GENDER GATES FOR AUTOMATIC SPEAKER RECOGNITION

Peter Barger, Stefan Slomka, Pierre Castellano and Sridha Sridharan

SPRC, School of Electrical and Electronic Systems Engineering
Queensland University of Technology, Brisbane 4001, Australia

ABSTRACT: The present work, based on telephone speech, proposes gender gates suitable as front-ends to ASR discrimination systems. The gates are composed of connectionist and/or statistical classifiers whose outputs are fused for increased robustness. While gender separation is simpler than speaker separation, the former is not a trivial problem as is commonly assumed.

INTRODUCTION

Automatic Speaker Recognition (ASR) is composed of Automatic Speaker Identification (ASI) and Verification (ASV). ASI is a closed set problem where unknown speech is authenticated with the reference speaker whose speech it most closely matches. ASV is open set and a speaker, within a given reference set, is retained only if his/her signal matches most closely unknown speech and if that matching exceeds a pre-set threshold. ASR is a process consisting of speech data collection, pre-processing of the speech signal (enrolment), pattern matching and adjudication. Many prominent discriminant models, used in the pattern matching step, such as the binary partitioned neural network approach (Rudasi and Zahorian, 1991) and the Augmented Multiple Binary Classifier Model (Slomka et al, 1996) consist of twin architectures. Each such architecture is trained to discriminate speech from one gender only. This entails, explicitly, that unknown speakers be discriminated on the basis of gender before further discrimination can proceed. This constraint may be satisfied, in principle, by resorting to a front end classifier or discriminant model known as a gender gate. No study has focused specifically on this aspect of the ASR process except for the recent work of (Parris and Carey, 1996). Such a gate reduces ASR complexity since past it, the size of speaker populations could be halved, assuming equal distributions of genders. The present work investigates a great number of gates for gender separation and their relative performances. These gates are based on fusion from outputs of multiple classifiers. In addition, pruning is used to remove the gates' least robust classifiers.

SPEECH DATA COLLECTION AND PRE-PROCESSING

Speech was extracted from the Switchboard telephone database which supports text-independent research (Godfrey et al, 1994). The data set compiled for the present study contained parametrised speech vectors from 79 males and 75 females. In our gender gates experiments, training sets consisted of speech from 30 males and 30 females and test data sets of speech from 22 males and 22 females. Validation sets, when utilised, consisted of speech from 27 males and 23 females. Training, test and validation speakers were distinct. Approximately one minute of speech was processed, for each speaker. Silent parts and low energy segments were removed. The resulting signal was high frequency pre-emphasised with transfer function $1 - 0.98z^{-1}$. Other pre-processing specifications include a 256 point Hamming window and an analysis filter of order 15. The parametrisations schemes used to reduce the speech signal included: Mel-based cepstra, Line Spectrum Pairs (LSPs), reflection and autocorrelation coefficients. All gender gate experiments, described in this paper, utilise 300 parametrised speech vectors for classifier training and 150 (distinct) vectors for testing, for each of the parametrisation schemes. Moreover, for each speaker, training and test data were extracted from different conversations recorded over different channels and handsets.
In addition to the speech parametrisation scheme discussed above, pitch information was also extracted from the Switchboard speakers' speeches. This information is potentially useful since pitch is generally higher for females (110-210 Hz) than for males (50-140 Hz) although it varies suddenly over time since it is driven by an individual emotional state and the need to adapt to context. The pitch detection algorithm utilised in this paper shall now be described:

Given a speech sample $s_m(n)$. The Kth real cepstrum coefficient at frame m is defined as:

$$C_m(K) = \mathcal{DTFT}^{-1}(log|\mathcal{DTFT}(s_m(n))|) \tag{1}$$

where $\mathcal{DTFT}$ is the Discrete Time Fourier Transform. If

$$C_m(K_{max}) > t_p \frac{1}{l+1} \sum_{K=K_1}^{K_1+l} |C_m(K)| \tag{2}$$

then the estimated pitch period is given by

$$P = K_{max} F_s \tag{3}$$

where $C_m(K_{max})$ is the first peak of the cepstrum in the interval $K_1 \leq K \leq K_1 + l$ which meets the abovementioned condition, F_s is the sampling frequency and:

$$K_1 = t_1 F_s \tag{4}$$

$$l = (t_2 - t_1) F_s \tag{5}$$

where t_1 is the minimum probable pitch period ($\approx$ 2ms), t_2 is the maximum probable pitch period ($\approx$ 15 ms) and t_p is an a priori threshold ($\approx$ 5 ms). The telephone, by acting as a bandpass filter can attenuate pitch frequency and higher harmonics. This renders the periodicity of the transmitted signal harder to detect than for wide band speech (Rabiner et al, 1976). However, this problem is usually minor for Switchboard data.

CLASSIFIERS USED IN THE PRESENT STUDY

Classifiers considered in the present study include the

- Mahalanobis distance classifier (statistical),
- Moody-Darken Radial Basis Function (MD-RBFN) (connectionist),
- Higher Order Neural Network (HONN) (connectionist) and
- Time Delay Higher Order Neural Network (TDHONN) (connectionist).

Assuming that speech for a multi-speaker problem is parametrised, a test vector in an N dimensional space, for speaker k and speech frame i may be expressed as $x_{N,k,i}$. A reference vector $ref_{N,k,s}$ for speaker k, is customarily produced by averaging each of its parameters over a segment of speech. Mahalanobis distance $d_{N,k,i}$, between $x_{N,k,i}$ and $ref_{N,k,s}$ is defined as

$$d_{N,k,i} = \left[(x_{N,k,i} - ref_{N,k,s})^T W^{-1} (x_{N,k,i} - ref_{N,k,s})\right]^{1/2}, \tag{6}$$

where W is a pooled covariance matrix and T is the transpose operator.
The MD-RBFN is characterised by a front end which clusters data and a back-end implementing gradient descent. Neurons in the hidden layer possess Gaussian transfer functions. Thus the role of the network's hidden layer can be viewed as implementing a connectionist version of Gaussian Mixtures.

The TDNN is aimed at extracting extra dimensions from data which an MLP is not able to do. These dimensions include temporal correlations of data vectors and a degree of invariance under time translations. An N dimensional input vector

$$X(t) = [x_1, x_2, .., x_N], \tag{7}$$

presented to the network at time time t is mapped to

$$H_{td}(t) = [X(t), X(t-1), .., X(t-T)], \tag{8}$$

where T is the TDNN's window or prediction order.

The HONN is a Multi-Layered Perceptron (MLP) whose input is mapped into a higher dimension with the intention of rendering it more easily separable. In this paper a second order mapping was utilised. An N dimensional input vector $X(t)$ (Equ. 7) is mapped to

$$H_{ho}(t) = [x_1, x_1x_2, x_1x_3, x_2, x_2x_3, x_2x_4, .., x_k, x_kx_{k+1}, x_kx_{k+2}, .., x_N, x_Nx_1, x_Nx_2]. \tag{9}$$

The TDHONN is an HONN with higher order inputs to hidden layer neurons (or to output layer neurons if no hidden layer is utilised) delayed in time. The mapped input, processed by the network, is a hybrid which may be expressed by considering both Equ. 8 and 9.

$$H_{tdho}(t) = [H_{ho}(t), H_{ho}(t-1), .., H_{ho}(t-T)] \tag{10}$$

where T is the TDHONN's window or prediction order. The sets of higher order inputs are thus overlapping in time. All classifiers discussed in this section are well established in ASR with the exception of the TDHONN.

GATES UTILISING FUSION OF OUTPUT FROM CLASSIFIERS OF THE SAME TYPE

A series of text-independent gender discrimination experiments were conducted, aimed at devising a discriminant system called a gender gate. This must be able to automate gender separation with high accuracy. Such a gate is either an individual classifier or an architecture consisting of a number of individual classifiers such as an MBCM (Castellano and Sridharan, 1994). Throughout this paper, a gender discrimination error refers to the case where a gate classifies a speaker as being of the opposite gender.

Mahalanobis distance

Experiments investigated the use of several Mahalanobis distance classifiers, whose outputs were fused. Fusion was accomplished by averaging classifier outputs over all classifiers utilised. Experiments were repeated for a range of speech parametrisation schemes and (for training set) general population sizes S of 1 to 10, 12, 15 and 30 speakers of the same gender. An MBCM like discriminant model, referred to as a Gender MBCM (GMBCM) was used in this case. The model incorporated (30 div. S) Mahalanobis distance classifiers (where div. is integer division). Each model was trained with 30 males and 30 females in all. Each individual classifier, within the model, was trained with S males and S females different for all classifiers. Individual classifier outcomes were fused (averaged). Classifiers were tested with unknown speech provided successively by male and female test sets. A single Mahalanobis distance classifier trained with parametrised speech (reflection coefficients) from 30 males and 30 females provided the best gender discrimination (see Table 1). Its performance was insufficient to correctly recognise all 44 speakers on the basis of gender.

Moody-Darken Radial Basis Function Network

A single MD-RBFN did not separate genders as accurately as a discriminant model consisting of several MD-RBFNs whose outputs were fused. This was so regardless of how many speakers were used to train the single ANN (see Table 1). The discriminant model used was the GMBCM. Best performance was obtained for S equal to 2 and 15 MD-RBFNs. (A GMBCM of 30 MD-RBFNs, each trained with a single male and a single female (S equal to 1), was only marginally inferior.) The retained GMBCM architecture was subjected to classifier pruning. Thus, in a first step, the 23 males and 27 female validation speakers were classified using the GMBCM. Classifiers, incorporated within the architecture, were graded from most accurate to least accurate on that basis. In a second step, the GMBCMs were reduced by one classifier at a time, starting with the least accurate. Pruning was terminated before GMBCM performance begun to degrade (through the use of too few classifiers). Pruning has the double merit of simplifying discriminant models while increasing their robustness. Figure 1 illustrates gender discrimination accuracy obtained with a gender gate based on GMBCMs (MD-RBFNs). This accuracy is given both in terms of the number of (unsorted) classifiers utilised (classifier fusion alone) and in terms of the reduced number of classifiers used following pruning based on the validation data. Results for selected GMBCMs, once pruned and for the various speech parametrisation schemes, were inferior to those delivered by the single Mahalanobis distance classifier, trained with parametrised speech from all available speakers, by approximately 5 per cent on average.

Higher Order Neural Network and Time-Delay

A series of experiments were undertaken aimed at determining whether the HONN was more accurate, on an individual classifier basis, when implemented with or without a hidden layer. Further experiments focussed on number of speakers used in general populations. (The general populations of speakers investigated were those used in the previous Mahalanobis distance classifier and MD-RBFN experiments.) Optimum gender discrimination was obtained without hidden layers and using a GMBCM of 30 HONNs, for S equal to 1. The GMBCM was then pruned in the same manner as for the MD-RBFNs above. The pruned GMBCM's gender discrimination performance, for the various speech parametrisation schemes, was superior to the best achieved with the MD-RBFN and comparable to that of the Mahalanobis distance classifier (see Table 1). Replacing the HONNs with TDHONNs added temporal information to the discrimination process. However, this addition consistently degraded results probably because the discriminative ability of the network significantly exceeds that required for the present problem.
In summary, a single Mahalanobis distance with reflection coefficients yielded the best gender discrimination accuracy of all discriminant systems investigated. Mean Gender discrimination performance, using pitch on its own, was as accurate as that obtained with the above pruned GMBCM based either on MD-RBFNs or TD-

Table 1: Gender discrimination accuracies for gates incorporating a single type of classifier, for 22 male and 22 female speakers

General population size	Number of classifiers fused	Number of gender misclassifications following fusion	Number of gender misclassifications following pruning	Number of gender misclassifications for best classifier
Classifier type: MD-RBFN Speech parameter type: Mel-based cepstrum				
1	30	8	5	8
2	15	6	5	12
3	10	8	8	9
4	7	9	8	9
5	6	8	8	8
6	5	9	9	9
8	3	9	8	10
10	3	11	11	9
15	2	11	8	8
30	1	N.A.	N.A.	16
Classifier type: Mahalanobis distance Speech parameter type: Mel-based cepstrum				
1	30	4	3	9
2	15	3	3	8
3	10	5	5	5
4	7	5	3	3
5	6	4	4	4
6	5	5	4	7
7	4	4	4	5
8	3	4	4	9
9	3	5	4	6
10	3	5	4	9
12	2	4	3	9
15	2	4	3	5
30	1	N.A.	N.A.	3
Classifier type: Mahalanobis distance Speech parameter type: Reflection				
15	2	2	N.A.	N.A.
30	1	N.A.	N.A.	2
Classifier type: Mahalanobis distance Speech parameter type: LSP				
15	2	3	N.A.	N.A.
30	1	N.A.	N.A.	2
Classifier type: Mahalanobis distance Speech parameter type: Autocorrelation				
15	2	4	N.A.	N.A.
30	1	N.A.	N.A.	2
Classifier type: Mahalanobis distance Speech parameter type: Mel-based cepstrum				
15	2	4	N.A.	N.A.
30	1	N.A.	N.A.	3
Classifier type: HONN Speech parameter type: LSP				
1	30	5	3	6
2	15	5	3	8
Speech parameter type: LSP Classifier type: HOTDN				
1	30	6	6	9
2	15	7	4	9

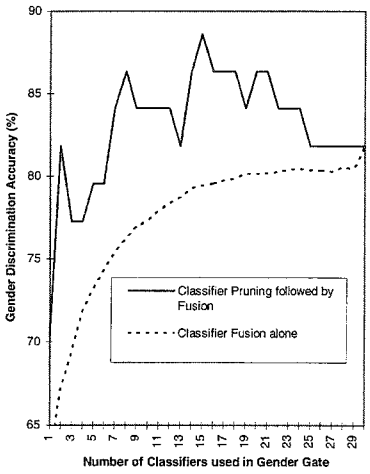

Figure 1: Gender discrimination accuracy obtained with a gender gate based on GMBCMs (MD-RBFNs)

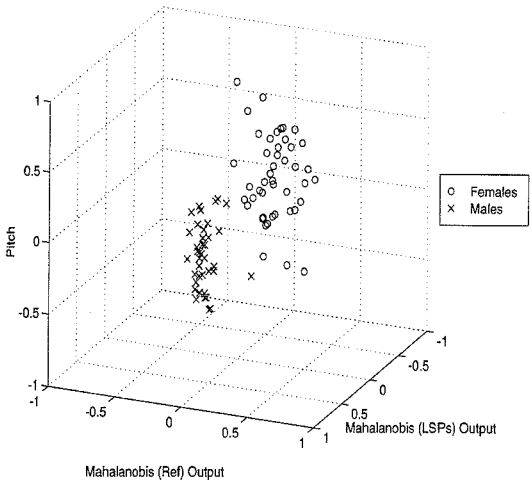

Figure 2: Linear gender separability for pitch information together with outputs from 2 Mahalanobis distance classifiers

HONNs (5 to 8 genders misclassified out of 44, depending on the speech parametrisation scheme employed). Pitch based discrimination was inferior to the pruned GMBCM with HONNs and LSPs (3 genders misclassified out of 44). No discriminant model was able to separate all 44 speakers, illustrating that gender separation is not a trivial problem.

GENDER GATES UTILISING CLASSIFIER AND FEATURE FUSION

Previously only outcomes from classifiers of the same type were fused. In the present section, fusion of classifier output was conducted on a wider scale. As a first step and as in the case of previous gender gate experiments, outcomes from some or all classifiers of a same type were fused. This step was repeated for the four speech parametrisation schemes previously utilised and for pitch which was considered on its own. Outcomes using some or all schemes, as well as pitch, were then fused in a second step. In a third step, Step 2 was repeated for gates based on different types of classifiers and extensive averaging was obtained by fusing the gates' outcomes. Such an approach quickly led to an explosion in the number of different gate architectures created at the end of Step 3. Thus, in step one, only the best performing gates derived from the previous section were examined, namely

- Single Mahalanobis distance classifier trained with 2 general populations of speakers (30 males and 30 females),
- Pruned GMBCM with up to 30 HONNs, each trained with LSPs extracted from speech from a single male and a single female (S equal to 1) and
- Pruned GMBCM with MD-RBFNs, each trained with Mel-based cepstrum coefficients extracted from 2 males and 2 females (S equal to 2).

It was observed that vectors of outputs from these gates, corresponding to males and females, were often (but not always) linearly separable. Linear gender separability is illustrated in Figure 2 for pitch information together with outputs from 2 Mahalanobis distance classifiers (inputs being reflection coefficients and LSPs respectively). (All outputs are normalised between -1 and $+1$.) Thus output fusion was conducted by inputting the output vectors into a Perceptron (well known for its ability to classify linearly separable data). This back-end Perceptron was trained using output from the front-end classifiers whose inputs were data from the 27 male and 23 female validation speakers. The addition of this classifier led to slightly better gender discrimination (by approximately 3 per cent) than simply averaging outputs from the front-end classifiers. Output from the Perceptron itself was either a 0 (male) or a 1 (female). Discrimination results delivered by 126 different gender gates, generated after Step 2 and Step 3 were produced in a search for improved gate

robustness. The 26 best performing gates were able to correctly discriminate 43 of the 44 speakers on the basis of gender. (One male was consistently identified as a female.) These gates ranged, in order of complexity, from pitch fused with the output of a single Mahalanobis distance classifier to pitch fused to the outputs of 4 Mahalanobis distance classifiers (each trained with a different type of coefficient) and a pruned GMBCM. The two gates with the simplest architectures, within the 26, consisted of a single Mahalanobis distance classifier, trained with either reflection or cepstrum coefficients, whose output was fused to pitch. The next simplest architecture contained the fusion of a single Mahalanobis distance classifier, trained with reflection coefficients and a pruned GMBCM which could be composed of either MD-RBFNs or HONNs.

Two gates within the 26 incorporated MD-RBFNs. Ten gates incorporated HONNs. As for the TDHONN, the MD-RBFN's poor performance was probably due to its discriminative power exceeding that required for the gender separation problem. Separability of all 44 speakers was not achieved. One explanation lies in input to the Perceptron not always being linearly separable.

The use of any of the above 26 gates, as a front end to an ASR system would greatly improve the latter's accuracy. None, however, would be suitable for, telephone based, financial transaction ASR which must be virtually flawless. However, other ASR telephone applications, which also have clientele of several million, are viable for speaker mis-recognition rates of 5 to 10 per cent. In this context, ASR compares favourably with other, telephone based, verification methods such as proffering or even the use of keypads (Naik, 1990). Thus our most accurate gates may enable ASR systems to reach sufficient accuracies for those applications.

CONCLUSIONS

Several investigations of gender gate architectures were conducted using Switchboard's band limited signal, four speech parametrisation schemes and pitch, 3 neural networks (HONN, TDHONN and MD-RBFN) and the Mahalanobis distance classifier. Training, test and validation data, used for gender gates experiments, were distinct and recorded during separate conversations. They consisted of speech from 30 males and 30 females, 22 males and 22 females and 27 males and 23 females, respectively.

Firstly, it was seen that a pruned GMBCM with HONN classifiers (S equal to 1), trained with LSPs was slightly less accurate than a single Mahalanobis distance classifier on average. The fact that the HONN (with second order links and no hidden layer) outperformed the other ANNs is indicative of gender discrimination being a simpler problem than ASR where the MD-RBFN, a more powerful classifier, has an edge. The addition of temporal information through the use of the TDHONN did not benefit the present problem.

A second investigation built upon the first by fusing the outcomes of Mahalanobis distance classifiers amongst themselves as well as with outcomes from GMBCMs (with ANNs), for several parametrisation schemes as well as pitch. Fusion was conducted through the use of a Perceptron. The Perceptron was trained using the output of the gender gates front-end classifiers when their input was speech from the validation speakers. This resulted in 126 different gender gates with superior mean performance than previous gates. The most important contributors to this performance included the use of pitch, and Mahalanobis distance classifiers.

REFERENCES

Rudasi, L. and Zahorian, S. (1991) "Text-Independent Talker Identification with Neural Networks", In Proc. of the International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 389-392.

Slomka, S., Castellano, P. and Sridharan, S. (1996) "An Augmented Multiple Binary Classifier Model for Speaker Recognition", In Proc. of the Seventh Australian Conference on Neural Networks, pp. 39-44.

Parris, E.S. and Carey, M.J. (1996) "Language Independent Gender Identification", In Proc. of the International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 685-688.

Godfrey, J., Graff, D. and Martin, A (1994) "Public Databases for Speaker Recognition and Verification", In Proc. of the Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, pp. 39-42.

Castellano, P. and Sridharan S. (1994) "Text-Independent Speaker Identification with a Tensor-Link Neural Network", Applied Signal Processing I, pp. 155-165.

Rabiner, L. R., Cheng, M. J., Rosenberg, A. E. and McGonegal, C. A. (1976) A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Transactions Acoustics, Speech and Signal Processing, AAP-24 (5), pp. 395-418.

Naik, J. M. (1990) "Speaker Verification: A Tutorial", IEEE Communications Magazine, pp. 42-48.