THE ROLE OF DYNAMICS IN VISUAL SPEECH BIOMETRICS

J.S.Mason and J.D.Brand

Department of Electrical Engineering
University of Wales Swansea
SA2 8PP, UK
email: J.S.D.Mason@swansea.ac.uk

1. ABSTRACT

This paper begins by introducing biometrics and their underlying performance factors. Biometrics are sometimes classed as either behavioural or physiological. Difficulties with these classes are discussed in terms of the importance of dynamics, highlighting the key point that definitions are clarified if the biometric information-bearing signal itself is considered. Emphasis is then given to visual speech in the form of lip profiles The case is made that these are a special case in that they privide a vehicle for a twin biometric: both behavioural and physiological. It is argued that lips might well be unique in providing a practical twin biometric. Illustration is presented in the form of practical experiments based around visual speech and lip profiles. Experimental results using short, test and training segments from video recordings give recognition error rates as: physiological lip-profiles 2% and behavioural lip-profiles 15%.

2. INTRODUCTION

Biometrics in its various forms provide the potential for reliable, convenient, automatic person recognition. Systems based on fingerprints, faces, hand geometries, iris images, vein patterns and voice signals are now all available commercially. While this is true, it is also true that the take-up of this technology has been slow. Biometrics applications act as a guard control, limiting access to something or somewhere; it is the something or the somewhere that in turn defines the potential merits of biometric-based solutions. Yet biometrics have not entered every-day life to any extent whatsoever. If it had, then conventional guards, physical keys and the all-pervading passwords and PIN numbers essential to gain access to mobile phones, computer software, buildings and so on, all would be in decline. This lack of take-up is even more surprising given the rate at which passwords and PINs are forgotten, as indicated by the prevalence of the web-page rescue message "Forgotten your password?" In short, why has biometric technology not made PINs, passwords and many physical keys things of the past? The answer must lie in factors such as cost, accuracy and convenience of use.

A recent report entitled "Biometric Product Testing" by Mansfield *et al* [1] presents what is believed to be the first major comparison of biometric systems based on different modes (fingerprints, faces, hand geometries, iris images, vein patterns and voice signals). It is interesting that voice performs relatively well, especially as voice also exhibits the benefits of low-cost and convenience. A good review of the commercial aspects of voice-based biometrics is provided by Boves [2]. In comparison to voice, other more sophisticated and potentially more accurate modes would seem to have signal capture difficulties in real application environments.

Furthermore, these potentially accurate biometrics, such as iris images and finger-prints, often possess negative attributes, such as high costs, sensitive sensors and intrusiveness to the user. Other biometrics can be cheaper, less intrusive, more covert and yet are still not widely adopted. A good example of this is the particular form of automatic speaker verification proposed by Auckenthaler *et al* [3]. This system continuously monitors the speech signal entering a mobile phone. After automatically learning the characteristics of the first person to use the device, it then continuously tracks usage, updating models as appropriate, and raises an alarm if usage does not include the said person for an excessive time. This is a good example of a covert, wholly non-intrusive biometric. Furthermore it would

0-7803-7402-9/02/$17.00 ©2002 IEEE

seem obvious that, with the advent of video-phones, visual speech could be added as a complementary biometric, thereby improving reliability.

It is this latter idea of visual speech as a biometric, combined with the more conventional audio-based speaker recognition, that is considered here. It is argued that visual speech is a particularly interesting case of a physiological and a behavioural biometric. In 1998 Roethenbaugh and Mansfield revised a glossary of biometric terminology originally compiled by the Association of Biometrics (AfB) in 1993. Prominent was the idea that systems can be thought of as either physiological or behavioural. Put simply, a physiological biometric is what you *are*, while a behavioural biometric is what you *do*. Thus a finger-print would be deemed a physiological biometric while a person's gait or voice would be a behavioural biometric. A potential difficulty with these terms or concepts stems from the unavoidable link between the two classes. It is fairly obvious that behavioural biometrics imply movement or dynamics, but these dynamics must always be dependent on the physiological make-up of the said person. One's gait must be dependent on the physical properties of one's legs! Some form of dependency will always hold for *all* behavioural biometrics.

Here a different emphasis is placed on the distinguishing characteristics between the two classes, namely the role played by dynamics. Table 2 summarises the implications of dynamics in physiological and behavioural biometrics. Consider first the row entitled 'dynamics'. In the case of a physiological biometric, movement might well be present but the key point is that the biometric signal is not itself a function of time. Furthermore, dynamics during the capture of the biometric signal might well be detrimental. Note, this is not to say that if dynamics are present the biometric is necessarily behavioural although the converse is true: without dynamics the biometric can only be physiological.

Consider next the 'signature variation' row in Table 2. The very existence of dynamics on whatever time scale causes unavoidable variations in the signatures. It is not possible to reproduce exactly the same speech signal, or exactly the same hand-written signature. Thus for physiological biometrics, signature variations can be very slow, small or even nil. Again this is in contrast to the behavioural case where variations across one person's signatures are inevitable. This is an underlying reason why behavioural biometrics are often labelled as less accurate. In practice it is difficult if not impossible to capture all the natural variations exhibited in a behavioural biometric.

An important practical consequence of this observation is shown in the final row of Table 2. As a direct consequence of the variations inherent in behavioural biometrics, it is necessary to acquire larger quantities of training data to represent the person. In order to acquire this data it is usually necessary to have multiple enrolment sessions, or preferably an adaptive system that updates itself through usage [3], thereby minimising user inconvenience. In speaker recognition for example, Furui [4] suggests that speech data should be collected over an interval of at least three months to capture typical variations. Any period of time shorter than this tends not to capture sufficient range of variations due to external influences such as ill health and mood fluctuations.

3. VISUAL SPEECH

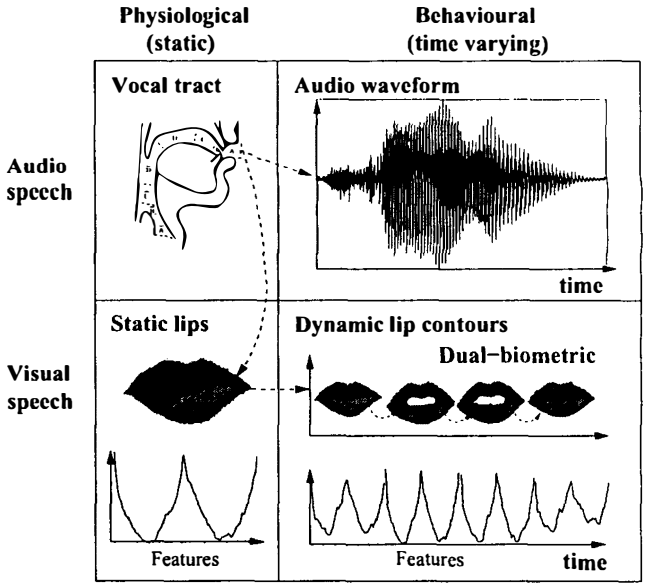

Fig. 1. Illustrating the inextricable link between behavioural biometrics and physiological components for audio and visual speech.

A key point in the classification of any biometric system lies in the biometric signal itself. It is argued that to be a behavioural biometric the biometric signal must be time dependent. Speech is a good example. Speech is unquestionably a behavioural biometric, albeit influenced by the physiological characteristics of

	Physiological	Behavioural
	Has	*Does*
Dynamics	Possible / Maybe detrimental	Essential
Signature variation	Possible / Slow / Small / Essentially nil	Inherent / Unavoidable
System enrollment/training	Possible to be one-off	Multi-session / Adaptive

Table 1. The implications of dynamics on physiological and behavioural biometrics.

the vocal tract. Of particular interest in this paper is visual speech in the form of lips and lip signatures. It is postulated that these signatures provide are a special case of a *twin* biometric, potentially both physiological and behavioural.

Figure 1 shows the components of visual speech along with an illustration of the speech production system. The acoustic signal, while obviously dependent on the vocal tract characteristics, is a clear example of a behavioural biometric: the person-specific information is born in a time-dependent signal. However, if the case for speech as a behavioural biometric is clear, then the case for the lips is far less clear. In the past, geometric lip profiles have been investigated for person recognition by a few researchers, including [5, 6, 7] while the majority of work using lip profiles has been for speech reading [8, 9, 10, 11].

Examples of lip profiles are shown in Figure 3. Measurements of lip profiles relate directly to the physical characteristics of the given person. A single profile is inherently stationary or instantaneous, without any time axis. Thus using one of these profiles in assessing a person's identity must lead to a physiological biometric. Repeating this process many times, across a series of profiles such as that in Figure 3 does not change the situation, it simply might increase the accuracy of the decision. However, if that sequence is recorded under prescribed conditions, such as during speech, then the sequence might itself be analysed along the time course. Movement and the underlying biometric behaviour can then be represented for a time series of *instantaneous* snap-shots by appropriate differentiation, resulting in dynamic features.[1] Using these changes across a time-set of these profiles in the form of dynamic features leads to a behavioural biometric. It is for this reason that such lip-based profiles are a special case, exhibiting the ready potential to be twin biometric, either behavioural or physiological.

Lip profile based recognition experiments highlight this interpretation. The database comprises video recordings of 9 persons who each utter a series of prompted digits. In total each person utters 144 digits. In testing, an identification decision is made on each single digit and the error rates are averaged across the 9 speakers, 12 digits. Thus the experimental results come from a total of 1080 digit tests. Visual features are the lip profiles illustrated in Figures 1 and 2. followed by a discrete cosine transform (DCT). See [7, 12] for more details and a comparison with the corresponding audio features.

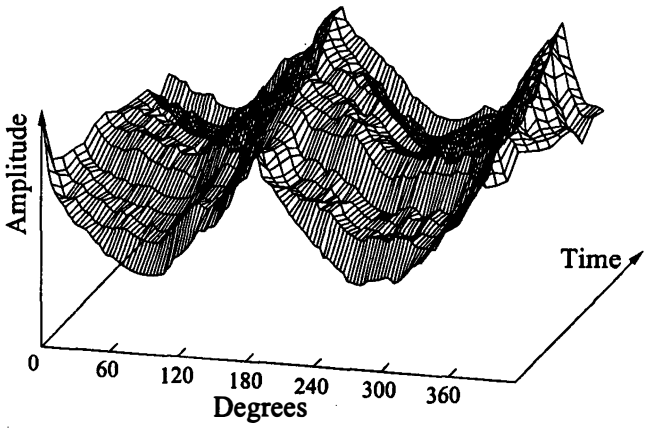

Fig. 2. A time series of lip profiles across one digit utterance.

Feature	Instantaneous	Dynamic
DCT of lip profiles	2%	15%

Table 2. Speaker identification error rates for visual speech in the form of discrete cosine transfoms of lip profiles: instantaneous are a physiological biometric, dynamics are behavioural.

1. The situation in the case of the audio speech is different at least in one important respect, namely the equivalent *instantaneous* snap-shots are not quite so instantaneous! In fact they come from a time window of speech recorded typically over 20ms or 30ms. Movement in the audio signal is inherent, reflected by the movement of the microphone.

Table 3 shows speaker recognition error rates using instantaneous and dynamic features. It can be seen that the instantaneous features achieve better recognition error rates than dynamic features. This is because of the inherent noise in the differentiating process to obtain the dynamic features.

4. CONCLUSION

This paper addresses biometric classification. It is argued that the simple classification scheme of behavioural and physiological can be difficult to apply in cases like visual speech where there is potential for both physical and behavioural traits to be employed. It is observed that while dynamics are essential for a behavioural biometric to exist, it does not follow that the presence of dynamics means that a biometric is behavioural. A key deciding factor is the nature of the information-bearing signal itself: if this is a function of time then the biometric is behavioural, otherwise the biometric is physiological.

Visual speech illustrates this definition particularly well and lip profiles provide what is thought to be a rare, if not unique, case of a practical twin biometric: one capable of being either behavioural or physiological.

Experimental results using lip profile sequences during spoken digits give recognition error rates as: physiological - instantaneous lip profiles 2% and behavioural - dynamic lip profiles 15%. Signatures from both classes of biometrics, physiological and behavioural, possess inherent variations from one occasion to the next. This is due to the fact that humans are unable to repeat a given physical action in *exactly* the same way: for example reproduce a spoken utterances with exactly the same time waveform. When the biometric is of the behavioural class, this natural variation is embedded into the biometric signature itself, thereby introducing dynamic or behavioural variations across signatures. As a consequence of these natural variations and in order to meet given performance specifications behavioural biometrics tend to require more training data than biometric systems of the physiological class.

5. REFERENCES

[1] T. Mansfield, G. Kelly, D. Chandler, and Y. Kane. Biometric Product Testing Final Report. *NPL intenal report*, 2001.

[2] L. Boves. Commercial Applications of Speaker Verification: Overview and Critical Success Factors. *RLA2C*, pages 150–159, 1998.

[3] R. Auckenthaler, E. Parris, and M. Carey. Improving a GMM Speaker Verification System by Phonetic Weighting. *ICASSP*, page 1440, 1999.

[4] S. Furui. Speaker-Independent Isolated Word Recognition using Dynamic Features of speech spectrum. *IEEE Trans. on ASSP*, 34:52–59, 1986.

[5] C. C. Chibelushi, J. S. Mason, and F. Deravi. Integration of acoustic and visual speech for speaker recognition. In *Proc. Eurospeech*, volume 1, pages 157–160, Berlin, 1993.

[6] P.Jourlin, J.Luettin, D.Genoud, and H.Wassner. Acoustic Labial Speaker Verification. *Proc AVBPA, Lecture Notes in Computer Science 1206*, pages 319–334, 1997.

[7] J. Brand, R. Auckenthaler, J. S. D. Mason, C. Chibelushi, and F. Deravi. Lip Signatures for Speaker Recognition. *AVBPA*, pages 142–147, 1999.

[8] A. Rogozan and P. Deleglise. Continuous Visual Speech Recognition Using Geometric Lip-Shape Models and Neural Networks. *EuroSpeech*, page 1999, 1997.

[9] I. Matthews, J. Bangham, R. Harvey, and S. Cox. Nonlinear Scale Decomposition Based Features for Visual Speech Recognition. *EUSIPCO98*, pages 303 – 305, 1998.

[10] I. Matthews, J. Bangham, R. Harvey, and S. Cox. A Comparison of Active Shape Model and Scale Decomposition Based Features for Visual Speech Recognition. *ECCV*, 1998.

[11] R. Gocke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. Automatic Extraction of Lip Feature Points. In *Australian Conference on Robotics and Automation*, pages 31–36, 2000.

[12] J. Brand, R. Auckenthaler, J. S. D. Mason, C. Chibelushi, and F. Deravi. Lip Signatures for Automatic Person Recognition. In *IEEE Workshop, MMSP*, pages 457–462, 1999.