Pattern Recognition Letters 32 (2011) 780–786

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Correlation based speech-video synchronization

Amar A. EL-Sallam [a,b,*], Ajmal S. Mian [b]

[a] School of Electrical, Electronic and Computer Engineering, The University of Western Australia, 35 Stirling Highway Crawley, WA 6009, Australia
[b] School of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 28 August 2009
Available online 9 January 2011
Communicated by H.H.S. Ip

Keywords:
Correlation
Lip sync
Formants
Estimation
AM, FM

ABSTRACT

This paper presents a novel Lip synchronization technique which investigates the correlation between the speech and lips movements. First, the speech signal is represented as a nonlinear time-varying model which involves a sum of AM–FM signals. Each of these signals is employed to model a single Formant frequency. The model is realized using Taylor series expansion in a way which provides the relationship between the lip shape (width and height) w.r.t. the speech amplitude and instantaneous frequency. Using lips width and height, a semi-speech signal is generated and correlated with the original speech signal over a span of delays then the delay between the speech and the video is estimated. Using real and noisy data from the VidTimit and in-house diastases, the proposed method was able to estimate small delays of 0.01–0.1 s in the case of noise-less and noisy signals respectively with a maximum absolute error of 0.0022 s.

1. Introduction

Lip Synchronization (Lip-Sync) is one of the important topics in speech and video analysis. It has wide applications in multimedia including; the film industry, synchronization of visual and audio signals during post production and transmission, animated characters (including computer facial animation), Automated Dialogue Replacement (ADR), lip dubbing, to name a few (ATSC, 2003; Li et al., 2003). Recently Lip-Sync became a serious problem for the multimedia industry. Lip sync problems can become annoying and lead to subconscious viewer stress which in turn leads to viewer dissatisfaction with the program or the multimedia service they are using. Accordingly, the television industry standards organizations have become involved in setting standards for lip sync errors (ATSC, 2003; David, 1993). Speech-video synchronization is also used as a pre-processing step to strengthen the performance of audio–video based speaker verification (e.g. in banking). Alternatively, the output of the speech-video synchronization process can itself be used to verify the speaker or the liveliness of the biometrics to avoid replay attacks (Garcia et al., 2004) or for audio visual automatic speech recognition systems (AVASR) (Rajitha et al., 2010). A very good servery can be found in (You et al., 2009). For Lip sync; data belong to the lips' movements and the associated speech are needed. This requires the shape of the lips to be tracked and from which data such as the shape or curvature can be extracted. Our main focus in this work is to derive a novel theoretical analysis for the relationship between the lip shape and the associated speech. To enable the reader to follow the proposed method, the image processing part which involve lips tracking is outside the scope of this work and several off-the-shelf techniques and implementations (e.g. OpenCV). In (Rajitha et al., 2010) for example Viola–Jones algorithm was used, and in (Kazuhiro, 2002) a lip tracking algorithm based on contrast analysis have been developed, while in (Duy and David, 2006) a naive Bayes classifier and Lip feature extraction that uses the contrast around the lip contour are employed to extract the height and width of the mouth. In (Lewis, 1991) Lip Sync is investigated using a prediction LPC model to estimate the upcoming lips shape. In (Mori and Sonoda, 1996; Barbosa and Yehia, 2001), linear and nonlinear mapping is used to estimate facial motions of speech acoustics, and similarly in (Huang, 1998) an algorithm is proposed to map the speech into visual parameters of mouth movements. In (Ogata et al., 2001), a multi-modal system is proposed to synchronize lips movement with synthetic voices. Some methods also studied the process of synchronizing lips movements using a driven speech based on a pre-defined trajectory and speech synthesis (Koster et al., 1994; Chen et al., 1995). In (Chan-Ho et al., 2003) a synchronization method is employed which uses training pilot signals injected between audio and video signals. In (Fabio, 1997), a time-delay neural networks learning method is proposed for estimating lip movements from speech analysis, but requires a long training sequence and the extraction of several face features. In (Zoric and Pandzic, 2005), a neural network based method is also employed to classify the speech into viseme classes then map them into a virtual character's face.

* Corresponding author at: School of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway Crawley, WA 6009, Australia. Tel.: +61 8 6488 3695; fax: +61 8 6488 1089.

E-mail addresses: elsallam@csse.uwa.edu.au, elsallam@ieee.org, elsallam@ee.uwa.edu.au (A.A. EL-Sallam), ajmal@csse.uwa.edu.au (A.S. Mian).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.01.001

A limitation of this method is that there are many-to-one mappings between phonemes and viseme. Another recent lip synch approach that uses phonemes recognition and a Head–Body–Tail (HBT) model is developed in (Junho, 2008). The approach is developed to minimize the computation complexity of some existing methods but still requires heavy computation. In (Fedina and Glasman, 2006) an analytical method was developed for lip-sync by measuring correlation between real visemes of the speech phonemes and the predicated ones. Recently, Liu and Sato (2009), performed audio–video synchronization using kernel density estimation and quadratic mutual information. They extracted visual features using a fixed window around the lips assuming a stationary speaker.

We propose a correlation based technique for estimating the delay between speech and video signals. Our method is different since it does not use training-based audio–video signals or their sync pilots, neither it fit them into a pre-defined model or trajectory (mapping). Our method does not require the identification of the speech phonemes (Fedina and Glasman, 2006). For Lip-Sync, our method uses a sum of AM–FM signals to model the speech signal. The proposed model is analyzed using Taylor series in a way that shows the relationship between lips movements and the speech. A correlation function is then formulated to estimate the delay between the speech signal and signals generated from lips movements. The proposed approach was tested on the VidTimit database (Sanderson, 2008) and was able to estimate delays as small as 0.1 s with a maximum error of 0.0031 s.

2. System model

Without loss of generality, there are six main elements of a speech signal namely; (i) Formants, (ii) pitches, (iii) amplitude due to (nasals, approximants, lips, etc.), (iv) noise (fricatives, unvoiced), (v) transients (stop-release bursts), and (vi) others (moving, timing) Sundberg and Nordström (1976), Ellis (2009). Accordingly, let us consider the speech model,

$$x(t) = \sum_{k=1}^{M} a_k(t) \cos\left(2\pi f_k t + \int_{\lambda=0}^{t} b_k(\lambda) d\lambda + n_k^{FM}(t) + \phi_k\right) + n^{AM}(t) \tag{1}$$

This model represents a combined AM–FM signal that satisfies the main six elements mentioned above and additional characteristics provided for speech models studied by Gordon and Harold (1952), Alexandros and Petros (1999) and Dimitriadis et al. (2005), Christopher and Chris (2006), where k represents the Formant frequency index, f_k is the kth Formant frequency, M is the number of Formants, $a_k(t)$ is a function that controls the amplitude of the kth Formant. We define $a_k(t)$ by a 2D function,

$$a_k(t) = f_{a_k}(m_l(t), m_i(t)), \tag{2}$$

where $m_l(t)$ is a function associated to the AM signaling part due to lip dimensions (width and height) and $m_i(t)$ is another function representing the AM modulation due to tongue, jaw, larynx, etc. $b_k(t)$ is a function which controls the bandwidth of each Formant frequency. The frequency signal $b_k(t)$ is considered to represent the bandwidth variation of each Formant frequency. Similarly, we define $b_k(t)$ by another 2D function,

$$b_k(t) = f_{b_k}(m_l(t), m_i(t)) \tag{3}$$

where $m_l(t)$ and $m_i(t)$ are as described above, $n^{AM}(t)$ and $n_k^{FM}(t)$ are additives AM and FM noise, respectively and ϕ_k is a constant phase.

In our model, we use $f_{i_k} = f_k + b_k(t)$ to represents one of the instantaneous frequencies. Also, $m_l(t), m_i(t), n_k^{AM}(t), n_k^{FM}(t)$ will be assumed independent and with finite means and variances. The frequency of change in $a_k(t)$ and $b_k(t)$ is small when compared with the Formant frequencies. This assumption is true since for example the number of movements a person can make using his/her lips are very slow when compared with the lowest audible frequency.

An example using VidTimit data for a speech signal, its video frames and their corresponding horizontal lips movement (width) and vertical lips movement (height) is shown in Fig. 2.

The main objective is to synchronize a speech signal with its associated video signal represented by lips movements. To achieve this, we will first establish a mathematical relationship between the two signals.

3. Mathematical analysis

According to Taylor series representations, an L-dimensional function $f(x_1,\ldots,x_L)$ can be represented by,

$$f(x_1,\ldots,x_L) = \sum_{n_{1=0}}^{\infty} \cdots \sum_{n_{L=0}}^{\infty} \frac{\partial^{n_1}}{\partial x_1^{n_1}} \cdots \frac{\partial^{n_L}}{\partial x_L^{n_L}} \frac{f(x_1,\ldots,x_L)}{n_1!\ldots n_L!}\bigg|_{(x_1=x_1^o,\ldots,x_L=x_L^o)} (x_1 - x_1^o)^{n_1} \ldots (x - x_L^o)^{n_L}, \tag{4}$$

where $x_1^o,\ldots,x_L^o, f(x_1^o,\ldots,x_L^o)$ and the partial derivatives of $f(x_1,\ldots,x_L)$ w.r.t $x_1,\ldots,x_L$ at $x_1^o,\ldots,x_L^o$, are denoted by $\frac{\partial^{n_1}}{\partial x_1^{n_1}} \cdots \frac{\partial^{n_L}}{\partial x_L^{n_L}} \frac{f(x_1,\ldots,x_L)}{n_1!\ldots n_L!}\big|_{(x_1=x_1^o,\ldots,x_L=x_L^o)}$ and are all known initial conditions (values). By applying Taylor series to $f_{a_k}(m_l(t), m_i(t))$ and $f_{b_k}(m_l(t), m_i(t))$ we have,

$$f_{a_k}(m_l(t), m_i(t)) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \frac{\partial^{n_1}}{\partial m_l(t)^{n_1}} \frac{\partial^{n_2}}{\partial m_i(t)^{n_2}} \frac{f_{a_k}(m_l(t), m_i(t))}{n_1! n_2!}\bigg|_{(m_l(t)=m_l(t_o), m_i(t)=m_i(t_o))} (m_l(t) - m_l(t_o))^{n_1} (m_i(t) - m_i(t_o))^{n_2}, \tag{5}$$

where, $m_l(t_o)$, $m_i(t_o)$ and $\frac{\partial^{n_1}}{\partial m_l(t)^{n_1}} \frac{\partial^{n_2}}{\partial m_i(t)^{n_2}} \frac{f_{a_k}(m_l(t), m_i(t))}{n_1! n_2!}\big|_{(m_l(t)=m_l(t_o), m_i(t)=m_i(t_o))}$ are the initial conditions at the initial instant t_o. Using superposition at $n_2 = 0$, $n_1 = 0$ and n_1, $n_2 = 1, 2, \ldots, \infty$, respectively, Eq. (5) can be partitioned into,

$$\begin{aligned} f_{a_k}(m_l(t), m_i(t)) = & \sum_{n_1=0}^{\infty} \alpha_{n_{1k}} (m_l(t) - m_l(t_o))^{n_1} \\ & + \sum_{n_2=0}^{\infty} \zeta_{n_{2k}} (m_i(t) - m_i(t_o))^{n_2} \\ & + \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \eta_{n_1 n_{2k}} (m_l(t) - m_l(t_o))^{n_1} (m_i(t) - m_i(t_o))^{n_2} \end{aligned} \tag{6}$$

where,

$$\begin{aligned} \alpha_{n_{1k}} &= \frac{\partial^{n_1}}{\partial m_l(t)^{n_1}} \frac{f_{a_k}(m_l(t), m_i(t))}{n_1!}\bigg|_{(m_l(t)=m_l(t_o), m_i(t)=m_i(t_o))} \\ \zeta_{n_{2k}} &= \frac{\partial^{n_2}}{\partial m_i(t)^{n_2}} \frac{f_{a_k}(m_l(t), m_i(t))}{n_2!}\bigg|_{(m_l(t)=m_l(t_o), m_i(t)=m_i(t_o))} \\ \eta_{n_1 n_{2k}} &= \frac{\partial^{n_1}}{\partial m_l(t)^{n_1}} \frac{\partial^{n_2}}{\partial m_i(t)^{n_2}} \frac{f_{a_k}(m_l(t), m_i(t))}{n_1! n_2!}\bigg|_{(m_l(t)=m_l(t_o), m_i(t)=m_i(t_o))} \end{aligned} \tag{7}$$

The first partition of Eq. (6) (at $n_2 = 0$), corresponds to Taylor terms for $m_l(t)$, the second partition (at $n_1 = 0$), corresponds to Taylor terms for $m_i(t)$ and the third partition (at n_1, $n_2 = 1, 2, \ldots, \infty$) corresponds to Taylor mixed/cross-terms between $m_l(t)$ and $m_i(t)$. By expanding Eq. (6) at $n_1 = 0, 1, \ldots, \infty$, the above derivation and terminology can be translated to,

$$f_{a_k}(m_l(t), m_i(t)) = \alpha_{o_k} + \alpha_{1_k} m_l(t) + \alpha_{2_k} m_l(t)^2 + \cdots + c_k(t) \tag{8}$$

where

$$c_k(t) = \sum_{n_2=0}^{\infty} \zeta_{n_{2k}} (m_i(t) - m_i(t_o))^{n_2} + \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \eta_{n_1 n_{2k}} (m_l(t) - m_l(t_o))^{n_1} (m_i(t) - m_i(t_o))^{n_2} \tag{9}$$