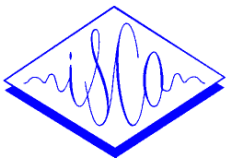

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

INTEGRATION OF ACOUSTIC AND VISUAL SPEECH FOR SPEAKER RECOGNITION

C.C. Chibelushi, J.S. Mason, F. Deravi

Department of Electrical and Electronic Engineering
University of Wales, Swansea, U.K.

ABSTRACT

This paper describes a speaker recognition system that uses both acoustic speech and visual speech (motion of visible articulators). Integration of acoustic and visual speech aims at improving recognition performance with regard to recognition accuracy, robustness against variability of input data, and protection against impersonation. As an initial step towards this goal, voice has been used together with still face images; this combination of vocal and facial information has resulted in better recognition accuracy than from either of the two constituents individually.

Keywords: *speaker recognition, face recognition, data fusion, visual speech*

10.21437/Eurospeech.1993-62

1 INTRODUCTION

There is a wide array of approaches to automatic person recognition; examples include the use of voice, face, finger–print, vein patterns, handwritten signature, and keyboard typing–rhythm. Large–scale adoption of such systems for real–world applications is being impeded by combinations of reasons such as their failure to operate with consistently high recognition performance (particularly for large user populations) and by their relatively poor performance in unconstrained environments, or by a sheer lack of acceptance by their intended users.

With the aim of improving speaker recognition performance, we propose a multi–sensor data fusion approach whereby acoustic and visual identifying characteristics are used jointly. Joint use of acoustic and visual information for person recognition has the potential advantage of higher accuracy than recognition in the acoustic or visual domain alone. In addition, performance is expected to improve in terms of robustness against variations of either acoustic or visual information, and protection against impersonation.

Supportive illustrations of the benefits of data fusion are given by Waltz [1], and Nahin and Pokoski [2]. Waltz performed a Monte Carlo simulation of a two–sensor data fusion system for tactical air–to–air IFFN (Identification, Friend, Foe, Neutral). He showed that the identification accuracy of the sensor fusion system was better than that of separate single–sensor systems. Nahin and Pokoski used a mathematical model to show that improvements in identification accuracy result from sensor fusion. Multi–sensor identity fusion may be performed at several levels: data level, feature level, or decision level; hybrid fusion methods are also possible [1] [3].

Acoustic speech and visual speech (i.e. motion of visible articulators) have been used jointly for speech recognition by a number of researchers, for example Bregler and co–workers [4]. However, to our knowledge, no automatic speaker recognition system has used the speaker identifying cues of visual speech. This paper presents our approach regarding the use of identity–bearing characteristics carried by lip motion during an utterance. We describe a speaker recognition method which uses both vocal and facial personal characteristics.

2 EXTRACTION OF VOCAL AND VISUAL FEATURES

2.1 Vocal features

Perceptual Linear Predictive (PLP) cepstral features are chosen to represent the vocal personal characteristics extracted along the time course of each utterance [5].

2.2 Visual features

The preprocessing of visual speech includes normalisation against affine image transformations arising from changes in head orientation and position, and changes in the distance between the camera and the talker. Location and tracking of the lips is based on deformable contour models [6] and motion segmentation. Motion segmentation is the process, based on image motion, of extracting regions of structural significance in a scene. The contour models use image edge information and *a priori* knowledge of lip shape to yield position information in image coordinates.

For recognition, distinctive personal features are extracted from either static or, alternatively, dynamic parameters of the lip contours. For better temporal feature stability and protection against mimics, the extracted identifying characteristics should be based on the biological individuality of the structures involved in visual speech production.

In their study of lipreading, Montgomery et al. [7] observed that the size and shape of inner lip–border tracings varied across talkers; they also noted much less variation for the same talker. Based on these findings, one of our feature sets consists of static features extracted from lip margins.

The velocity of muscular contraction depends on the magnitude of the load being displaced by the muscle [8]. Given that the muscular mass around the mouth and its underlying skeletal structure are presumed unique to each individual, we hypothesize that the motion of the lips during speech production also carries information about personal identity. Hence, dynamic features extracted from lip motion parameters are considered for use as identifying features.

Two dynamic feature extraction models are proposed: an excitation–modulation model, and a lumped–parameter kinetic model.

(a) Excitation–modulation model

Here, lip motion is considered to result from the response of a bio–mechanical system to excitation from motor neurons. Assuming that the excitation and the transfer function are separable in the cepstral domain, homomorphic filtering can be used to separate the transfer function of the bio–mechanical system from the excitation component of the observed motion. We postulate that the identity–carrying characteristic are embedded in the transfer function.

(b) Lumped–parameter kinetic model

A kinetic model of lip motion is used for extracting the identifying characteristics of the changes in lip shape arising from time–dependent non–rigid elastic deformations. The model consists of concentrated masses (point masses) distributed around the lip margins; the masses are connected to massless springs and dampers. Each mass is subjected to disturbing forces as a result of muscular action. Restoring forces in the springs try to return the point mass to its rest position, and velocity–dependent damping forces in the dampers reduce the oscillations. The resultant of all the forces acting on a mass gives rise to motion in accordance with Newton's second law of motion. A kinetic description of the translational motion of each point mass is of the form:

$$m_i \frac{d^2\bar{p}(t)}{dt^2} + \gamma_i \frac{d\bar{p}(t)}{dt} + \kappa_i \bar{p}(t) = \bar{F}_d(t)$$

where:

	m_i	i^{th} point mass,
	γ_i	viscous damping coefficient,
	κ_i	spring stiffness coefficient,
	t	time,
	$\bar{p}$	displacement vector of the point mass from its rest position,
	$\bar{F}_d$	disturbing force.

Feature vectors comprise the parameters m_i, γ_i, κ_i, and $\bar{F}_d$.

3 CLASSIFICATION

Each known person is allocated an artificial neural network model. Multi-layer perceptrons (MLPs) trained as predictive or discriminative networks (Figure 1) are considered.

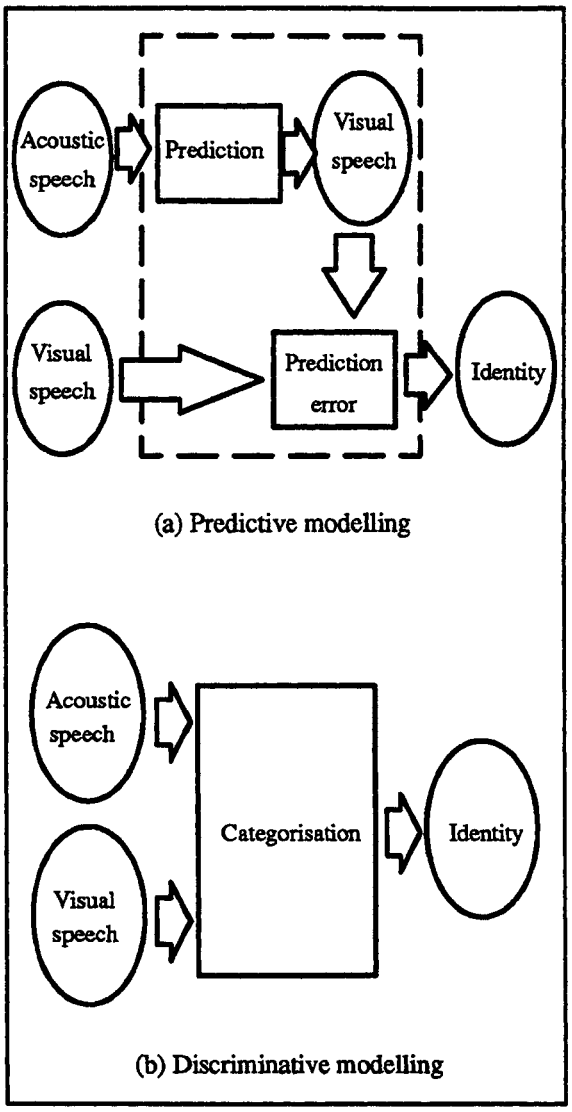

Figure 1: Predictive and discriminative classifier modes

For speedier and incremental training, fuzzy ARTMAP models are preferred. Each fuzzy ARTMAP model is trained to make a many-to-one mapping from acoustic speech to visual speech for its allotted person (Figure 2). The predictive error across an utterance acts as recognition measure.

A fuzzy ARTMAP neural network architecture is composed of a pair of fuzzy Adaptive Resonance Theory modules (modules A and B in Figure 2) [9]. Each module creates recognition categories corresponding to its input data. The two modules are interconnected by a map field which makes predictive associations between the categories of modules A and B. The map field tries to redress any predictive mismatch by reorganising the category structure through a strategy known as match tracking.

Owing to the critical importance of the fusion method used, we evaluate various fusion schemes. In the work described herein, voice and image sensors are neither identical nor commensurate, hence data level fusion is precluded. Principal component analysis or linear discriminant analysis is used for feature level fusion. In addition, the following decision-level identity fusion schemes are considered: Bayesian inference, Dempster-Shafer theory, possibility theory (based on fuzzy set theory), and artificial neural networks.

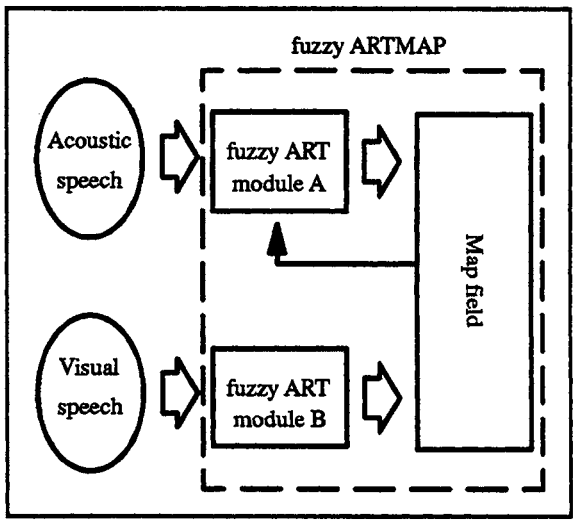

Figure 2: Fuzzy ARTMAP speaker model

4 RESULTS

Results based on face profile image and voice data fusion show that the combination of acoustic and visual identifying cues yields improvements in recognition accuracy. Identification error rates reduce from 6.25 % for voice recognition and 12.5 % for face recognition to 1.25 % for an integrated voice-face recognition system. Similarly, an improvement in verification equal-error rates is observed (voice recognition: 3.4 %, face recognition: 3.0 %, integrated system: 1.5 %). These improvements are achieved with a simple linear combination scheme [10].

Anticipated results for our investigations on acoustic and visual speech integration are:

- an improvement in recognition performance resulting from data fusion for normal input data and for a range of degraded input data conditions.
- discriminative models will show better performance than predictive models. This will result from the more stringent data alignment (lip sync) requirements and the lack of classifier input information redundancy in the predictive scheme. Also, predictive modelling is expected to show less robustness to input data variability.
- fuzzy ARTMAP models are expected to outperform MLP predictive models with regard to accuracy. This is because the fuzzy ARTMAP map field realises a minimax learning rule that conjointly allows the minimisation of predictive error and the maximisation of predictive generalisation [9]. Standard back-propagation learning does not provide such a mechanism. In addition, tremendous savings in training times are expected with the fuzzy ARTMAP approach.

4 CONCLUSIONS

Humans often rely on multiple senses – particularly hearing and vision – for many recognition tasks. We have proposed the joint use of acoustic and visual information for reliable automatic speaker recognition. The initial step towards this goal has combined static facial image information with voice, and the results have shown that performance improvements can be achieved even with a relatively simple integration scheme. We believe that visual speech does convey personal identity information, and that its use in conjunction with acoustic speech results in improved automatic speaker recognition performance in terms of accuracy, robustness, and protection against impersonation.

ACKNOWLEDGEMENT

Sincere thanks are expressed to the Beit Trust for the financial support in form of a Beit Trust Fellowship awarded to the first author.

REFERENCES

1 Hall, D.L., **Mathematical Techniques in Multisensor Data Fusion**, pp. 23 – 25, Artech House Inc., 1992

2 Nahin, P.J., Pokoski, J.L., NCTR Plus Sensor Fusion Equals IFFN or Can Two Plus Two Equal Five?, **IEEE Transactions on Aerospace and Electronic Systems**, Vol. AES–16, No. 3, pp. 320–337, 1980

3 Zhang, X., Mason, J.S., Andrews, E.C., Multiple Dynamic Features to Enhance Neural Net Based Speaker Verification, **Eurospeech 91**, Vol. 3, pp. 1411 – 1414, 1991

4 Bregler, C., Hild, H., Manke, S., Waibel, A., Improved Connected Letter Recognition by Lipreading, **ICASSP 93**, Vol. I, pp. I–557 – I–560, 1993

5 Xu, L., Oglesby, J., Mason, J.S., The Optimization of Perceptually–based Features for Speaker Identification, **ICASSP 89**, Vol. I, pp. 520 – 523, 1989

6 Waters, K., Terzopoulos, D., The Computer Synthesis of Expressive Faces, **Philosophical Transactions of the Royal Society – Series B: Biological Sciences**, Vol. 335, No. 1273, pp. 87 – 93, 1992

7 Montgomery, A.A., Jackson, P.L., Physical Characteristics of the Lips Underlying Vowel Lipreading Performance, **Journal of the Acoustical Society of America**, Vol. 73, No. 6, pp. 2134 – 2144, 1983

8 Luciano, D.S., Vander, A.J., Sherman, J.H., **Human Function and Structure**, McGraw–Hill, 1978

9 Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., Rosen, D.B., Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps, **IEEE Transactions on Neural Networks**, Vol. 3, No. 5, pp. 698 – 713, 1992

10 Chibelushi, C. C., Deravi, F., Mason, J.S., Voice and Facial Image Integration for Person Recognition, **IEEE International Symposium on Multimedia Technologies and Future Applications**, April 1993