AUTOMATIC GENDER IDENTIFICATION UNDER ADVERSE CONDITIONS

Stefan Slomka and Sridha Sridharan
Speech Research Laboratory, Signal Processing Research Center
Queensland University of Technology, GPO Box 2434, Brisbane, 4001, Australia
E-mail: slomka@markov.eese.qut.edu.au and s.sridharan@qut.edu.au

ABSTRACT

This paper evaluates 63 Automatic Gender Identification (AGI) systems for text-independent clean speech segments, coded speech and speech segments affected by reverberation. The AGI systems contain a Linear Classifier (LC) with inputs from a combination of two average pitch detection methods and paired Gaussian Mixture Models trained with mel-cepstral, autocorrelation, reflection and log area ratios parameterised speech data. An AGI system is built which is able to handle the LPC10, CELP and GSM coders with no significant loss in accuracy and reduce the impact of even severe reverberation by subjecting the training data of the LC with a different room response. Using speech segments with an average duration of 890ms (after silence removal), the best AGI system had an accuracy of 98.5% averaged over all clean and adverse conditions.

1. INTRODUCTION

In Automatic Gender Identification (AGI), a computerised system is used to identify an individual's gender by analysis of his/her speech signal. The need for AGI arises in several situations, some of which are: 1) sorting telephone calls by gender (e.g. for gender sensitive surveys), 2) as part of an automatic speech recognition system to enhance speaker adaptation, and 3) as part of Automatic Speaker Recognition (ASR) systems. With regards to the latter application, some ASR systems discriminate speakers from one gender only [10]. This entails that unknown speakers be discriminated on the basis of gender before further discrimination can proceed. AGI may also be useful as an additional step to reduce cross gender errors present in other ASR systems such as that by Reynolds [1]. In the past, AGI has been investigated for clean speech by Wu and Childers [14]. Recently Parris and Carey [2] studied AGI for different languages using telephone speech data.

This paper investigates AGI under degraded conditions using text-independent speech data from the TIMIT database [6]. The conditions examined are reverberation and speech coding. The effect of these adverse conditions have been investigated recently for ASR [7] (reverberation) [11] (coding). The following speech parameter types are investigated for their suitability to AGI under adverse conditions: Autocorrelation, Reflection, Mel-Cepstral, Log Area Ratio Coefficients [5], and two estimates of the average pitch. The suitability to AGI of the Gaussian Mixture Model (GMM) Classifiers (this classifier has been extensively used in ASR) and fused systems comprising of GMMs and/or average pitch are also tested.

2. PROPOSED AGI SYSTEMS

2.1 AGI System

To construct our proposed AGI systems we use the concept of fusion of knowledge sources as used in ASR [8] and in AGI by Parris and Carey [2]. In their AGI system, Parris and Carey use a linear classifier to fuse information provided by acoustic analysis with pitch information. The systems presented in our paper are similar in that they use linear classifiers to fuse various knowledge sources. In our work the Linear Classifiers (LCs) are implemented using the simple perceptron learning rule [8]. The knowledge sources investigated are two methods of average pitch estimation and paired GMMs (PGMMs) each trained and tested with a different speech parameterisation scheme. To calculate the Average Estimated Pitch (AEP)[12] the speech signal is first divided into N overlapping frames. A frame is taken as voiced and the pitch value is determined if there is a detected peak (within a certain interval in which a reasonable pitch is expected) in the cepstrum greater than a designated threshold. To improve the accuracy, the threshold is not determined as fixed for different frames but is determined on the basis of the sum of cepstrum coefficients in a search interval. The AEP is the average pitch over all the frames identified as voiced. In an attempt to improve the accuracy of average pitch calculation we use a second method [13], which focuses on pitch calculation at predominantly steady points in the speech signal (points in voiced speech which are far from voiced/unvoiced boundaries). The Average Estimated Pitch at Steady Points AEPSP is the average of all the pitch values calculated using this second method. An example AGI system comprising of P PGMMs and AEP is shown in Figure 1. The parameterisation types used are described in Section 4 and PGMMs later in this section.

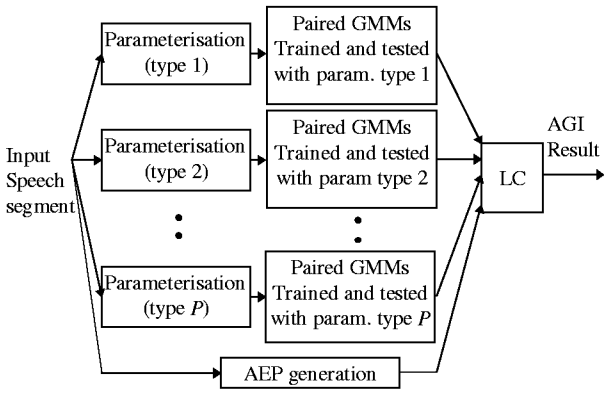

Figure 1. AGI system comprising P PGMMs and AEP

2.2 Gaussian Mixture Models for AGI

The distribution of parameterised feature vectors, for a particular speaker, can be modelled using a GM density given by [1]:

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

where M is the number of mixtures and $b_j(\bar{x})$ are uni-modal Gaussian densities, each characterised by mean vector μ_j and covariance matrix Σ_j; p_j are corresponding mixture weights. In our AGI system paired GMMs (PGMMs) are used for each speech parameter type. For each pair, one GMM (g = *female*) is trained with parameterised data from a general population of female speakers and the other from a general population of male speakers (g = *male*). Given T parameterised feature vectors for speech segment X, the following expression is computed for each GMM in a pair:

$$S_g = \sum_{t=1}^{T} \log(p(\bar{x}|(p_j,\mu_j,\Sigma_j)) \quad (2)$$

A score $\Gamma(X) = S_{female} - S_{male}$ is then taken as the output of the PGMM.

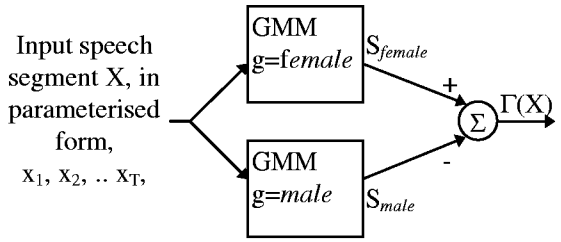

Figure 2. PGMM component of AGI system.

3. ADVERSE CONDITIONS

In the era of rapidly expanding mobile telecommunication services AGI systems may receive input from mobile terminals. These systems may use coded speech. The proposed AGI systems are evaluated with coded speech. Several standard coders are used. These are: LPC10 [5], GSM [3] and CELP [4] coders which have bit rates of 2.4, 4.8 and 13 Kbits/s. The AGI systems are also investigated for speech affected by reverberation, an obstacle often encountered in forensic applications as well as in secure database access with hands free telephones. For this study we make use of Allen and Berkley's image method [9] to simulate rectangular room responses in the time domain. An acoustically reverberated speech signal $r(n)$ can be expressed mathematically as the convolution of a clean speech signal $s(n)$ with an enclosure (room) impulse response $h(n)$: $r(n) = s(n) * h(n)$. where $h(n)$ is characterised by room dimensions (x_r,y_r,z_r), speaker location (x_s,y_s,z_s), microphone location (x_m,y_m,z_m) as well as a wall reflection coefficient for each wall β_k where k, $k \in [1..6]$)

4. SPEECH DATA COLLECTION AND PRE-PROCESSING

For each gender, 177 speakers are selected from the TIMIT database [6]. This database provides clean wideband speech data sampled at 16 kHz for speakers from various dialectic regions in the United States of America. For each of the selected speakers the three text-independent speech segments provided by TIMIT are used. The 177 speakers are divided up into training, test and validation sets (each speaker appeared in only one of these sets). To increase the amount of test and validation segments all speech segments in these sets are split into two equal segments. All speech segments are decimated to 8kHz as the speech coders used have been designed for processing speech sampled at 8kHz. Silent parts and low energy segments were removed since they are poor indicators of speaker identity. The details of the speech set allocation are as summarised in Table 1. The speech parameterisation schemes used to reduce the speech signal for the PGMMs included: Mel-based cepstral coefficients (Mel-cep), Reflection coefficients (Ref.), Autocorrelation coefficients (Auto.) and Log Area Ratios (LAR). For these parameterisation schemes an analysis filter of order 15 is used and speech segments are divided into overlapping 32ms speech frames with 22ms overlap between adjacent frames.

5. EXPERIMENTS

In the following experiments, clean speech and speech affected by 7 adverse conditions are evaluated for a variety of AGI schemes. For all experiments, PGMMs are trained using training data, LCs are trained using validation data, and quoted test results are obtained by testing the AGI systems with the test set. AGI systems used are constructed using all possible combinations of

Table 1: Speech set allocation

	Training		Test		Validation	
	male	fem	male	fem	male	fem
Speakers	33	33	90	90	54	54
Speech segs per speaker	3	3	3	3	3	3
Speech segs per speaker after splitting	N/A	N/A	6	6	6	6
Total number of speech segs	99	99	540	540	324	324
Average duration of (split) speech segs	3.22s	3.41s	1.57s	1.63s	1.65s	1.73s
Average duration of (split) speech segs after silence rem	1.62s	1.62s	0.88s	0.90s	0.92s	0.91s
Max and min duration (split) speech segs after silence rem	N/A	N/A	2.03s 0.21s	1.92s 0.15s	2.11s 0.15s	1.94s 0.21s
Amount of speech used per speech seg (after silence rem and splitting)	0.7s	0.7s	All Avail	All Avail	All Avail	All Avail

inputs to the LC. Thus given 4 parameter types for the PGMMs and two average pitch detection schemes a total of 63 AGI systems may be evaluated. Three of the adverse conditions are generated by passing the clean speech through the three coders. The remaining 4 adverse conditions are generated by convolving the clean speech with 4 different room impulse responses. The reverberant environments of interest contain a number of constants which include room dimensions (4.5x3.3x3.4 m^3), speaker position (center of the room) and microphone position (4m,0m,2m). Given reverberation times of 0.3, 0.6, 1.2, and 2.3 s the speech signal ranged from lightly degraded to virtually unintelligible, respectively. These times correspond to a β_k of 0.8, 0.9, 0.95 and 0.97 respectively, given our model of the room. Adverse condition test sets are generated by degrading the clean test set by one of the adverse conditions. Four sets of experiments are conducted as follows:

Experiment 1 is designed to evaluate the ability of the 63 AGI systems to cope with adverse conditions given that they are trained entirely with clean speech (i.e. clean data is used for training and validation). The AGI systems are tested with the clean test set as well as with test sets for each of the seven adverse conditions.

Experiment 2 is designed to provide an indication of how well the 63 AGI systems can handle each of the adverse conditions given that the problem space is narrowed down to that condition only. This experiment has eight subsections. Each subsection uses speech affected by only one different condition (i.e. training, test and validation speech data are affected by the same condition). Conditions are no adverse environment (clean speech) and the 7 adverse environments.

Experiment 3 is designed to provide an indication of how well the 63 AGI systems can handle each of the adverse conditions given that the problem space is narrowed down to that condition and clean speech. This experiment has four subsections. The first three subsections each use a different coder. For each of these three subsections the training data is clean data, the validation data consists of clean data plus coded data, and the test data is also clean as well as coded data. The fourth subsection is for investigating reverberation. We cannot simply use the same room impulse response for validation and testing. This is because in a real life situation the position of the speaker (e.g. hands free telephony) or the room reflection characteristics may not be known. Therefore we use a different speaker position (0.5m, 1.65m, 0.5m)) and β_k = 0.9 to generate an additional room impulse response. For this subsection the training data is clean speech, the validation data is clean speech plus data affected by the additional room impulse response, and the testing data is clean speech, as well as speech affected by the four room impulse responses.

Experiment 4 is designed to see if the 63 AGI systems can be improved to handle all of the different adverse conditions. In this experiment the training speech is clean speech, the validation speech is clean speech plus coded speech using the LPC10 coder, plus coded speech using the CELP coder plus coded speech using the GSM coder plus reverberated speech using the additional room impulse response of Experiment 3. The test speech is the clean speech plus seven adversely affected speech sets (one for each of the seven adverse conditions).

Table 2. Single coefficient results (the LC has only 1 input, that shown in each column) averaged over clean plus all adverse conditions.

Exp. no.	PGMM Mel-cep	PGMM Auto	PGMM Ref	PGMM LAR	AEP	AEPSP
1	97.8%	92.9%	94.1%	96.2%	96.1%	94.0%
2	97.6%	95.2%	96.7%	96.9%	96.4%	94.5%
3	97.2%	93.6%	93.2%	94.6%	96.1%	94.2%
4	97.1%	91.7%	93.9%	93.8%	96.1%	93.7%

From Table 2 it may be seen that the best single input to the LC is the output of the Mel-cepstral coefficient trained PGMM. The LCs with input from one PGMM trained with either autocorrelation or reflection coefficients, have the highest improvement from Experiments 1 to 2. This may indicate that the PGMMs are not able to generalise well from the clean to the adverse condition affected test data. A possible explanation for this could be that these two parameterisation schemes exhibit greater variability between the clean and adverse condition cases than the other two parameterisation schemes. From Table 3 it may be seen that PGMMs trained with Mel-cepstral

coefficients and at least one of the average pitch detection methods feature in all the most accurate AGI systems for the four experiments. Furthermore, it is interesting to note that the most accurate systems all consist of systems with at least three inputs to the LC.

Table 3: Top 12 AGI systems performance summary

	Experiment Number			
	1	2	3	4
Highest accuracy	97.8%	98.6%	98.3%	98.5%
Most accurate AGI system's LC inputs	Mel-cep Ref. AEPSP	Mel-cep Auto. AEP	Mel-cep Auto. LAR AEPSP	Mel-cep Ref. LAR AEP AEPSP
Average accuracy of the top 12	97.4%	98.5%	98.2%	98.3%

PGMMs trained with Mel-cepstral coefficients also featured in every one of the top 12 AGI systems. The increase in the LCs amount of training conditions from Experiments 1 to 3 to 4 is accompanied by a respective increase in the average number of inputs into the LC from 3.5 to 3.7 to 4.3 for the top 12 AGI systems. From Table 4 it may be seen that AGI systems trained solely with clean speech have a marked decrease in accuracy when tested with speech subjected to adverse conditions (Experiment 1). From the table it may also be seen that AGI systems trained and tested with speech from the same adverse conditions are able to produce high accuracy (Experiment 2). Finally it can be seen that AGI systems with LCs trained with data affected by adverse conditions have an accuracy somewhat between these two extremes (Experiments 3 and 4).

Table 4: Averaged accuracy results across the 63 AGI systems for the four experiments

	Experiments			
Test Data	1	2	3	4
CLEAN	98.1%	98.1%	98.1%	98.0%
GSM	98.0%	98.2%	98.1%	98.0%
CELP	97.7%	97.9%	97.7%	97.9%
LPC10	97.7%	98.1%	98.1%	98.1%
$\beta_k = 0.8$	94.8%	97.9%	96.8%	96.8%
$\beta_k = 0.9$	93.8%	97.4%	96.1%	95.9%
$\beta_k = 0.95$	92.8%	96.7%	95.0%	95.2%
$\beta_k = 0.97$	91.3%	96.8%	94.4%	94.8%

6. CONCLUSIONS

This paper evaluates 63 AGI systems for clean speech segments and speech segments affected by coding or reverberation. The average duration of the speech segments (after silence removal) used for testing the AGI system is 890 ms (minimum duration 150ms, maximum duration 2030ms). We have shown that it is possible to build an AGI system able to handle the LPC10, CELP and GSM coders with no significant loss in accuracy. It is also possible to build AGI systems to significantly reduce the impact of even severe reverberation by subjecting AGI training data with a different estimated room response. The best AGI system had an accuracy of 98.5% averaged over all possible conditions (clean and adverse).

7. REFERENCES

[1] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", Speech Communication, v17, pp91-108, 1995.

[2] E.S. Parris and M.J. Carey, "Language Independent Gender Identification", ICASSP, pp 685-688, 1996.

[3] P. Kroon and E.F. Deprettere, "A Class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbits/s", IEEE J. on Sel. Areas in Communications, v6, no. 2, pp353-363, 1988.

[4] J.P. Campbell Jr, T.E. Tremain, and V.C. Welch, "The DoD 4.8 KBPS Standard (Proposed Federal Standard 1016)", Advances in Speech Coding, Kluwer, 1991.

[5] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, "Discrete-time Processing of Speech Signals", Maxwell Macmillian Int., 1993.

[6] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status", DARPA Workshop Speech Recog., pp93-99, 1986.

[7] P.J. Castellano, S. Sridharan, and D. Cole, "Speaker Recognition in Reverberant Enclosures", ICASSP, May 1996, pp 117-120.

[8] S. Haykin, "Neural Networks: A Comprehensive Foundation", Maxwell Macmillian Int., 1994.

[9] J.B. Allen and B.A. Berkley, "Image Method for Efficiently Simulating Small Room Acoustics", J. Acoust. Soc. Am. Vol. 65, No. 4, pp 943-950, 1979.

[10] L. Rudasi and S. Zahorian, "Text-Independent Talker Identification with Neural Networks", ICASSP, Vol. 1, pp 389-392, 1991.

[11] J. Leis, M. Phythian, and S.Sridharan, "Speech Compression with Preservation of Speaker Identity", ICASSP, pp 1711-1714, 1997.

[12] S. Ghaemmaghami, M. Deriche, and B. Boashash, "Formant Detection Through Instantaneous-Frequency Estimation using Least Square Algorith", ISSPA, pp 81-84, Aug. 1996.

[13] S. Ghaemmaghami and M. Deriche, "A New Approach to Efficient Interpolative Determination of Pitch Contour using Temporal Decomposition", IEEE TENCON, pp 125-130, 1996.

[14] K. Wu and D.G. Childers, "Gender Recognition from Speech. Part I: Coarse Analysis", J. Acoust. Soc. Am., Vol. 90, No. 4, pp 1828-1840, 1991.