Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 21 (2006) 1–12
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Partial linear regression for speech-driven talking head application

Chao-Kuei Hsieh, Yung-Chang Chen*

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan 30013, ROC

Received 15 September 2004; accepted 25 April 2005

Abstract

Avatars in many applications are constructed manually or by a single speech-driven model which needs a lot of training data and long training time. It is essential to build up a user-dependent model more efficiently. In this paper, a new adaptation method, called the partial linear regression (PLR), is proposed and adopted in an audio-driven talking head application. This method allows users to adapt the partial parameters from the available adaptive data while keeping the others unchanged. In our experiments, the PLR algorithm can retrench the hours of time spent on retraining a new user-dependent model, and adjust the user-independent model to a more personalized one. The animated results with adapted models are 36% closer to the user-dependent model than using the pre-trained user-independent model.

Keywords: Partial linear regression (PLR); Speaker adaptation; Speech-driven face animation; Audio-to-visual conversion

1. Introduction

With the rapid development of multimedia technology, the virtual avatar has been widely used in many areas, like cartoon or computer game characters and news announcers. However, huge amount of manpower is needed in adjusting the avatar frame by frame to achieve a vivid and precise synthetic facial animation, since the asynchronism between mouth motion and voice pronunciation would be a fatal defect of realism. Therefore, a real-time speech-driven synthetic talking head, or so-called audio-to-visual synthesis system, is expected, which can provide an effective interface for many applications, e.g. image communication [1,24], video conferencing [12,7], video processing [8], talking head representation of agents [26], and telephone conversion for people with impaired hearing [22].

*Corresponding author. Tel.: +886 3 5731153; fax: +886 3 5715971.
E-mail addresses: frost@benz.ee.nthu.edu.tw (C.-K. Hsieh), ycchen@ee.nthu.edu.tw (Y.-C. Chen).

0923-5965/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2005.04.002

In an audio-to-visual synthesis system, it needs a model established for describing the correspondence between the acoustic parameters and the mouth-shape parameters. In other words, the corresponding visual information is to be estimated for some given acoustic parameters, such as the phonemes, the cepstral coefficients or the line spectrum pairs. The visual information could be images or mouth movement parameters. Mouth images were used in the work of Bregler et al. [6] to provide a factual representation. However, the stitching perplexity and the limited view angle abated the practicability.

A number of algorithms have been proposed for the task of mapping between acoustic parameters and visual parameters. The conversion problem is treated as one of finding the best approximation from given sets of training data. These approaches were briefly discussed in Chen and Rao [10], including vector quantization [25], Hidden Markov Models (HMM) [2,3,9,13,31], and neural networks [19,20,30]. However, the speech-driven systems were generally made to be user-independent for satisfactory average performance, which means a decrease in accuracy rate for a specific user. To maintain high performance, a time-consuming retraining procedure for a new user-dependent model is unavoidable since there is no reported adaptation method for this application in the literature.

On the other hand, speaker adaptation methods have been extensively studied in the speech recognition field. There are two main categories in the adaptation methods. The first is the eigenvector-based speaker adaptation method [4,5], which uses the normalization on both the training-end and the recognition-end to deal with a variety of the acoustic characteristics due to different vocal channels. The other is based on the acoustic model, and is simpler than the former since the normalization for the training data is not necessary. A user-independent model is statistically established with the training data of several speakers in the beginning, and the parameters are then modified with certain adaptation data of a new user. The adaptation schemes include maximum a posteriori (MAP) estimation [11,17,27,28], maximum likelihood linear regression (MLLR) [18,21,32], VFS [29], and nonlinear neural network [16]. In these methods, they tried to adjust the model parameters to maximize the occurrence probability of the new observation data. Among them, the MLLR method is more widely adopted for its simplicity and effectiveness when the set of adaptation data is small.

In this study, we try to integrate the MLLR adaptation approach with the audio-to-visual conversion of Gaussian mixture model, because the MLLR is first used for speaker adaptation of continuous density Hidden Markov Models and GMM is the kernel distribution used in an HMM. If the adaptation of audio-to-visual conversion model can be carried out with both audio and visual adaptation data, it will be exactly the same task as that in [21]. However, to obtain the precise visual adaptation information of a new user is not feasible in a usual environment, since some markers, infrared cameras, and post-processing (same as in the training phase) are needed. This makes the MLLR not fully adequate to adapt only the audio parameters while keeping the visual part the same. In other words, we require another appropriate adaptation, by means of which the new model will map the new audio parameters of a new user to the original visual movement.

A new adaptation method, called partial linear regression, is proposed in this paper. It is derived from the MLLR and put into practice in an audio-driven talking head system (Fig. 1). Rather than a time consuming retraining procedure, a simple adaptation with a small amount of additional data will be sufficient to adjust the model so as to be more applicable to the new user.

The rest of the paper is organized as follows. In Section 2, we describe the audio-driven talking head system which uses the Gaussian mixture model to represent the relationship between audio and video feature vectors. The audio-to-visual conversion is also mentioned. Section 3 provides a review of MLLR and a detailed description of the proposed PLR model adaptation algorithm. Some experimental results are described in Section 4, and Section 5 concludes the paper.