IMPROVED SPEECH READING THROUGH A FREE-PARTS REPRESENTATION

Simon Lucey, Patrick Lucey+*

*Advanced Multimedia Processing Laboratory
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh PA 15213, USA
+Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2424, Brisbane 4001, Australia
*slucey@ieee.org, +p.lucey@qut.edu.au

ABSTRACT

Motivated by the success of free-parts based representations in face recognition [1] we have attempted to address some of the problems associated with applying such a philosophy to the task of speaker-independent automatic speech reading. Hitherto, a major problem with canonical area-based approaches in automatic speech reading is the intrinsic lack of training observations due to the visual speech modality's low sample rate and large variability in appearance. We believe a free-parts representation can overcome many of these limitations due to its natural ability to generalize by producing many observations from a single mouth image, whilst still preserving the ability to discriminate between various visual-speech units. This approach additionally requires a modification to traditional techniques employed for the estimation of hidden Markov Models (HMMs), whose resultant models we currently refer to as free-parts HMMs (FP-HMMs). Results will be presented on the CUAVE audio-visual speech database.

1. INTRODUCTION

A problem of immense importance across the broad gamut of pattern recognition tasks at the moment is the ability to produce robust and well trained classifiers from sparse amounts of training observations. The task of speaker independent automatic speech reading is a prime candidate for the development and analysis of this generic pattern recognition problem, as the nature of the task demands an ability to generalize for an infinite number of inter-speaker permutations from a finite and typically small visual-speech database.

It is largely agreed upon that the majority of visual speech information stems from a subject's mouth [2]. As a result a large proportion of the work that has been conducted in automatic speech reading has been towards the goal of finding a suitable mouth representation for recognition purposes. The more discriminant and compact the mouth representation, generally the easier the recognition task. From literature [3] mouth features can be categorized into two types, namely: contour and area based representations. Area based representations are concerned with transforming the whole region of interest (ROI) mouth pixel intensity image into a meaningful feature vector(s). Contour based representations, like those proposed by Luettin et al. [4], Matthews et al. [5] or Wark and Sridharan [6] are concerned with parametrically atomizing the mouth, based on a priori knowledge of the components of the mouth (i.e. outer and inner labial contour, tongue, teeth, etc.). In a recent paper by Potamianos et al. [7] a review was conducted between area and contour features for the tasks of speechreading (i.e. speech recognition using only the visual modality) on a large AV database. In this paper it was shown that area representations obtained superior performance. However, it has also been noted [8] that the visual speech modality is inherently under-trained using area-based features; limiting the performance of current automatic speech reading algorithms. This limitation can be related to the intrinsic lack of training observations being used to describe the visual speech modality; due to the modality's low sample rate and large variability in pixel appearance.

In this paper we propose a novel representation of the mouth, which we refer to as *free-parts*, based on recent success this representation has enjoyed in the task of face recognition [1]. Much improvement has been noted in speech recognition literature [9] by relaxing temporal structure in the speech signal (e.g. Gaussian mixture models (GMMs), Hidden Markov Models (HMMs)) when compared to more temporally rigid models (e.g. Dynamic time warping (DTW)). An advantage of dealing with models based on low-dimensional distributions (e.g. GMMs, HMMs) over models based on single points existing in a high-dimensional space (e.g. DTWs)

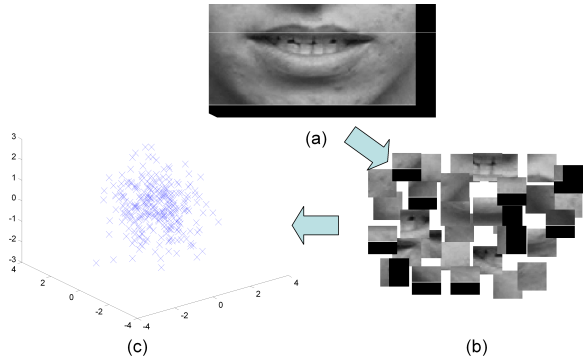

Fig. 1. Depiction of the process of structural collapse of the mouth, from an (a) single monolithic mouth image, through the (b) structural relaxation process (i.e. removing positional information); to finally be left with (c) a cloud of free-parts observations describing the subject's mouth.

is their inherent ability to generalize. This generalization stems from the increase in the number of training observations and a decrease in the dimensionality of these observations; both being caused from the relaxation of structural constraints in the signal. Extending this idea to the spatial domain, we propose the employment of a free-parts representation that relaxes both spatial and temporal structure in the visual signal. Free-parts representations assume that the position/structure of patches within the mouth image can be relaxed so they can "freely" move to varying extents. An example of this structural relaxation can be seen in Figure 1. A caveat however, must be placed on the removal of structure for improved generalization; as the removal of structure will always come at the cost of discrimination in the speech reading task. From this perspective, the job of the engineer or scientist in designing an automatic speech reading system is to find the appropriate balance of generalization vs discrimination for the task at hand.

In this paper we propose two novel concepts to improve current automatic speech reading performance. First, the introduction of the free-parts representation to alleviate current limitations caused by under-trained visual HMMs. Second, the formulation of a modified form of the Viterbi and EM algorithms to enable the estimation of FP-HMMs. This second contribution, comes as a consequence of certain inherent characteristics of a free-parts representation of the mouth, namely: (i) That multiple observations will occur at the exact same time instant. (ii) A state transition within a FP-HMM can only occur when *all* observations for that mouth image have been evaluated (i.e. a state transition cannot occur given that only half the observations within a mouth image have been evaluated). We will be presenting results on the CUAVE [10] audio-visual speech database for the task of speaker-independent speech reading.

2. REFERENCES

[1] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[2] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard hearing people," *IEEE Trans. Rehabilitation Engineering*, vol. 3, no. 1, pp. 90–102, March 1995.

[3] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing,*. MIT Press, 2004.

[4] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," in *International Conference on Spoken Language Processing*, 1996, vol. 1, pp. 58–61.

[5] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham, "Lipreading using shape, shading and scale," in *Auditory-Visual Speech Processing*, Sydney, Australia, 1998, pp. 73–78.

[6] T. Wark and S. Sridharan, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *International Conference on Pattern Recognition*, 1998, vol. 1, pp. 123–125.

[7] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *International Conference on Image Processing*, 1998, vol. 3, pp. 173–177.

[8] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in *International Conference of Audio- and Video-Based Person Authentication (AVBPA)*, Guildford, U.K., 2003, pp. 260–267.

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[10] Z. Tufekci E. K Patterson, S. Gurbuz and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal humancomputer interface research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.