GENDER GATES IN DEGRADED ENVIRONMENTS

Stefan Slomka, Peter Barger, Pierre Castellano and Sridha Sridharan
SPRC, School of Electrical and Electronic Systems Engineering
Queensland University of Technology, Brisbane 4001, Australia

ABSTRACT: The present paper extends the investigation of gender gates proposed in (Barger et al, 1996) to speech degraded by coding and, separately, room reverberation. Coded speech did not degrade gate accuracies relative to the uncoded case in (Barger et al, 1996). Reverberation slightly degraded gate accuracies although this was only weakly dependent on reverberation time.

INTRODUCTION

In our first paper (Barger et al, 1996), gender gates are proposed for Automatic Speaker Recognition (ASR) and investigated under reasonably favourable experimental conditions. In that work, the main source of degradation is signal band limitation, channel effects and variability of available handsets, all of which are due to the use of the Switchboard telephone database (Godfrey et al, 1994). However, in real world situations, the speech signal utilised for ASR is often further degraded by other factors such as

- noise
- coding and
- reverberation.

The present work directly extends that discussed in (Barger et al, 1996) by considering the effects of speech coding and, separatly, reverberation on gender gate performance. Speech coding is conducted with standard coders (GSM and LPC10). This speech is of importance since ASR may be needed for the rapidly expanding cellular mobile telephone market. The situation would primarily address applications such as information release over a mobile telephone. It has been confirmed that speech coding slightly degrades ASR (Castellano and Sridharan, 1996). Reverberated speech is obtained using an image method. The method closely models that obtained in a rectangular (unfurnished) enclosure. This type of signal obtained is that which must be dealt with when forensic recordings are made in enclosures or for voice activated security access when a microphone governs entrance into one room from another room (Castellano and Sridharan, 1996). A series of text-independent experiments were conducted to determine the effects of coding and, separately, reverberation upon gender separation using the gender gates proposed in our first paper (Barger et al, 1996) as well as the same individual classifier types.

GENDER DISCRIMINATION WITH CODED SPEECH

GSM and full rate LPC10 coders

Two speech coding schemes were studied: the full rate GSM coder, commonly used for mobile telephony and the LPC10 coder used in military applications. The GSM coder uses the Analysis-By-Synthesis Linear Predictive Coding (ABS-LPC) scheme (Kroon and Deprettere, 1996). During encoding, an excitation sequence minimising weighted mean square error between original and synthesised speech is determined. This sequence consists of pulses spaced evenly over a frame. During decoding, speech is reconstructed from the quantised excitation and prediction parameters. The LPC10 coder differs from the GSM model in that its excitation is periodic for voiced speech and Gaussian noise for unvoiced speech. GSM and LPC10 coders are used at 13 and 2.4 kbits/s coding rates respectively.

Experiment

Gender gates evaluated in this paper were limited to those which performed best in (Barger et al, 1996). These incorporated Mahalanobis distance classifiers with or without pruned GMBCMs with HONNs (trained exclusively with LSPs). In all, 63 different gates were retained out of the initial 126 described in (Barger et al, 1996). For each of the LPC10 and GSM coders, 2 series of gender discrimination experiments were conducted depending on whether the front-end classifiers were trained with or without coded speech. These experiments encompassed 63 gates, pitch, 4 speech parametrisation schemes (for the Mahalanobis distance classifiers only), 60 training speakers, 44 test speakers and 50 validation speakers (used to manufacture training data for the

Coder type	Training data type used for Perceptron	Mean number of speakers being misclassified on the basis of gender
Uncoded training data, coded test data		
LPC10	uncoded	10.4
	coded	9.7
GSM	uncoded	7.1
	coded	7.4
Coded training and test data		
LPC10	coded	9.8
GSM	coded	8.7

Table 1: Mean discrimination accuracies over 63 gates for 22 males, 22 females and coded speech

Perceptron). As in ,(Barger et al, 1996) the Perceptron was trained with output data from the front-end classifiers whose input was speech from the 50 validation speakers. When the front-end classifiers were trained with coded speech the validation speech, used to generate input for the Perceptron, was also coded. When the front-end classifiers were trained with uncoded speech, that validation speech was either coded or uncoded as shown in Table 1. Mean discrimination accuracies, expressed for all 63 gates are shown in that table. Mean number of speakers correctly identified on the basis of gender, over the 63 gates was similar to what could be obtained for uncoded telephone speech in (Barger et al, 1996). The most accurate gates for each coder and training data type for the back-end Perceptron are reported in Table 2 along with their performances. The main points to notice are that

- gender gates based on a single classifier were outperformed by those based on classifier fusion,
- unlike, in the previous series of experiments (with uncoded speech), complex gates relying on the fusion of several Mahalanobis distance classifiers outputs were systematically outperformed by gates based on less fusion (except in the case of the GSM coder and coded training speech for the back-end Perceptron),
- gender discrimination based on GSM coded speech slightly outperformed that based on LPC10 coded speech, given the present limited data and
- The most accurate gates were as accurate as the best gates tested with uncoded speech.

While the use of connectionist technology (HONN, see (Barger et al, 1996)) did not degrade the better performing gates, it did not enhance gender discrimination accuracies either. (The only exception to this is for the LPC10 coder and uncoded training speech both for the front and back end classifiers, where the use of a Mahalanobis distance classifier was detrimental to accuracy.) The GSM, with its high coding rate, is a better quality coder than the LPC10. It was thus expected that the former would lead to superior gender separation accuracies.

GENDER DISCRIMINATION IN REVERBERANT ENVIRONMENTS

The reverberant model

An acoustically reverberated speech signal $x(n)$ can be expressed mathematically as the convolution of a clean speech signal $s(n)$ with an enclosure (room) impulse response $h(n)$:

$$x(n) = s(n)oh(n). \tag{1}$$

That response is defined by the transmission properties between source and receiver. It is dependent on enclosure volume and therefore size, wall surface reflection coefficient, source and physical content of the enclosure. Reverberation smears all sound and therefore speech. Any low frequency energy present in the speech signal is especially affected (Nabelek et al, 1993). The present study makes use of Allen and Berkley's image method (Allen and Berkley, 1979). The method is able to simulate a rectangular room response in the

Coder type	Gate configuration using feature and classifier fusion. Symbols are abbreviations*	Training data type used for Perceptron	Number of speakers misclassified on the basis of gender
Uncoded training data - coded test data			
LPC10	GMBCM	uncoded	2
	cep pitch	coded	2
GSM	ref pitch	uncoded	2
	lsp pitch	uncoded	2
	lsp GMBCM	uncoded	2
	ref pitch	coded	1
Coded training - coded test data			
LPC10	cep pitch	coded	3
	ref cep	coded	3
	ref autoc	coded	3
	ref GMBCM	coded	3
GSM	lsp cep pitch	coded	1

* *ref: Mahalanobis distance and reflection coefficients*
cep: Mahalanobis distance and Mel-based cepstrum coefficients
lsp: Mahalanobis distance and Line Spectrum Pairs (LSPs)
autoc: Mahalanobis distance and autocorrelation coefficients
pitch: pitch alone
GMBCM: pruned HONNs (S=1) and LSPs

Table 2: Most accurate gate for 22 males, 22 females and coded speech, as a function of coder type and training data type for back-end perceptron

time domain. A speaker is modelled as a point source. Assuming a rigid smooth wall, the associated boundary condition is satisfied by positioning an image symmetrically behind the wall. This image, being a point source, is also imaged so that for a 6 wall system an infinite number of images are produced. In practice, walls are non rigid and the present image method is not exact but it provides a good approximation to the real world situation. An enclosure impulse response is given by:

$$p(t,X,X^{'}) = \sum_{p=0}^{1} \sum_{r=-\infty}^{+\infty} \beta_{x1}^{|n-q|}\beta_{x2}^{|n|}\beta_{y1}^{|l-j|}\beta_{y2}^{|l|}\beta_{z1}^{|m-k|}\beta_{z2}^{|m|} * \frac{\delta\left[t-|R_p+R_r|/c\right]}{4\Pi|R_p+R_r|}, \qquad (2)$$

where: t is time,
p is (q,j,k),
r is (n,l,m),
X is speaker location in 3 dimensional space: (x,y,z),
$X^{'}$ is microphone location in 3 dimensional space: $(x^{'},y^{'},z^{'})$,
R_p is $(x-x^{'}+2qx^{'},y-y^{'}+2jy^{'},z-z^{'}+2kz^{'})$,
R_r is $2(nL_x,lL_y,mL_z)$, $((L_x,L_y,L_z)$ being enclosure dimensions),
c is the speed of sound at sea level,
β_k, is wall reflection coefficient: $(1-\alpha_k)^{1/2}$ (α being the Sabine energy absorption, coefficient for wall k, $k \in [1..6]$) and
$\delta()$ is the delta function.

Allen and Berkley's image method implementation allows for a comprehensive range of enclosure parameters to be accurately controlled.

Experiment

The reverberant environment of interest here consisted of a number of constants which include room dimensions (8x6x4 m^3) and speaker to microphone separation (3 m). β_k was varied between 0.75 (low reflectiveness) and 0.97 (high reflectiveness) and was taken to be the same for all 4 walls, ceiling and floor, at any one time. (For a β_k equal to 1, an impulse response was not attenuated when being reflected off a solid surface.) The same 63 gates as well as males and females, investigated for speech coding were again studied here. (All parametrisation schemes, as well as pitch, were also retained.) Given reverberation times of 0.18, 0.31

to 0.72 and 2.25 s the speech signal was, respectively, lightly degraded, moderately degraded and virtually unintelligible. (These times correspond to a β_k of 0.75, 0.8, 0.85, 0.9 and 0.97 respectively, given our model of the room). For all gender gates, the back-end Perceptron was trained using two possible data sets which included

- output from front-end classifiers whose input was non reverberant speech from the 50 validation speakers (N.R.) and

- as above but with validation speech reverberated with a standard impulse response corresponding to a corner of the room (two metres away from each wall) (S.R.).

Thus the characteristics of the standard impulse response were different from those which affected the test speech since the impulse response, in this case, was taken to be the centre of the room. (This discrepancy was necessary since, in a real world situation, the exact characteristics of reverberation contaminating test speech are not known because the associated impulse response is not known.)

Figure 1 illustrates mean percentages of correct discriminations for a back-end Perceptron trained using either N.R. or S.R.. The use of N.R. led to a sharp degradation in mean gate accuracy with increasing reverberation time. By substituting N.R. with S.R., mean gate accuracy did not change significantly as a function of reverberation time, except for heavy reverberation (2.25 s). However, even for the latter reverberation time, mean discrimination accuracy was only 2 per cent below that recorded in the anechoic case (see Figure 1). Moderate reverberation (0.31 to 0.45 s) was even seen to benefit the gender discrimination problem, given S.R. It may be that this degree of reverberation smears low energy speech with high energy speech such as vowels. Unlike the latter, the former is poorly indicative of speaker identity (Rudasi and Zahorian, 1991).

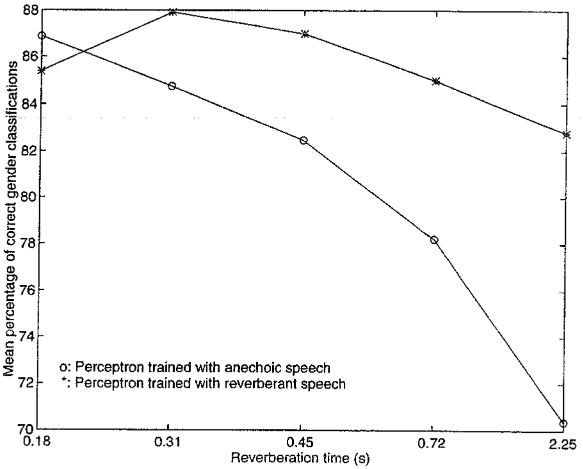

Figure 1: Mean percentage of correct gender classifications, over 63 gender gates, for 22 male and 22 female speakers, as a function of reverberation time. The gates' back-end Perceptron is trained with output from front-end classifiers whose input is either anechoic or reverberant speech from 50 separate speakers.

The most accurate gates, under the present reverberant conditions are listed in Table 3. Conclusions are analogous to those reached in the case of coded speech, notably

- gender gates based on a single classifier were outperformed by those based on classifier fusion,

- complex gates relying on the fusion of outputs from many classifier types were systematically outperformed by gates based on less fusion, irrespective of reverberation time and

Reverberation time (s)	Training data type used for Perceptron	Gate architecture*	Number of speakers misclassified on the basis of gender
0.18	N. R.	ref pitch	2
	N. R.	GMBCM pitch	2
	S. R.	cep pitch	1
0.31	N. R.	ref pitch	2
	N. R.	lsp pitch	2
	N. R.	cep pitch	2
	N. R.	autoc pitch	2
	S. R.	cep pitch	1
0.45	N. R.	cep pitch	1
	S. R.	ref pitch	1
	S. R.	lsp pitch	1
	S. R.	cep pitch	1
0.72	N. R.	cep pitch	1
	S. R.	pitch	1
2.25	N. R.	cep pitch	2
	S. R.	ref autoc	1
	S. R.	lsp autoc	1
	S. R.	cep autoc	1

* *ref: Mahalanobis distance and reflection coefficients*
cep: Mahalanobis distance and cepstrum coefficients
lsp: Mahalanobis distance and LSPs
autoc: Mahalanobis distance and autocorrelation coefficients
pitch: pitch on its own
GMBCM: pruned HONNs (S=1) and LSPs

Table 3: Most accurate gates for reverberant speech, 22 male and 22 female speakers as a function of reverberation time and training data type for back end perceptron

- gates were more robust when their back-end Perceptron was trained with output from front-end classifiers whose input was reverberant rather than anechoic.

The accuracies obtained were probably sufficiently high for associated gender gates to be used in a forensic context as front-ends to ASR systems. This view is supported by three arguments which are

1. If a speaker's gender is misclassified by a gate the speaker may still be rejected at the speaker verification stage,
2. In most cases, no other means of person authentication are available and
3. ASR evidence is accepted as supportive only, in a court of law.

Another problem which has not been covered here is that of reverberation contaminating speech spoken into a handset during a telephone conversation. This reverberation would be caused by the enclosure containing the speaker and handset (thus be separate from channel induced distortion). Distance between source (mouth) and receiver is, at most, 10 cm. In the case of ASR, such a short separation results in minimal degradation in speaker discrimination accuracy even when classifiers are trained with non reverberant data (Castellano and Sridharan, 1996). Since we believe gender separation to be a simpler problem, we expect the impact of enclosure reverberation upon gender discrimination based on hand held telephones to be very small indeed.

However, this may not be the case for hands free telephony.

CONCLUSION

Gender separation experiments were conducted under degraded acoustical environments. The same 63 gender gates retained in our other paper (Barger et al, 1996) were again investigated in the present work. Switchboard's test signal was first coded using, separately, LPC10 and GSM coders. The latter series of experiments were repeated using Mahalanobis distance and HONN classifiers only (63 gates). Coding the classifiers' training data did not enhance results. The most accurate gates were as accurate as the best gates tested with uncoded speech.

Finally, the (uncoded) and previously unreverberated Switchboard signal was contaminated with reverberation and experiments repeated once more for the 63 gates. The environment modelled that of a room with fixed speaker to microphone separation using an image method. Training data were anechoic. By training the Perceptron with reverberated data affected by a different room impulse response to that present in the test data (because the latter is unknown in practice), mean gate accuracy did not change significantly with reverberation time (except for heavy reverberation).

The most robust all-round gates, in (Barger et al, 1996) and in the present work, consisted of 2 Mahalanobis distance classifiers with fused outputs or pitch fused to the output on one such classifiers. The best all-round speech parameters were reflection and Mel-based cepstrum coefficients. Future work should use a greater number of test speakers and investigate whether a more powerful back-end classifier than a Perceptron would benefit the more complex gates discussed in this study.

REFERENCES

Allen, J. B. and Berkley, B. A. (1979) "Image Method for Efficiently Simulating Small Room Acoustics", J. Acoust. Soc. Am. 65(4) ,pp. 943-950.

Barger, P., Slomka, S., Castellano, P. and Sridharan, S. (1996) "Gender gates for Automatic Speaker Recognition", In Proc. of the Sixth Australian International Conference on Speech Science and Technology.

Castellano, P. and Sridharan, S. (1996) "Effects of Speech Coding on Speaker Verification", Electr. Let. 32(6), 517-518.

Castellano, P. and Sridharan, S. (19960 "Speaker Recognition in Reverberant Enclosures", In Proc. of the International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 117-120.

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.

Kroon, P. and Deprettere, E. F. (1988) "A Class of Analysis-By-Synthesis Predictive Coders for high Quality Speech coding at Rates Between 4.8 and 16Kbit/s" J. Sel. Areas Comm. 6(2), pp. 353-363.

Nabelek, A. K., Czyzewski, Z. and Crowley H. J. (1993) "Vowel Boundaries for Steady-State and Linear formant Trajectories", J. Acoust. Soc. Am. 94 (2), pp. 675-687.

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.