SPECTRAL PATTERNS AND SPEAKER IDENTIFICATION ASYMMETRY

S. Ong and P. Castellano

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT - This paper investigates the phenomenon of Automatic Speaker Identification asymmetry. Automatic Speaker Identification attempts to identify one or several speakers in a multi-speaker environment by analysing the speech signal. In this situation, some speakers may routinely be mistaken for others while the latter are rarely identified as the former. The MUltiple SIgnal Classification algorithm was used to provide an eigenvector approach, partitioning the problem space into signal plus noise and noise eigenvalue subspaces. A speaker associated with high eigenvalues in the first subspace (well defined spectral patterns) was not routinely mistaken for another. The opposite was found for those with low eigenvalues in that subspace (weak spectral patterns). Hence, a disparity in spectral pattern strengths from speaker to speaker may strongly influence asymmetry in Automatic Speaker Identification.

INTRODUCTION

The aim of Automatic Speaker Identification (ASI) is to identify the person most likely to have spoken from within a reference database of speakers. This is a maturing science which draws on probabilistic techniques and Artificial Neural Networks (ANNs). While good results have been reported for some time (Matsui and Furui, 1991; Rudasi and Zahorian, 1991), the core problem of minimising "false rejections" and "false acceptances" (O'Shaughnessy, 1986) is a matter of ongoing research.

This paper focuses on asymmetry arising from a speaker being erroneously identified as another while the latter is less frequently (if ever) mistaken for the former. This phenomenon is important since it influences ASI scores, thus ASI outcome. A similar type of asymmetry has been reported in the field of computer vision by (Mumford, 1991).

CHOICE OF PARAMETRIC REPRESENTATION FOR SPEECH

A broad range of parametric representations for speaker recognition, using the Linear Prediction Coding (LPC) method, are available. LPC parameters are adversely affected by the frequency response of the recording device. However, for cepstral coefficients, this effect may be nullified by subtracting time averages over an entire utterance. Cepstral parameters are independent of the recording device characteristics in addition to allowing the best ASI performances (Sethuraman and Gowdy, 1989). Because of these considerations, a mel-based cepstral representation was selected for the present study.

The speech signal used was provided from tapes belonging to an audio-visual library. These tapes contained the voices of twenty English speaking males. The signal was digitised at 10 kHZ (12 bits). It consisted of 55 second frames, each frame having been subdivided into 40 segments. The segments were high-frequency pre-emphasised with transfer function $1\text{-}0.98z^{-1}$ then windowed using a non overlapping 256 point Hamming Window.

EXPERIMENTAL FRAMEWORK

Classifiers for Automatic Speaker Identification

Interactive Laboratory System (ILS). This is a well established general purpose signal processing and analysis package (Signal Technology Inc, 1985). It supports speaker pattern analysis and identification. ASI is a two step process. Firstly, the Statistical Measures (SME) command is used to process raw data according to a Means and Inverse Covariance Matrix algorithm. The Best Fit

Pattern Recognition (BPA) algorithm is then applied to the output. BPA performs weighted Euclidean distance computations between the set of reference patterns (generated by the SME) and the test data. The latter is typically an independent set.

Higher Order Neural Network (HONN). This a feed-forward ANN architectures also known as Functional-link networks (Pao, 1989). The flat version of this network (Castellano and Shridharan, 1993) was used in this study. HONN allows for each acoustic parameter, in a vector, to be expanded into a higher dimensional data space (Hush and Salas, 1989). This was made possible by third order tensor link transformations applied to the input of the ANN. Each cepstral component x_i ($0<i<14$) was expanded into the four components: x_i, $x_i x_{i+1}$, $x_i x_{i+2}$ and $x_i x_{i+1} x_{i+2}$. x_{14} and x_{15} were expanded as x_{14}, $x_{14}x_{15}$, $x_{14}x_1$, $x_{14}x_{15}x_1$ and x_{15}, $x_{15}x_1$, $x_{15}x_2$, $x_{15}x_1x_2$ respectively. The expanded vector was then processed at the output layer where the number of processing elements was set equal to the number of speakers present in the data.

Eliminating classifier bias

ILS and the HONN were considered separately in the ASI role. Being radically different technologies, any results they share are expected to be free of both systems' inherent bias. Let P be the number of speakers present in the data and N the number of ASI systems available. Let $(i,j) = (\{1,2,...,P\})^2$ and n a positive integer less than N. A composite confusion matrix $\{M_{comp}: M_{comp,i,j}\}$may be deduced from N confusion matrices $\{M_n: M_{n,i,j}\}$(one for each ASI system available) by applying the simple formula

$$M_{comp,i,j} = \frac{N \prod_{n=1}^{N} M_{n,i,j}}{\sum_{n=1}^{N} M_{n,i,j}} \qquad (1)$$

In this case P=20 and N=2. The composite matrix will only include results shared by (and averaged over) all N matrices. The magnitude of matrix components will reflect the extent to which the corresponding spectral characteristics are shared amongst the N matrices. Once the composite matrix is obtained, a measure of intra-speaker asymmetry with respect to the diagonal can be deduced. For a given speaker x, this is expressed as

$$ASYM_x = \sum_{j=1}^{P} \left| (\text{column } x,\ \text{row } j) - (\text{row } x, \text{column } j) \right| \qquad (2)$$

The Music Algorithm - Review

In parallel with the computation of matrices, the MUltiple Signal Classification algorithm (MUSIC) will be used to orthogonalise speech frames (both training and testing). MUSIC is a high resolution frequency estimation algorithm (Haykin, 1991). When used in ASI, the operation yields eigenvalues representative of each speaker's spectral patterns. MUSIC is briefly reviewed here.
For an input signal consisting of L uncorrelated zero-mean complex sinusoids (angular frequencies: $\omega_1,..,\omega_L$ (ω: omega) and average powers $P_1,..,P_L$) and zero-mean additive white noise with variance σ^2 (σ: sigma), the ensemble-averaged correlation matrix R is given by

$$R = SDS^H + \sigma^2 I \qquad (3)$$

I is an $(M+1)^2$ identity matrix. D the diagonal matrix and S an (M+1)-by-L *frequency matrix* are given respectively by:

$$D = \text{diag}(P_1,..,P_L) \qquad (4)$$

$$S = \begin{bmatrix} 1 & \cdots & 1 \\ \exp(-j\omega_1) & & \exp(-j\omega_L) \\ \cdot & \cdots & \cdot \\ \cdot & & \cdot \\ \exp(-jM\omega_1) & \cdots & \exp(-jM\omega_L) \end{bmatrix} \tag{5}$$

Let $\lambda_1,..,\lambda_{M+1}$ be the eigenvalues of R and $\mu_1,...,\mu_{M+1}$ be those of SDS^H then, from eq. (3),

$$\lambda_i = \mu_i + \sigma^2 \qquad i = 1,..,M+1 \tag{6}$$

If $q_1,..,q_{M+1}$ are the eigenvectors of R, the (M+1 -L) eigenvectors corresponding to the smallest eigenvalues of R are given by:

$$Rq_i = \sigma^2 q_i \tag{7}$$

or, equivalently

$$(R - \sigma^2 I)q_i = 0 \tag{8}$$

where i = L+1, .. , M+1. It can be shown that the subspaces spanned by the eigenvectors $q_1,..,q_L$ and $q_{L+1},..,q_{M+1}$ are the orthogonal complements of each other. The first subspace is known as the noise subspace. It corresponds to the (M+1-L) smallest eigenvalues of R. These are ideally equal to zero since the L sinusoids have distinct frequencies and therefore the L columns of S are linearly independent. The second subspace is called the signal plus noise subspace and corresponds to the L largest eigenvalues of R. MUSIC can be used in ASI to identify those features present in spectral patterns which most facilitate the patterns' separation by solving an eigenvalue problem.

RESULTS

Automatic Speaker Identification

ASI results obtained for ILS and the ANN are illustrated in Tables 1 and 2 respectively. The mean percentage of correctly recognised speech frames for ILS, given one hundred test frames, was 60. The lowest ASI score was 31 (twice). The mean percentage of correctly recognised speech frames for the ANN, given one hundred test frames, was 53. The lowest ASI score was 33. ILS was able to classify frames with 7 percent more accuracy on average than the ANN. Despite this, the ANN's ASI threshold at 33% (lowest ASI score found over the speaker population) was 2% higher than for ILS.

Table 1. SI performance for the ILS system and twenty speakers

Speakers

Speakers	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	31	18	2	5	2	3			3		18		16						2	
B	11	59	1	2	1	1			3		17		3				1		1	
C	4	3	36	3	32	2	1	1			12		1					1	4	
D	12	10	3	44	1		1				25		3					1		
E	1	6	6	2	63	3		1			9		1					4	4	
F	1	1			1	96							1							
G					1	3	90					1	1					4		
H			1		3		2	69		1	8		1			3		7	5	
I		2	1		3	22	2	2	41	4	5		18							
J						2		8	7	49	25		5		1	1		1	1	
K		2						1			96							1		
L	1		4	4	5	1	11				1	56						16	1	
M				1			2	5	1		36		40					12	3	
N			2		3	1	3	27			2		1	31	10	20				
O					2			12		2	1			13	45	23		1	1	
P							1	29			10				2	52		6		
Q		4									1		6				89			
R						1		2			2							91	4	
S			1	1	1				2	1	6		1					3	84	
T			5				3	38			5				1	2		6	1	39

Table 2. SI performance for the Functional-link ANN and twenty speakers

Speakers

Speakers	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	53	15	1			1					8	1	8				7		5	1
B	19	43		2	4	1					9		9				11			2
C			42	1	20	1					11	2	10				3		7	3
D	9	4	1	49	1						27	3	3				2		1	
E		6	5	1	47	3				1	10		10		1		5	7	2	2
F						91	4		5											
G				4	2	13	53					6				1		1	19	1
H		1	1	1				39		2	17					2	4	18	12	3
I						17			33	7	10		11		10	5	2		3	2
J					1	1		13	9	38	12		11			1	12	1	1	
K		8	2	3	4				1	1	64		4			1	1	11		
L			5	7	6	2	13				1	44		2			2	18		
M	4		2	1	3			5		6	24		35			7	1	10	2	
N					1	10	1	11		6		9		36	3	9		9	1	4
O					4		1	13	1					5	42	31	1	1		1
P		1						6			9		19		23	41				1
Q											2						98			
R								4			10	8	13	2			3	44	16	
S								2	4									1	93	
T								12	2							2		1		83

Inspection of Tables 1 and 2 reveals that intra-speaker asymmetry, as defined by eq. (2), was significant for both ILS and ANN results. Some of this asymmetry was particular to one ASI method and not the other. This is reflected in the results for speaker T given ILS and speaker Q given the

ANN. A composite matrix shown in Table 3 was deduced from the above two tables by applying eq.(1). Finally intra-speaker asymmetry was measured for each speaker by applying eq. (2) to the results of Table 3 to obtain Table 4 These tables indicate significant asymmetry for speakers E, H, K, M, P and R.

Table 3. Composite results (ILS and ANN)

Speakers	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	39	16	1			1					11		10						3	
B	14	50		2	2	1					12		5				2			
C			39	1	25	1					11		2						5	
D	10	6	1	46	1						26		3							
E		6	5	1	54	3					9		2					5	3	
F						93														
G					1	5	67					2						2		
H			1					50		1	11					2		10	7	
I						19			36	5	7		14							
J						1		10	8	43	16		7			1		1	1	
K		3									77							2		
L			4	5	5	1	12				1	50						17		
M				1				5			29		37					11	2	
N					1	2	1	16						33	5	12				
O					3		12							7	43	26		1		
P								10			9				4	46				
Q											1						93			
R								3			3							59	6	
S									3									1	88	
T								18								2		2		53

Table 4. Intra-speaker asymmetry for twenty speakers

Speaker	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Asymmetry	38	27	45	47	56	34	31	82	46	39	141	43	89	34	40	54	3	52	29	22

Orthogonalisation of feature sets

The MUSIC algorithm was applied in turn to each speaker's frame. A vector of 15 eigenvalues, corresponding to the speech frame size, was deduced from eq. (6), for each speaker. This eigenspectrum is shown in Fig. 1. The signal plus noise subspace corresponds to eigenvalue numbers 1 to 11 inclusive. As expected, in the noise subspace (numbers 11 to 15 inclusive), eigenvalues were small (< 1.5). The signal plus noise subspace itself may be separated into two regions, on the basis of eigenvalue magnitude. Eigenvalues are high for numbers 1 to 5 and for speakers H,K,M,R (and to a lesser extent, P and E). Eigenvalues are low for the remaining speakers (although not as low as in the noise subspace).

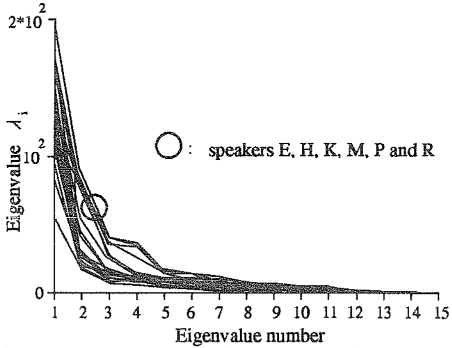

Figure 1 . Eigen-spectrum for twenty speakers.

DISCUSSION

The composite matrix of Table 3 exhibits regions of both high symmetry and asymmetry as defined by eq. (2). The symmetry may be explained in terms of the non-separable regions in overlapping patterns (Morgan and Scofield, 1991). This occurs when feature representations are shared by several classes.

Asymmetry with respect to the diagonal in ASI confusion matrices is less well understood. An examination of Table 4 and Fig. 1 indicates that those speakers for which ASI is highly asymmetrical are also those speakers for which eigenvalues are the highest in the signal plus noise subspace (eigenvalue numbers 1 to 11). This is obvious for eigenvalues 2 to 5 inclusive and is the case for speakers E, H, K, M, P and R. Hence these speakers exhibit both the sharpest spectral features and the greatest ASI asymmetry. If spectral patterns have prominent features then the classifier makes use of these features. If other patterns are without these features they are not mistaken for the first patterns. Now, if a nondescript pattern is the reference then the classifier experiences difficulty focussing on any single feature and is more likely to mistake another pattern for it.

It would be worth investigating whether asymmetry measurements or suitable transformations thereof could be substituted for unavailable a priori probabilities estimates in current probabilistic ASI approaches.

CONCLUSION

Asymmetry in Automatic Speaker Identification is expressed by some speakers being routinely mistaken for others while the latter are rarely (if ever) mistaken for the former. While bias inherent to computerised speaker identification technologies may play a part in this phenomenon it cannot fully explain it.

This study has linked high eigenvalues in the signal plus noise part of the eigen-spectrum with Speaker Identification asymmetry. It is believed that the prominent spectral features to which those eigenvalues correspond have a strong influence on the ASI discrimination process. Hence, if a speaker's pattern exhibits prominent features then the ASI solving process will look for those features so that if a pattern (belonging to another speaker) is without them it will not be retained. However, if the reference is a relatively featureless speech pattern, the ASI process will experience difficulty finding a match. It seems that in this case, many patterns become candidates for selection and in particular the better defined ones.

This study was aimed at shedding light into a lesser studied aspect of ASI. It remains to be established how future ASI systems should be designed to cater for it.

ACKNOWLEDGMENTS

The authors wish to thank David Cole for his contribution to the practical implementation of the Music algorithm.

REFERENCES

Castellano, P. & Sridharan, S. (1993) "Speaker identification with artificial neural networks", Proc. WoSPA 93, 153-159.

Haykin, S. (1991) Adaptive Filter Theory, 2nd ed., Prentice Hall, Englewood Cliffs, New Jersey, 452-454.

Hush, D. R. & Salas, J. M. (1989) "Classification with neural networks: a comparison", ISE Eleventh Annual Ideas in Science and Electronics, 107-114.

Matsui, T. & Furui, S. (1991) "A text-independent speaker recognition method robust against utterance variations", Proc. ICASSP 91 1, 377-380.

Morgan, D. P. & Scofield, C. L. (1991), "Neural Networks and Speech Processng", Kluwer Academic Publishers, Norwell, Massachusetts, 63-69.

Mumford, D. (1991) "Mathematical theories of shape: do they model perception?", Geometric Methods in Computer Vision, Society of Photo-Optical Instrumentation Engineers 1570, 2-9.

O'Shaughnessy D. (1986) "Speaker recognition", IEEE ASSP Magazine 3, 4-17.

Pao, Y. (1989) "Adaptive Pattern Recognition and Neural Networks", Addison-Wesley Publishing Company, 197-222.

Rudasi, L. & Zahorian, S. A. (1991) "Text-independent talker identification with neural networks", Proc. ICASSP 91 1, 389-392.

Sethuraman, R. & Gowdy, J. N. (1989) "A cepstral based speaker recognition system", Proc. the Twenty-First Southeastern Symposium on System Theory, 503-507.

Signal Technology Inc. (1985) Interactive Laboratory System ILS, V 5.0, Ch. 4 & 5. Goleta, California.