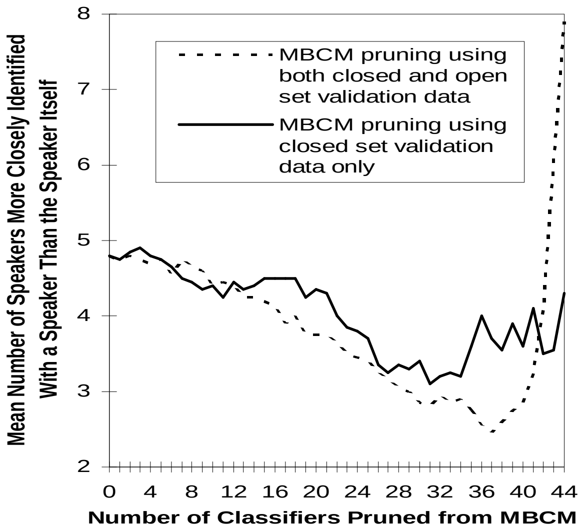

Figure 3. Closed set ASI accuracy using MBCMs pruned separately using closed set and both open and closed set validation data

taken, in this instance, from a different database is further investigated in Figure 3. As for Table 1, the figure illustrates a slightly lower mean number of speakers more closely identified with another reference speaker than the speaker itself, for the $MBCM_F$ as opposed to the $MBCM_E$. Interestingly, the latter's performance degrades less abruptly when more than 36 classifiers are pruned from the original 45 dimensional model. Thus, it is by no means clear from the present ASI problem whether speaking discrimination accuracy has benefited from the additional use of data taken from outside the problem. Other approaches to closed set ASI involve the use of a single individual classifier trained for all reference speakers. These consider all speakers simultaneously and thus can directly model differences in their speeches [13]. Such approaches have led to fairly robust closed set ASI for small speaker reference sets [14]. Their main disadvantage is the scaling problem. A comprehensive study comparing the ASI accuracy of single multi class classifiers to multiple classifier discrimination models is presently lacking in the literature.

6. CONCLUSION

This study has compared the discrimination accuracies of GMMs and MBCMs in the context of a text-independent ASR problem. The problem involved narrow band speech with training and test data sets separated by 4.5 to 6 months in time.

The MBCMs were found to be the most robust for ASV. Two factors account for this outcome. Firstly, the GMMs, only one of which is allocated to each reference speaker, generalise poorly, on occasion, between training and test due to the non-stationary nature of the signal. The MBCMs, through extensive fusion of individual classifier outcomes, lessen the impact of poor individual outcomes provided these are in the minority. Secondly, MBCMs may be pruned so the worst performing individual classifiers are removed, significantly boosting speaker discrimination accuracy. In the context of closed set ASI, pruned MBCMs, restricted to closed set data, outperformed GMMs.

REFERENCES

[1] Gish H. and Schmidt M., “Text-Independent Speaker Identification”, IEEE Signal Proc. Mag., Vol. 11, No. 4, Oct. 1994, pp. 18-32.

[2] Castellano P., “Text-Independent Speaker Verification with a Multiple Binary Classifier Model”, Proc. Sec. Aust. and New Zeal. Conf. Intell. Info. Sys., 29 Nov.-2 Dec. 1994, , pp. 51-55.

[3] Castellano P. and Sridharan S., “Text-Independent Speaker Identification with a Tensor-Link Neural Network”, Applied Signal Proc., Vol. 1, No. 3, 1994, pp. 155-165.

[4] Castellano P., Sridharan S. and Cole D., “Speaker Recognition in Reverberant Enclosures”, Proc. Inter. Conf. Acous. Speech Sign. Proc., Vol. 1, April 1996, pp. 117-120.

[5] Slomka S., Barger P., Castellano P. and Sridharan S., “Gender Gates in Degraded Environments”, To appear in Proc. of Sixth Aust. Int. Conf. on Speech Sci. and Tech., Dec. 1996.

[6] Godfrey J., Graff D. and Martin A., “Public Databases for Speaker recognition and Verification”, Proc. Work. Auto. Speaker. Rec. Iden. Ver., 5-7 April 1994, pp. 39-42.

[7] Reynolds D. A. and Rose C., “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans. Speech Audio Proc., Vol. 3, No. 1, Jan. 1995, pp. 72-83.

[8] Mak M. W., Allen W. G. and Sexton G. G., “Comparing Multi-layer Perceptrons and Radial Basis Functions networks in Speaker Recognition”, J. Micro. Appl., Vol. 16, April 1993, pp. 147-159.

[9] Castellano P. and Sridharan S. “Speaker Identification with Concomitant Open and Closed Decision Boundaries”, Aus. J. Intell. Info. Proc. Sys., Vol. 2, No. 2, Winter 1995, pp. 47-53.

[10] Slomka S., Castellano P. and Sridharan S., “An Augmented Multiple Binary Classifier Model for Speaker Verification”, Proc. Sev. Aus. Conf. Neu. Net., April 1996, pp. 39-44.

[11] Doddington G. R., ”Speaker Recognition - Identifying People by their Voices”, IEEE, Vol. 73, No. 11, 1985, pp. 1651-1664.

[12] Farrell F. and Mammone R. J., “Data Fusion Techniques for Speaker Recognition”, in Modern Methods of Speech Processing - Ed. Ramachandran R. P. and Mammone R. , Kluwer Academic Publishers, 1995, pp. 279-297.

[13] Oglesby J. and Mason J. S., “Optimisation of Neural Models for Speaker Identification”, Proc. Inter. Conf. Acous. Speech Sign. Proc., Vol. 1, April 1990, pp. 261-264.

[14] Castellano P., “A Study of LVQ Learning Schedules for ANN Speaker Identification”, Proc. IEEE Region 10's Ninth Annual International Conference, August 1994, pp. 902-906.