Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 48 (2006) 598–615

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A comparison of acoustic coding models for speech-driven facial animation

Praveen Kakumanu [a], Anna Esposito [b], Oscar N. Garcia [c], Ricardo Gutierrez-Osuna [d,*]

[a] *Department of Computer Science and Engineering, 358 Russ Engineering Center, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435-0001, United States*

[b] *Department of Psychology at the Second University of Naples, Italy*

[c] *College of Engineering, University of North Texas, United States*

[d] *Department of Computer Science, Texas A&M University, College Station, TX 77843, United States*

Received 23 January 2004; received in revised form 21 July 2005; accepted 21 September 2005

Abstract

This article presents a thorough experimental comparison of several acoustic modeling techniques by their ability to capture information related to orofacial motion. These models include (1) Linear Predictive Coding and Linear Spectral Frequencies, which model the dynamics of the speech production system, (2) Mel Frequency Cepstral Coefficients and Perceptual Critical Feature Bands, which encode perceptual cues of speech, (3) spectral energy and fundamental frequency, which capture prosodic aspects, and (4) two hybrid methods that combine information from the previous models. We also consider a novel supervised procedure based on Fisher's Linear Discriminants to project acoustic information onto a low-dimensional subspace that best discriminates different orofacial configurations. Prediction of orofacial motion from speech acoustics is performed using a non-parametric *k*-nearest-neighbors procedure. The sensitivity of this audio–visual mapping to coarticulation effects and spatial locality is thoroughly investigated. Our results indicate that the hybrid use of articulatory, perceptual and prosodic features of speech, combined with a supervised dimensionality-reduction procedure, is able to outperform any individual acoustic model for speech-driven facial animation. These results are validated on the 450 sentences of the TIMIT compact dataset.

Keywords: Speech-driven facial animation; Audio–visual mapping; Linear discriminants analysis

1. Introduction

Lip reading plays a significant role in spoken language communication. It is not only essential for the hearing impaired but also used by normal listeners as an aid to improve the intelligibility of speech in noisy environments. Summerfield (1979) has shown

[*] Corresponding author. Tel.: +1 979 845 2942; fax: +1 979 847 8578.

E-mail addresses: kpraveen@cs.wright.edu (P. Kakumanu), iiass.anna@tin.it (A. Esposito), ogarcia@unt.edu (O.N. Garcia), rgutier@cs.tamu.edu (R. Gutierrez-Osuna).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.09.005

experimentally that using only lip movements it is possible to raise the word intelligibility in noisy conditions from 22.7% to 54% on average, and up to a maximum of 74%. In addition, lip movements are useful for understanding expressions and developing tools for human–machine communication (Bernstein and Benoit, 1996). Jourlin et al. (1997) has shown that lip trajectories contain information about a person's identity, and that they can be used to improve the performance of speaker verification systems. Moreover, the integration of lip trajectory information when added to automatic speech recognition for text-to-speech systems improves system performance and enhances the intelligibility of synthetic speech (Benoit and Le Goff, 1998; Massaro, 1997; Rogozan and Deléglise, 1998). Central to any of these applications is the development of robust and accurate models of lip movements, audio–visual sampling at a reasonable rate and, most importantly, a careful synchronization with the produced speech.

A number of lip modeling techniques have been proposed in the past two decades. Several of these models have been developed for the purpose of joint audio–visual speech recognition, and are based directly on images or statistics of lip motion (Bregler and Omohundro, 1995; Coianiz et al., 1995; Kass et al., 1988). Since they use static images, changes in lip poses have to be accomplished by interpolation, missing important dynamic features. Few models are based on the anatomy of the lips, attempting to simulate the muscles in the mouth region (Essa, 1995; Waters and Frisbie, 1995). Yet, the muscles surrounding the lips are extremely complex, and have proved to be difficult to model and subsequently control accurately. Finally, when embedded in more complex systems for synthesis and facial animation, lip movements are often modeled by static articulatory parameters which are then converted into dynamic coefficients by an active lip-shape model or by hand (Abry et al., 1989; Caldognetto et al., 1989; Leps¢y and Curinga, 1998; Luttin et al., 1996; Beskow, 1995; Goldschen, 1993; Parke, 1982; Lee et al., 1995). Some of these models do not accurately reproduce the dynamics and the synchronism between utterances and the corresponding synthesized facial movements since the oral trajectories are not estimated directly from the corresponding sampled speech signal.

Previous work on predicting lip trajectories from speech can be grouped into two major approaches. The first of these essentially associates the acoustic parameters with lip parameters by exploiting key lip positions (e.g., visemes, action units, codewords, control parameters) or phonetic segmentation (Nakamura and Yamamoto, 2001; Ezzat and Poggio, 2000; Pelachaud et al., 1996; Waters and Levergood, 1993; Cohen and Massaro, 1993; Morishima and Harashima, 1991; Arslan and Talkin, 1999; Yamamoto et al., 1998). There is, however, no uniformly predictable simple correspondence between phonetic-level acoustics and the commonly used visual units because of coarticulatory effects and the non-linear relationship between the phonetics of the produced speech and the perceived orofacial movements often influenced by prosody or mood. Therefore, the mapping from purely acoustic sequences to visual-phonetic units introduces additional uncertainty, producing a loss of information (similar to that generated by unaccounted prosody) and resulting in a less natural lip motion modeling. An alternative more recent approach is to construct a direct mapping from sub-phonemic speech acoustics onto orofacial trajectories, using either three-dimensional coordinates of facial points or dynamic articulatory parameters (Massaro et al., 1999; Hong et al., 2002; Brand, 1999; Tekalp and Ostermann, 2000). McAllister et al. (1998) were able to predict the position of the mouth for English vowels by transforming the fundamental frequency (F0) contour into a probability density function, and using the first and second moments as inputs to a bivariate predictor function. Since the method uses F0 as the basic source of information, it cannot be extended to the articulation of voiceless sounds. Lavagetto (1995) synthesized lip parameters from Linear Prediction Coefficients (LPC) using a Time Delay Neural Network (TDNN) to capture the temporal relationship between acoustics and lip movements. Massaro et al. (1999) predicted facial parameters from Mel Frequency Cepstral Coefficients by training a Multilayer Perceptron and using a certain number of past and future frames to model coarticulatory effects. Hong et al. (2002) used a family of multilayer perceptrons, each trained on a specific phoneme, and a seven-tap delay line to capture coarticulation. Brand (1999) predicted the trajectories of 3D facial points using a combination of LPC and RASTA–PLP coefficients, along with an entropy-minimizing algorithm that learned simultaneously both the structure and the parameters of a Hidden Markov Model (HMM). However, methods based on TDNNs or HMMs rely on an iterative estimation of non-linear relationships that result in computationally intensive training phases.

At this juncture, we evaluate quantitatively various acoustic models and learning approaches by their ability to predict lip trajectories. Several significant questions are addressed in this paper: To what extent can lip motion be predicted from speech acoustics for a given measured accuracy? Which are the acoustic features that encode better the relationships between speech and lip dynamics? Does acceptable realism require either articulatory- and/or perceptually based acoustic features, or both? And the central question is: Which speech acoustics model, among those frequently used, is most suitable to represent these features? To address these questions, four acoustic modeling algorithms were considered for the purpose of feature extraction: Linear Prediction Coefficients, Perceptual Critical Band Features, Mel Frequency Ceptrum Coefficients, and Linear Spectral Frequencies (Markel and Gray, 1976; Aversano et al., 2001; Duttweiler and Messerschmitt, 1976; Itakura, 1975). Our approach predicts lip movements directly from these acoustic features, without any intermediate transformation, using a simple and effective *k*-nearest-neighbor (KNN) procedure (Duda et al., 2001, pp. 177–186). This approach helps to put the models in an equal basic footing for comparison. To include coarticulation effects, context is explicitly taken into account by means of a tapped-delay line. In addition, some prosodic information is also considered by adding energy and F0 to the aforementioned acoustic features (Parsons, 1986).

2. System overview

The audio–visual system employed in this research consists of two color cameras (Kodak ES310), two dedicated PCs and frame grabbers capable of acquiring two 648 × 484 video streams to a hard drive at 60 frames per second (fps). Speech is captured on one of the PCs using a shotgun microphone (Sennheiser K6/M66 with Symetrix 302 preamplifier) and saved to disk using a proprietary file format that interleaves 1/60 s of the corresponding audio between video frames to ensure accurate synchronization. Once the data has been saved to disk, the 3D coordinates of various facial points are tracked using stereo correspondence, from which the corresponding MPEG-4 Facial Animation Parameters (FAP) are subsequently extracted (Tekalp and Ostermann, 2000). A number of acoustic features are then extracted from each 1/60 s window of the audio track. These procedures are described in the following subsections.

2.1. Video processing

To facilitate accurate tracking of the facial dynamics, 27 markers are placed on the face of a subject at various facial key positions defined by the MPEG-4 standard (Tekalp and Ostermann, 2000), as shown in Fig. 1. Each marker position is independently tracked by finding the maximum of the average cross-correlation across the three color planes. This search is performed in a local region of interest centered on the marker position in the preceding frame. The initial position of each marker is manually entered by means of a graphical user interface. This process is performed independently on each of the two stereo sequences. To obtain the 3D coordinates of the MPEG-4 facial points from two stereo images, we apply a calibration procedure comprising a prism with fiduciary markers and a

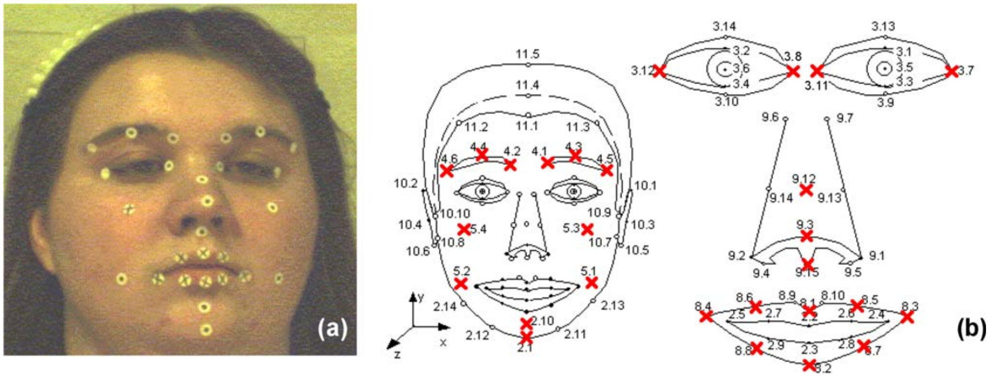

Fig. 1. (a) Neutral face of the subject with visual markers. (b) The 27 MPEG-4 feature points used in this work (denoted by cross-marks) (Tekalp and Ostermann, 2000).

calibration tool to establish the calibration correspondences and calibration matrix (Bryll et al., 1999). The algorithm is based on Tsai's stereo calibration algorithm that takes radial lens distortion into consideration (Tsai, 1987). Although the subject is asked to stand still during data collection, head motion is unavoidable due to the speaker's natural tendency to move her head during speech. Hence, the 3D coordinates of the tracked points also contain movements due to head translations and rotations. Head motion is removed by estimating head pose from eight MPEG-4 facial points (points 5.3, 5.4, 9.3, 9.12, 3.7, 3.8, 3.11 and 3.12) (Arun et al., 1987). Finally, the coordinates of the neutral face in the first frame of each sequence are subtracted from the remaining frames to yield a vector of 81 (27×3) relative displacements following the MPEG-4 Facial Animation Parameters (FAP).

2.2. Data collection

The database used in this study consists of 450 phonetically balanced sentences (TIMIT compact set, Garofolo et al., 1988), spoken by a female American English speaker born and raised in Ohio. On each recording session, the speaker is given a list of sentences, and asked to produce them in series of five sentences. To reduce intra-speaker variations, each sequence of five sentences is recorded starting from a neutral facial expression, such as that of Fig. 1(a), which serves as a baseline for each take.

This dataset is divided into three separate collections of sentences: a training set with 316 (70%) sentences, a validation set with 67 (15%) sentences, and a test set with 67 (15%) sentences, for a total of over 97000 frames. The training set is used to create a lookup table for the KNN algorithm, whereas the validation set is used to indicate appropriate structural parameters for each acoustic model. The performance of the final models is evaluated on the independent test set.

2.3. Speech processing

The audio signal is sampled at 16 kHz and filtered with a spectral subtraction algorithm to reduce electrical noise, mostly 60 Hz (Boll, 1979). Spectral subtraction operates by estimating the frequency spectrum of the background noise and subtracting it from the noisy speech spectrum. The filtered signal is then processed in blocks of 1/60 s with a 33% overlap between consecutive blocks. Each block is pre-emphasized with an FIR filter ($H(z) = 1 - az^{-1}$; $a = 0.97$) and weighted with a Hamming window to avoid spectral distortions (Rabiner and Schafer, 1978). Four different acoustic models (see Section 3) are used to extract information from each frame. In addition, F0 and signal energy, which capture some prosodic information, are also considered as a separate acoustic model. This prosodic model will be denoted by PP (for power and pitch) in what follows. F0 is computed using the cube clipping method (Parsons, 1986).

2.4. Audio–visual mapping

To account for coarticulation, context is included by associating each video frame $v(t)$, where t denotes the frame index, with an acoustic vector $a_n(t)$ containing n past and future audio frames:

$$\begin{aligned} a_n(t) = [&a(t-n), \ldots, a(t-2), a(t-1), \\ &a(t), a(t+1), a(t+2) \ldots, a(t+n)] \end{aligned} \tag{1}$$

The audio–visual pairs $[a_n,v](t)$ form the training database (i.e., a lookup table.) For $k = 1$ neighbors, the KNN audio–visual mapping is used as follows. To synthesize a new animation frame $\hat{v}(t)$ from an audio vector $\hat{a}_n(t)$, the procedure finds the example in the audio–visual lookup table whose audio vector $a_n(j)$ is closest to $\hat{a}_n(t)$ in Euclidean distance, and uses its video vector $v(j)$ as the prediction:

$$\hat{v}(t) = \{v(t); t = \arg\min_j \|\hat{a}_n - a_n(j)\|\} \tag{2}$$

The absence of a training phase in the KNN procedure allows us to perform a systematic comparison of the different acoustic models, as well as study the effect of the context duration n, which would be impractical if a training-intensive iterative learner (e.g., a time-delay neural network) were used. The performance of the KNN mapping is measured by the Mean Square Error (MSE) and Correlation Coefficient (CC) between the predicted and original video trajectories, defined as:

$$\mathrm{MSE} = \frac{1}{\sigma_v^2}\frac{1}{N}\sum_{t=1}^{N}(\hat{v}(t) - v(t))^2,$$

$$\mathrm{CC} = \frac{1}{N}\sum_{t=1}^{N}\frac{(\hat{v}(t) - \mu_{\hat{v}})(v(t) - \mu_v)}{\sigma_{\hat{v}}\sigma_v} \tag{3}$$

where N is the total number of frames in the dataset, and μ and σ are the mean and standard deviation of the video parameters. These measures of performance are not computed on the predicted

3D trajectories, but rather on the following three articulatory parameters, which have been shown to contain the most important features for automatic visual phones (visemes) recognition (Montgomery and Jackson, 1983; Finn, 1986): mouth height (MH = 8.1y − 8.2y in the MPEG-4 standard), mouth width (MW = 8.3x − 8.4x) and chin height (CH = 9.3y − 2.10y).

2.5. Animation

To verify the accuracy of our tracking and the resulting predictions, we have developed an MPEG-4 compliant player capable of rendering a facial animation at 60 fps on mid-range performance computers with minimal OpenGL acceleration. The player is based on a one-layer pseudo-muscle model consisting of a mesh with 876 triangles, 28 pseudo-muscles to allow facial expressions and movement, and an underlying non-penetrable ellipsoidal structure that approximates the skull and the jaw. This novel ellipsoidal structure is used to detect and prevent penetration of the skull/jaw by facial points, thereby giving a sense of volume and a more realistic dynamic behavior. The details of this model have been presented elsewhere (Gutierrez-Osuna et al., 2002, 2005). The complete audio–visual capture, tracking, and prediction procedure are depicted in Fig. 2.

3. Speech coding models

In using speech to predict lip motion, it is essential to identify a technique that efficiently captures information correlated to the lip motion, since the prediction will rely strongly on context surrounding this information. For this purpose, five individual acoustic models are selected for evaluation: two that model articulatory parameters of speech production, two that model perceptual cues related to the human auditory system, and a prosodic model. In addition, two hybrid vectors models that combine information from the previous models are also considered.

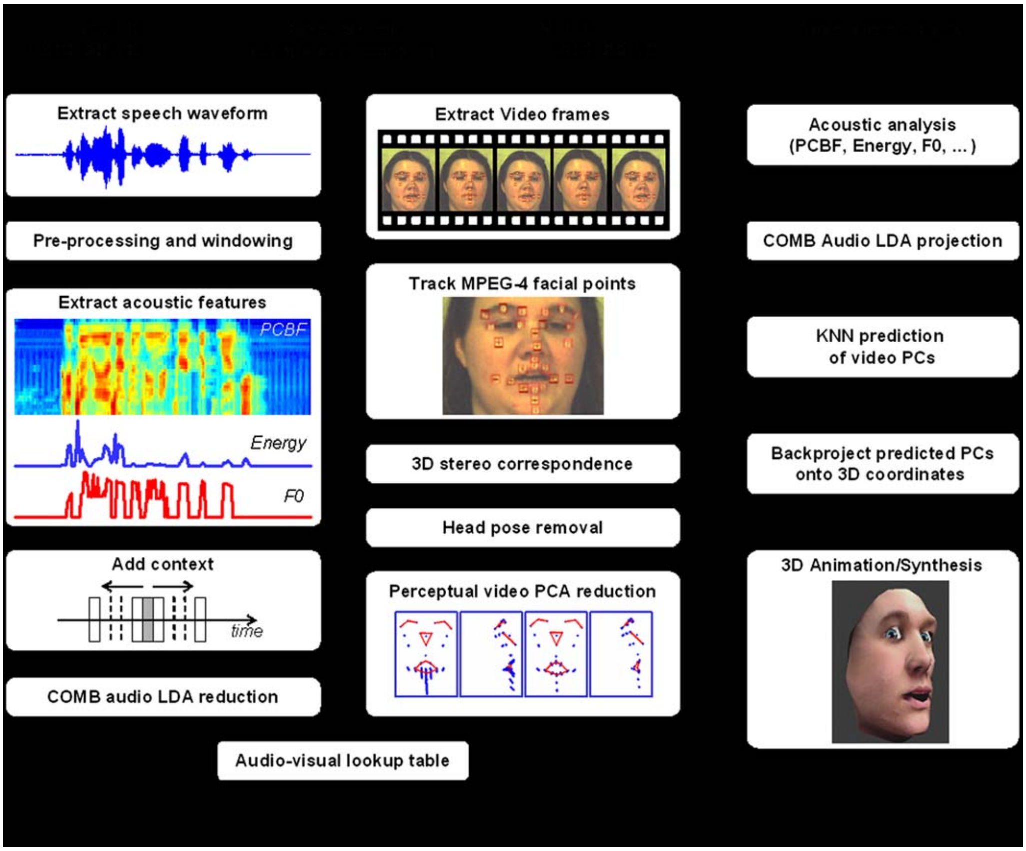

Fig. 2. Speech-driven facial animation procedure.

3.1. Articulatory models

The rationale for using an acoustic model of the speech production dynamics is that the model may also be able to capture the visual articulatory movements associated with these dynamics. For this reason, two articulatory models are considered in this work: Linear Predictive Coding and Linear Spectrum Frequencies, a brief overview of which is presented next.

3.1.1. Linear Prediction Coding (LPC)

Linear Predictive Coding models speech as a two-source/filter system excited at the glottis (the signal source) by noise (voiceless sounds) and by a periodic impulse train (voiced sounds). This model is articulated through a frequency dependent transformation representing the vocal tract configuration (the filter). Since the vocal tract configuration changes as a function of the speech sound being produced, the filter parameters are computed over a short window of usually no more than 20–30 ms approximating a static filter over this short period. The computation of assumed time-invariant filter parameters (in our case on a 16.6 ms window) is performed such as to minimize the error between the real signal and the one predicted from the model using the autocorrelation method (Markel and Gray, 1976). Since lip and tongue positions affect vocal tract configuration and the LPC coefficients are designed to encode articulatory movements, they can be expected to provide good estimates for lip-syncing graphic models driven by speech.

3.1.2. Linear Spectrum Frequencies (LSF)

The Linear Spectrum Frequencies model, also known as Line Spectrum Pairs (LSP), was introduced by Itakura (1975) as an alternative parametric model to LPC. The LSF parameters are computed through two polynomials, $P(z) = A(z) + z^{-(p+1)}A(z^{-1})$ and $Q(z) = A(z) - z^{-(p+1)} \times A(z^{-1})$, where $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$ is the inverse filter transformation of the LPC filter. The roots of these two polynomials determine the line spectral frequencies of $A(z)$. Soong and Juang (1993) have shown that if $A(z)$ is the minimum phase, then the zeros of $P(z)$ and $Q(z)$ are on the unit circle and they are interlaced with each other. The LSFs correspond to the frequencies w_i of these zeros, which are in ascending order, $0 < w_i < \pi$, $1 \leqslant i \leqslant p$ (p is the order of the prediction), ensuring thus the stability of the LP filter, which is an important requirement for speech coding applications. In our case, the LSF parameters are calculated from $P(z)$ and $Q(z)$ by applying a discrete cosine transformation (DCT) (Soong and Juang, 1993). LSF has a number of advantages compared to the LPC representation, including a bounded range, a sequential ordering of the parameters, and a simple check for the filter stability. In addition, the temporal evolution of the LSF coefficients has been proven to follow the formant frequency trajectories from one phone to another (Kim and Lee, 1999).

3.2. Perceptual models

A rationale for the use of perceptually related transformations of the signal, such as a logarithmic frequency scale, is the fact that speech is a learned skill which is affected not only by the production mechanism of the vocal tract but also by the manner in which speech has been perceived by the human learners who reproduce the learned sequences. According to this rationale, the production of speech would be a fortiori biased by the fact that speech production is learned through the human perception mechanism. This mechanism was modeled by Seneff (1988). This powerful hypothesis strongly relates recognition of speech to the human perception mechanisms.

Moreover, the previously described articulatory models are based on the assumption that formant frequency values and their trajectories are the cues for a speech production model. Therefore, they extract features that are based on the unambiguous identification of important peaks in the spectral envelope, and require a robust method for disambiguating these from additional peaks that may occur for a variety of reasons. However, it has been proven (Klatt, 1982) that neither the overall level, nor the exact location of the pattern in the frequency domain or the overall tilt of the spectrum are important for phonetic perception. Rather, it is the changes in the shapes and relative locations of the major spectral features that are of greater import (Klatt, 1982). In light of these considerations, it is also important to consider two perceptually based processing techniques, using concepts from the psychophysics of hearing in order to obtain an estimate of the auditory spectrum.

3.2.1. Perceptual Critical Band Features (PCBF)

The first perceptually based model consists of processing the speech signals through a critical-band resolution of the Fourier spectrum described by $\text{Bark} = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f^2}{7500^2}\right)$, where the acoustic frequency f is mapped onto a perceptual frequency scale referred to as *critical band rate* or *Bark*. This approximate perceptually modified spectrum is used to capture dynamic aspects of the spectral envelope patterns, resulting in a vector of Perceptual Critical Band Features (Aversano et al., 2001). Note that RASTA–PLP also contains a perceptual component, but then the signal is further processed by including a predictive modeling (PLP) and a temporal filtering (RASTA) process to make the extracted features closer to the formant trajectory (Hermansky and Morgan, 1994).

3.2.2. Mel Frequency Cepstral Coefficients (MFCC)

Another perceptually based alternative to PCBF is to use MFCC (Duttweiler and Messerschmitt, 1976). In this case, an LPC or a DFT spectrum of the signal is frequency warped through a Mel-scale transformation $\text{Mel} = 2595\log_{10}\left(1 + \frac{f}{700}\right)$, and amplitude-warped using a logarithmic transformation. This is done using a bank of N band-pass filters arranged linearly along the Mel scale. The bandwidth of each filter is chosen equal to the Mel-scale bandwidth of the corresponding filter center frequency. The log-energy output of these N filters X_k, is then used to compute a given number M of MFCC coefficients $C_n = \sum_{k=1}^{N} X_K \cos\left[n - \left(k - \frac{1}{2}\right)\frac{\pi}{20}\right]$, for $n = 1, \ldots, M$.

Fig. 3 illustrates the results of each of the six individual speech coding models for the TIMIT sentence "*She had your dark suit in greasy wash water all year*". Each result is shown as an image, where each column represents a time step, each row represents the trajectory of a coefficient over time, and pixel intensities represent the feature values. These images show that each model uses a different encoding to characterize the information embedded in the speech signal. In addition, the dynamic range of the coefficients (not shown in the figure) is also unique to each model, suggesting that the acoustic features are also weighted differently. These images are, in general, difficult to interpret by an untrained eye. Note, however, how the PCBF spectrogram shares a close resemblance to a short-time FFT spectrogram. The average cross-correlation coefficients between every pair of features $E\lfloor \text{CC}(X_i, Y_j) \rfloor_{\forall i,j}$, where X and Y are the individual acoustic models (LPC, LSF, MFCC, PCBF), are shown in Table 1. Note that the diagonal elements need not be equal to one, since they are the average cross-correlation between pairs of features from the same acoustic model. These figures are indicative of the degree of redundancy that exists between the different acoustic features. PCBF and LSF, in particular, appear to have highly correlated features, an observation

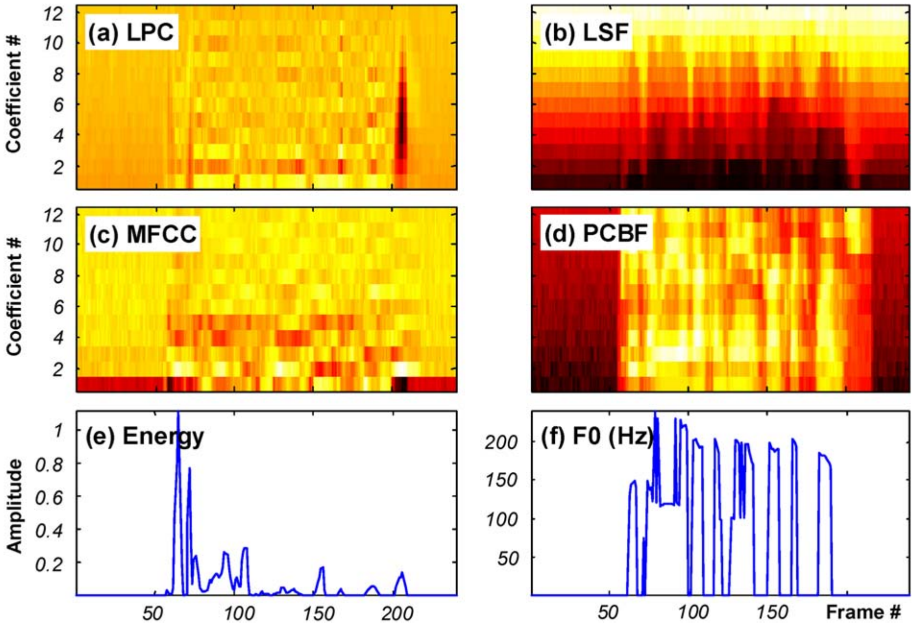

Fig. 3. Result of the six preprocessing algorithms.

Table 1
Average cross-correlation coefficients among pairs of features

	LPC	LSF	MFCC	PCBF	PP
LPC	0.3132	0.3120	0.1052	0.2817	0.2819
LSF		0.5809	0.1948	0.5247	0.3926
MFCC			0.1962	0.2460	0.1883
PCBF				0.8965	0.6810
PP					0.8126

that is consistent with the images in Fig. 3. These cross-correlation coefficients drop when we consider pairs of features from different models, which suggests that a combined feature vector with information from multiple models may be able to outperform the individual models. This issue is investigated next.

4. Combining coding models

The use of hybrid vectors with features from multiple acoustic models is common practice in speech recognition, speaker verification, and speech-driven facial animation problems (Tibrewala and Hermansky, 1997; Sharma et al., 1998; Brand, 1999; McAllister et al., 1998). For this reason, we also explore possible enhancement of the audio–visual predictions by combining information from the previously mentioned speech models. In a first instance this is accomplished by concatenating the individual feature vectors as a 150-dimensional[1] hybrid vector, a model that we will refer to as COMB (for combination). However, KNN procedures are notoriously sensitive to dimensionality. To address this issue, a novel supervised dimensionality-reduction procedure is used.

Given that our objective is to find audio features that are discriminative of orofacial configurations and their dynamics, we use a signal classification criterion such as Fisher's Linear Discriminant Analysis (LDA) to project the hybrid vector onto a lower-dimensional subspace (Duda et al., 2001; pp. 117–124). In order to obtain a LDA transformation of the input space (audio), the audio–visual vectors must first be assigned to distinct output classes. Since the output space (video) is not categorical, we cluster training data into groups with similar orofacial configurations by means of vector quantization (VQ) (Duda et al., 2001; pp. 526–528). Our VQ implementation uses a standard recursive binary splitting procedure: starting from one cluster containing all the samples, VQ recursively splits each cluster into two sub-clusters (along the direction of largest variance) until a desired number of codewords is obtained. Once a codeword index has been assigned to each audio–visual pair, an LDA projection matrix is computed to project the audio vector into a subspace that maximizes the separation of the different video clusters. Section 6.3 will report on the performance of the resulting vector (COMB–LDA) as a function of both the number of video codewords and of the number of LDA eigenvectors preserved for the projection.

5. Principal component analysis of orofacial motion

The video processing procedure described in Section 2.1 yields an 81-dimensional vector containing 3D coordinates of 27 facial points for each frame. This representation is clearly redundant since the motion of these facial points is highly coupled and also contains noise inherent to the video capture procedure. Therefore, we seek to form a low dimensional projection that captures the principal directions of orofacial motion, and also filters out some of the experimental noise in the process. This is accomplished through Principal Component Analysis (PCA) (Duda et al., 2001; pp. 568). The PCA procedure accepts an 81-dimensional video vectors (on the 316 training sentences), and produces a series of eigenvectors (81-dimensional vectors) that are aligned with the principal components (PC) or directions of variance. A subset of these PCs is then selected as the new representation for the video. To synthesize a new animation, the KNN audio–visual mapping is used to predict these selected PCs from the audio. The predicted PCs are then projected back into the original 81-dimensional space through the inverse Karhunen-Loéve transform in order to reconstruct the 3D orofacial trajectories. The details of this procedure can be found in (Gutierrez-Osuna et al., 2002, 2005).

An appropriate number of PCs for the video vector was determined with a pilot perceptual test involving nine human subjects, all members of our research group not working on this project. Subjects were asked to evaluate the perceptual realism of five different animations containing the 2, 4, 6, 8 and 12

[1] (12 features/frame × 3 frames/model × 4 models) + (1 F0 + 1 Energy) × 3 frames.

Table 2
Perceptual evaluation of the five PCA models

Pair-wise comparisons	Model B (second)				
	1:2	1:4	1:6	1:8	1:12
Model A (first)					
1:2	AAAABAAAA	BABABBBBB	ABABBBBAB	BBAAABBBB	ABBBABBAA
1:4	BAABAAAAA	BABBBAABA	BBBAABABA	BBBAAABBB	ABAAAABBB
1:6	ABABAAABA	AAAABAABA	BBBABBBBB	BBAAABBBB	AAAAAABBA
1:8	BAAABAABA	ABBABAABA	BABBBBABB	BBAABBBBB	BBAAAABBB
1:12	ABBAABAAA	BBBABBABA	BBAABBABA	BBBBBBBAA	BAAAABBBB

Each cell in the table represents a paired test, where model A (row) was presented first, followed by model B (column). The response of the 9 subjects (in order) is shown on each cell: an A indicates that the subject chose Model A.

largest PCs, respectively. Each of the five animations consisted of four sentences from the validation set. For each validation sentence, the 81 video coordinates were projected onto a low-dimensional space (i.e., 2, 4, 6, 8 and 12 dimensions) using the PCA eigenvectors from the 316 training sentences, and then back-projected onto the original 81 dimensions. Therefore, this test evaluated the ability of PCA to capture the perceptually relevant directions of orofacial motion, not the prediction capabilities of the KNN audio–visual mapping.

The evaluation was performed as a series of paired tests, for a total of 25 comparisons (5 models × 5 models) per subject. For each paired test, the subject was presented with the four bimodal sentences from the first model (model A), followed by the four sentences from the second model (model B), and then asked to choose the more realistic of the two animations. Note that each pair of models was presented twice, including each model against itself, in order to identify potential biases towards the order of presentation. The results are summarized in Table 2. Aggregating these results, the models with 2, 4, 6, 8, and 12 PCs received 32, 47, 55, 51 and 40 favorable votes, respectively. Though the perceptual choice between 6 and 8 PCs was not statistically significant (t-test, $\alpha = 0.05$), the distribution of votes indicates that a model with 6 PCs provided the best perceptual realism. The performance dropped for 2 and 4 PCs, indicating that fewer eigenvectors cannot capture sufficient orofacial motion. Similarly, the performance also dropped for 8 and 12 PCs, which suggests that the additional eigenvectors capture noise in the data.

6. Evaluation of the audio–visual models

Having presented the various acoustic coding models, and having explored a perceptually based representation of orofacial motion, we are now ready to investigate the performance of the nearest-neighbor audio–visual model as a function of two parameters: (1) temporal context, defined by the length and sampling rate of the coarticulation window, and (2) spatial locality, defined by the number of nearest neighbors in the audio–visual mapping. In addition, we also analyze the performance of the LDA–COMB hybrid model in terms of visual codebook size and number of audio LDA dimensions.

6.1. Sensitivity to temporal context

It is known that facial dynamics are strongly influenced by backward and forward coarticulation (Kühnert and Nolan, 1999) a phenomenon that is incorporated in our model with a moving window of fixed duration. This section investigates the performance of the KNN audio–visual mapping as a function of this temporal context, as captured by the tapped-delay line in Eq. (1). Throughout these experiments, a fixed number of $k = 15$ neighbors is used (the sensitivity to this parameter is analyzed in Section 6.2). Fig. 4(a) and (c) illustrates the performance of the different acoustic models on the validation set. Context duration was systematically increased until a global optimum (min MSE or max CC) was found for all models. Note that, for COMB and COMB–LDA, results beyond $n = 8$ are not included. For these two models, the dimensionality of the resulting audio vector (e.g., 950 features for $n = 9$), combined with the number of training samples (over 67000 frames from 316 sentences), prevented the dataset from even being loaded into main memory. This limitation prompted us to consider alternatives for reducing the initial dimensionality of the acoustic vector.

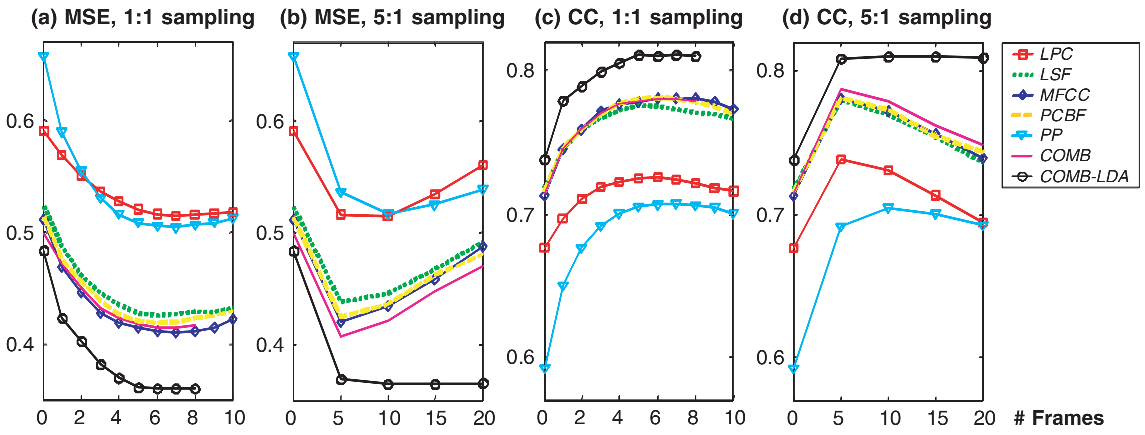

Fig. 4. Effect of context duration with 1:1 sampling and 5:1 subsampling (on validation data).

6.1.1. Subsampling the coarticulation window

The logical solution to the dimensionality issue was to establish a trade-off between context duration and temporal resolution. For a given context duration n, the coarticulation window can be subsampled by discarding all but the last of every T consecutive frames (T:1 subsampling). For 1:1, the scheme is equivalent to the acoustic vector in Eq. (1). For higher values of T, a longer context duration can be employed at no cost in dimensionality. A reasonable choice was experimentally found to be a 5:1 ratio:

$$a_n(t) = [a(t-n), \ldots, a(t-5), a(t), a(t+5), \ldots, a(t+n)], \quad n = 0, 5, 10, \ldots \tag{4}$$

With 5:1 subsampling and $n = 5$, the acoustic vector is thus reduced to 3 frames: one at −83 ms (5 × 16.6 ms), one at 0 ms, and one at +83 ms. Similarly, with 5:1 subsampling and $n = 10$, the acoustic vector is reduced to 5 frames: one at −167 ms, one at −83 ms, one at 0 ms, one at +83 ms, and one at +167 ms. The performance of the 5:1 subsampled audio vector is shown in Fig. 4(b) and (d). The best context length (using $k = 15$) for both 1:1 and 5:1 subsampling are summarized in Table 3. Several conclusions can be extracted from these results:

- All models perform poorly for $n = 0$, since the mapping does not incorporate coarticulatory effects, and sharply increase their performance with larger context durations until a global optimum is reached. Performance degrades for higher values of n as the mapping considers a temporal window that is larger than the duration of the coarticulatory effects and our stationary parameter assumption may not hold. A maximum value is obtained for $n = 10$ frames, or 167 ms of both forward and backward coarticulation.[2] This result is consistent with the duration of most of the stressed and unstressed canonical syllable forms (V, VC, CV, CVC), as noted by Greenberg et al. (2003).
- COMB–LDA consistently provides the best performance (Table 3). The worst performers are PP, as could be expected since it only captures prosody, and LPC. The remaining acoustic models, LSF, MFCC, PCBF and COMB, perform similarly.

Table 3
Best context duration and spatial locality (on validation data)

Audio model	Context duration ($k = 15$)				Spatial locality (5:1, $n = 5$)	
	1:1 sampling		5:1 sampling			
	Max CC	n	Max CC	n	Max CC	k
LPC	0.7256	7	0.7379	5	0.7379	15
LSF	0.7758	6	0.7797	5	0.7798	13
MFCC	0.7807	8	0.7887	5	0.7887	15
PCBF	0.7815	7	0.7812	5	0.7812	15
PP	0.7072	8	0.7043	10	0.6956	16
COMB	0.7819	7	0.7873	5	0.7911	20
COMB–LDA	0.8109	8	0.8110	10	0.8117	38

[2] Interestingly, coarticulation effects seem to be language-dependent. In English, forward coarticulation is more pronounced, whereas in French or Italian backward coarticulation is more dominant. In fact, unpublished results by our group indicate that prediction of orofacial motion is more accurate with forward than with backward coarticulation.

- 5:1 subsampling does not affect performance, as compared to 1:1 sampling. In fact, with the exception of PP and PCBF, all acoustic models benefit from subsampling (Table 3). This suggests that most of the coarticulatory effects are *redundantly spread* during a syllable. This is an interesting result worthy of further study.

6.1.2. Comparison with delta features

The proposed tapped-delay line in Eqs. (1) and (4) represents an alternative to the more conventional delta and delta–delta features. For validation purposes, we compared the predictive accuracy of the two approaches. Following Picone (1993), delta features were computed as:

$$\dot{a}(t) = \frac{\mathrm{d}}{\mathrm{d}t}a(t) \approx \sum_{n=-F}^{F} n * a(t+n) \tag{5}$$

Delta–delta feature were computed by reapplying Eq. (5) to the previously computed delta features, resulting in a new audio feature vector $a_{\mathrm{dd}}(t) = [a(t), \dot{a}(t), \ddot{a}(t)]$. Fig. 5 shows the results of the comparison as a function of F, the window size used to compute the derivative in Eq. (5). As a reference, the last sample on each curve corresponds to the tapped-delay features vector in equation (4) for $n = 5$. These results show that (i) our tapped-delay subsampling technique clearly outperforms delta and delta–delta features for every speech coding model, and (ii) projection of delta and delta–delta features onto an LDA subspace dramatically improves the prediction of orofacial motion from audio.

Based on these results, a 5:1 subsampling and context duration of $n = 5$ will be used in the remaining sections of this paper for all the acoustic models. Although the optimum for COMB–LDA occurs at $n = 10$, the marginal improvement in performance with respect to $n = 5$ (Fig. 4(d)) does not warrant a nearly two-fold increase in dimensionality for the acoustic vector (3 frames for $n = 5$ vs. 5 frames for $n = 10$).

6.2. Sensitivity to spatial locality

Having analyzed the impact of coarticulation on audio–visual prediction, and having determined an appropriate context duration, we turn our attention to the spatial locality imposed by the KNN rule. A value of $k = 1$ allows the KNN rule to produce a highly localized mapping since only the nearest example in audio space is used for the video prediction. Larger values of k control the locality of the mapping by averaging the prediction across a few nearest neighbors, reducing the sensitivity of the algorithm to noise (placement of markers, uncertainty in different takes, etc.) and providing smoother decision boundaries (Wilson and Martinez, 2000). Extension of the KNN audio–visual mapping in Eq. (2) to k neighbors is straightforward:

$$\hat{v}(t) = \left\{ \frac{1}{k}\sum_{i=1}^{k} v(k_i); \left(k_1 = \arg\min_{j} \|\hat{a}_n(t) - a_n(j)\|\right) \wedge \left(\|\hat{a}_n(t) - a_n(k_i)\| < \|\hat{a}_n(t) - a_n(k_j)\| \forall j > i\right) \right\} \tag{6}$$

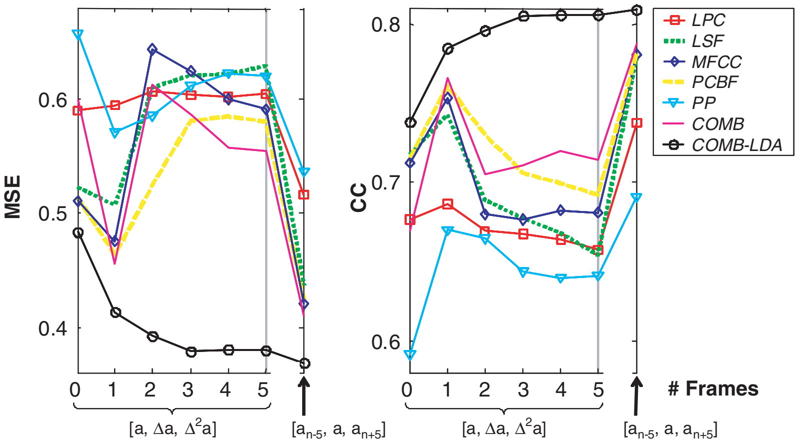

Fig. 5. Performance of the proposed tapped-delay vector as compared to delta and delta–delta features (on validation data). Results are shown as a function of the window size of the derivative.

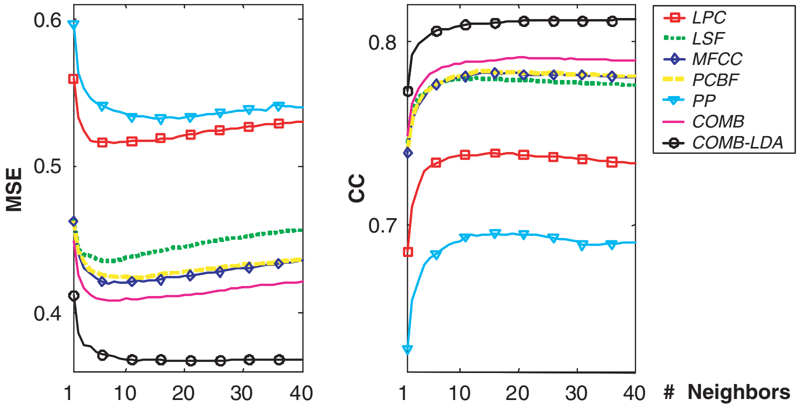

Fig. 6. Effect of spatial locality (on validation data).

where k_i is the index of the ith nearest neighbor. That is, the predicted orofacial parameters are estimated as the average video over the k nearest audio neighbors.

Fig. 6 illustrates the prediction results on the validation set for different values of k. The relative performance of the different models is the same as in the previous section: COMB–LDA is the best model, PP and LPC are the worst performers, and the remaining models are nearly equivalent. The performance of each model is lowest at $k = 1$ neighbors, and increases steadily. Performance reaches a plateau and/or begins to drop in the neighborhood of $k = 15$, although the optimum for COMB–LDA occurs at $k = 38$. Thus, these results indicate that all the audio–visual models benefit from averaging across video vectors in the training set that have similar acoustic features. Based on these results, a value of $k = 15$ is chosen for all models for the remaining sections in the manuscript.

6.3. Sensitivity to codebook size and LDA eigenvectors

The success of the COMB–LDA acoustic model lies in the supervised nature of the projection procedure: the model finds a low-dimensional representation of the audio data that best discriminates a subset of facial-configuration codewords. The performance of the model depends on two parameters: the codebook size Q, which controls the accuracy of the video-quantization procedure, and the number of LDA eigenvectors D, which controls the dimensionality of the acoustic vector, or the amount of audio variance preserved for the audio–visual mapping. In previous sections, reasonable values $Q = 128$ codewords and $D = 13$ eigenvectors were used. This section explores the sensitivity of the model to these two parameters.

Fig. 7(a) illustrates the relationship between Q and D for six different codebooks (16, 32, 64, 128, 256 and 512 codewords), in terms of the percent of total audio variance that is preserved in the LDA projection. Note that the number of non-zero LDA eigenvectors is limited by (1) the dimensionality of the COMB audio space (150 dimensions for 5:1 subsampling and $n = 5$), and (2) by the number of video codewords due to the rank of the between-class scatter matrix (Duda et al., 2001; pp. 124) ($D \leqslant \min[150, Q - 1]$). This explains why, for $Q = 16$ codewords, only 15 eigenvectors are needed to capture 100% of the total audio variance. As the codebook size increases, so does the number of eigenvectors required to capture a given percentage of the audio variance. These results also indicate that, as expected, the COMB audio vector is highly redundant, since a large percentage of the total variance is contained in the first few eigenvectors (99%, 98%, 95%, 91% and 84% for $Q = 32$, 64, 128, 256 and 512, respectively, in the first 20 eigenvectors).

The behavior of the COMB–LDA model on the validation set is shown in Fig. 7(b). The performance increases with the first few LDA eigenvectors in a very pronounced manner, and eventually saturates since most of the discriminatory information is contained within those first few eigenvectors. The performance also increases with codebook size, but only moderately, reaching a maximum at $Q = 128$ codewords and $D = 13$ eigenvectors. Hence, these values will be used in the next section to evaluate the final performance of the models on an independent test set.

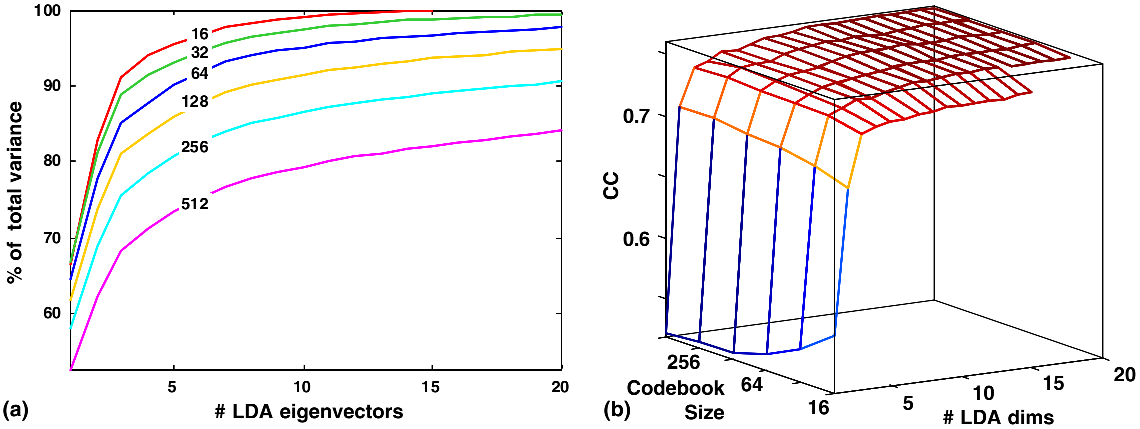

Fig. 7. (a) Percent of total acoustic variance as a function of video codebook size. (b) COMB–LDA performance vs. video codebook size and audio LDA dimensionality (on validation data).

7. Objective and perceptual comparison across acoustic models

The previous subsections have evaluated the audio–visual predictions as a function of various model parameters. Performance figures were obtained by generating an audio–visual lookup table from the 316 training sentences, and generating predictions for an independent validation set with 67 sentences. Based on these performance figures on validation data, final parameter values were chosen as follows: context duration $n = 5$ frames, context subsampling 5:1, facial point coordinates compression to six PCA eigenvectors, and $k = 15$ nearest neighbors. Model parameters for COMB–LDA were $Q = 128$ codewords, and $D = 13$ audio LDA eigenvectors. To evaluate the predictive accuracy and perceptual realism of these final models, video predictions were generated on an independent test with 67 sentences.

7.1. Predictive accuracy

The predictive accuracy of the models is summarized in Table 4 in terms of the average CC plus/minus one standard deviation. To determine if these differences in average performance were statistically significant, a paired t-test ($\alpha = 0.05$) was performed between every two models. The test indicated that all differences were statistically significant except for the pairs (LSF, MFCC), (LSF, PCBF), (LSF, COMB), (PCBF, COMB), and (PCBF, MFCC). Based on these results, we can conclude that (1) COMB–LDA provides the best predictive accuracy of all models, (2) PP and (surprisingly) LPC provide the worst performance, and (3) differences in performance across the remaining models are, in general, statistically not significant.

An important question is whether the differences in performance can be attributed to the source of the model (perceptual vs. articulatory) or are simply due to the parameter encoding. This issue would be critical if perceptually based models (MFCC and PCBF) had consistently outperformed articulatory-based models (LSF and LPC), or vice versa. However, since LSF performs similarly to MFCC or PCBF, the only valid conclusion is that both perceptual and articulatory information are necessary. The very fact that LSF performs well is an indication that articulatory models contain important

Table 4
Final performance (CC) of each audio model on an independent test set

	MH	MW	CH	AVG
LPC	0.8288 ± 0.0649	0.7066 ± 0.1372	0.8449 ± 0.0500	0.7901 ± 0.0623
LSF	0.8579 ± 0.0669	0.7601 ± 0.1381	0.8812 ± 0.0527	0.8330 ± 0.0609
MFCC	0.8699 ± 0.0625	0.7173 ± 0.1349	0.8921 ± 0.0437	0.8264 ± 0.0595
PCBF	0.8710 ± 0.0550	0.7075 ± 0.1652	0.8927 ± 0.0391	0.8237 ± 0.0606
PP	0.8305 ± 0.0552	0.5759 ± 0.1416	0.8362 ± 0.0589	0.7475 ± 0.0648
COMB	0.8761 ± 0.0606	0.7262 ± 0.1350	0.9009 ± 0.0385	0.8344 ± 0.0554
COMB–LDA	0.8817 ± 0.0588	0.7902 ± 0.1212	0.9072 ± 0.0385	0.8597 ± 0.0527

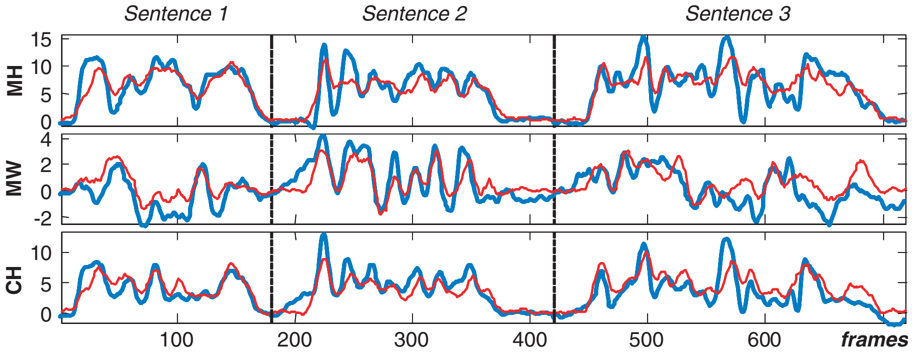

Fig. 8. Predicted (thin red trace) vs. actual (thick blue trace) trajectories for the three orofacial articulators on three unseen test sentences. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

information for visual speech synthesis. Thus, the difference in performance between LSF and LPC is not due to their origin (articulatory in both cases) but due to the particular parameter encoding: LPC does not extract information that is useful to predict orofacial motion.

The performance of the best model (COMB–LDA) is illustrated in Fig. 8 in terms of the predicted (thin red trace) versus correct (thick blue trace) trajectories of the three articulators on three independent test sentences: (a) "Clear pronunciation is appreciated", (b) "Barb's gold bracelet was a graduation present", and (c) "Those who are not purists use canned vegetables when making stew".

7.2. Perceptual realism

The perceptual realism of the models was evaluated following an experimental procedure similar to that in Section 5. Nine students not working on this project participated in the evaluation of eight models: ORIG (the ground truth), PP, LPC, LSF, MFCC, PCBF, COMB and COMB–LDA. The evaluation was performed as a series of paired tests, for a total of 64 comparisons (8 models × 8 models). For each paired test, the subject was presented with four bimodal sentences with predictions from the first model (model A), followed by the same four sentences predicted with the second model (model B), and then asked to choose the more realistic of the two animations. The results of this bimodal perception are summarized in Table 5. Aggregating these results, the total number of votes per model was: ORIG (106 votes), PP (38), LPC (51), LSF (74), MFCC (63), PCBF (83), COMB (70) and COMB–LDA (91), for a total of 576 votes (64 pairs × 9 subjects). The statistical significance (t-test; $\alpha = 0.05$) of these perceptual evaluations is shown in Table 6. Several interesting conclusions can be extracted:

- ORIG is perceived as the most perceptually realistic animation. This is to be expected, since the animation is driven directly from video data without any AV prediction stage.
- Among the actual AV predictions, COMB–LDA received the highest number of votes, followed by PCBF. It is interesting to note that the differences between the ground truth and each of these two models are not statistically significant and strongly support our proposed AV modeling approach.
- The higher performance of COMB–LDA is statistically significant with respect to all AV models except for PCBF, whereas the higher performance of PCBF is only statistically significant when compared with PP, the worst performer. Thus, COMB–LDA is the overall best AV model, both in terms of total number of votes and in terms of statistical significance. This result is consistent with the CC/MSE figures in Table 4.
- The lower performance of PP is statistically significant with respect to all AV models except for MFCC. This result is also consistent with the CC/MSE figures in Table 4.
- Differences among LPC, LSF, MFCC and COMB are found to be not statistically significance.

8. Discussion

This work has presented a quantitative and qualitative comparison of different acoustic models by their ability to predict orofacial motion from acoustic data. The models considered were (1) LPC and

Table 5
Perceptual evaluation of the seven AV models and the ground truth (ORIG)

Pair-wise comparisons	Model B (second)							
	ORIG	PP	LPC	LSF	MFCC	PCBF	COMB	LDA
Model A (first)								
ORIG	BBAABABBB	AAAABBAAB	AAAAAAAAB	BBAAABAAA	BAAABAAAA	AABAABABB	AAAABBAAB	AAAABBAAA
PP	BBABABBBB	ABABBBABB	BBBBAAAAB	BBABBBBAB	AAABBBBBB	BBBBABABB	BBABBBABA	BBBBBBBBA
LPC	BBBBABBBB	ABBABABAB	AAAAAAAAB	BBAAAAABA	BBAABBABB	BBABABABB	ABABBBAAB	BBABBBBAB
LSF	BBBBABBBB	AABAAAAAA	BAABAABAA	AABBABBBA	AABABAAAB	BBBBBBBBB	AABABBABA	BBBBABBBB
MFCC	AABBBBBBB	AABAAAAAA	AABAAABBA	BBBBBBBBB	BBBAABAAB	BBBBABABA	BBBBBBBBB	BBBBABBBA
PCBF	BBBBABBBB	AAAAAABAA	AAAAABBAA	AABABABBB	BAAABBBBA	AABBABBBB	AAAAABBBB	BABABBBBB
COMB	BBBBABBBB	AAAAAAABA	AABABAAAB	BBABBBBBB	AAABBABBB	BBBABBAAA	BBBAABAAB	BAABBABBB
LDA	ABBAABBBB	AAAAAAAAA	AAAAAAABB	BAAABAAAB	BBAABBABB	BBBBABABA	AABABABAA	AABAABAAA

Each cell in the table represents a paired test, where model A (row) was presented first, followed by model B (column). The response of the 9 subjects (in order) is shown on each cell: an A indicates that the subject chose Model A.

Table 6
Statistical significance of the perceptual differences between pairs of AV models

Pair-wise comparisons	Model B (second)							
	ORIG	PP	LPC	LSF	MFCC	PCBF	COMB	LDA
Model A (first)								
ORIG		0.0005*	0.0003*	0.0362*	0.0317*	0.1210	0.0165*	0.1525
PP			0.0499*	0.0090*	0.0579	0.0144*	0.0028*	0.0005*
LPC				0.0727	0.3637	0.0506	0.0599	0.0005*
LSF					0.0743	0.4714	0.6091	0.0474*
MFCC						0.2208	0.4751	0.0255*
PCBF							0.2557	0.5155
COMB								0.0094*
LDA								
# Votes	106	38	51	74	63	83	70	91

For each pair of models, the number of votes received by each model on each of the 9 subjects is used as a sample population. The null hypothesis is "the mean of the two populations (AV models) is equal". Small values in the *P*-value (marked with an asterisk) indicate that the null hypothesis must be rejected at a significance level $\alpha = 0.05$.

LSF, which encode articulatory features related to the dynamics of the speech production system, (2) MFCC and PCBF, which encode perceptual cues exploited by the human auditory system in processing speech signals, (3) energy and F0, which capture prosodic aspects of speech and (4) COMB and COMB–LDA, which combine information from all the above models. The models were trained, validated and tested on the 450 sentences from the TIMIT compact set. The general conclusion from our results is that the combination of information from multiple models, coupled with a supervised dimensionality reduction stage, yields statistically significant improvements in *predictive accuracy from acoustic data*. The results also indicate that LSF, MFCC, PCBF and COMB perform similarly, while LPC and PP give the lowest performance. More importantly, the COMB–LDA model received the highest ratings in terms of *perceptual realism*. Differences in perceptual realism between the ground truth and COMB–LDA (or PCBF) were found not to be statistically significant, a clear indication that our audio–visual model is able to synthesize credible orofacial motion from speech acoustics.

Several additional considerations can be drawn from our results. First, information about orofacial motion is encoded in both, perceptual and articulatory features. No significant differences in performance were found between these two groups of models. This suggests that both perceptual and articulatory hypotheses posed in Section 3 are valid: both the production system and the perceptual system play a role in shaping the speech signal. Speech (visual and acoustic) must be conceived as a signal jointly produced by the vocal tract and facial dynamics to be accurately perceived by human auditory learning that influences production and the evolved lipreading that takes place.

Second, proper use of context duration plays a critical role in accurately predicting orofacial motion. Our results show, as could be expected, that all the proposed acoustic models have a performance minimum at $n = 0$, when no coarticulatory effects are considered. The performance of individual models reaches a maximum at $n = 5$ (83 ms carryover, 83 ms anticipatory), and drops sharply for longer contexts. The COMB–LDA model is able to extend the optimum to $n = 10$ frames, though the improvements in performance beyond $n = 5$ can be considered negligible. This context length is consistent with the duration of most stressed and unstressed canonical syllable forms (Greenberg et al., 2003).

Third, the parameters of the audio–visual mapping (number of nearest neighbors k) also play a role in the final prediction accuracy. The performance of the individual models shows an optimum at around $k = 15$ neighbors, and drops moderately afterwards. In the case of COMB–LDA the maximum is reached at $k = 38$ neighbors, although the performance levels off near $k = 15$. A methodological concern that could be raised at this point, and is open for further investigation, is whether more sophisticated computational models (such as TDNNs or HMMs) could improve the prediction results. Surprisingly, preliminary experiments performed by our group using more sophisticated A/V mappings (input-output HMMs, radial basis functions, and support vector machines) provide results that are not better than those obtained by the KNN procedure presented here, and at the expense of a much higher computational complexity during training (Fu, 2002, 2005; Balan, 2003). Moreover, the computational load of KNN can be significantly reduced with the use of improved search (i.e., bucketing, k–d trees) or editing techniques (reducing the training set to a small number of prototype vectors). Preliminary but yet unpublished results by our group indicate that the entire database can be edited down to a small number of prototype vectors (e.g., 1024) without severely reducing the overall synthesis accuracy.

No comparisons are currently possible between our results and previously published studies or with intelligibility tests (e.g. CVC, VCV, etc.) as the corpus collected unfortunately did not include that data. Given that the interest in speech-driven facial animation is to implement realistic and natural talking faces, most of the results reported in the literature are presented in terms of video sequences of the animation. Although this allows the animation to be evaluated by its perceptual quality, there is no possibility to quantify the accuracy of the prediction numerically as is done here. Few quantitative results have been already reported in the literature (Lavagetto, 1995; Massaro et al., 1999; Hong et al., 2002; Brand, 1999). A comparison with these results was, however, not possible because (1) different metrics were used and (2) the audio–visual data was not publicly available. To this end, all of our audio–visual data is publicly available on the group's webpage at Wright State University (http://www.cs.wright.edu/~kpraveen/fa/).

Our study gives objective and perceptual measures that indicate the extent to which it is possible to predict facial motion directly from speech acoustics. We have shown that, although complete recovery of lip motion is not yet possible, a significant portion can be directly predicted from sub-phonemic speech features. Our results indicate that orofacial motion can be predicted with an average CC of 0.86 on novel speech sequences (test data), a relatively high performance considering the simplicity of the KNN audio–visual mapping. These results have been obtained on the complete TIMIT compact set, which contains 450 phonetically balanced sentences.

Finally, the present work shows that an appropriate encoding of the speech signal allows animation of synthetic faces in synchrony with speech through a computationally simple procedure. No complex associations between phonemes and visemes and no automatic segmentation and phoneme recognition procedures are required since the proposed method works at a sub-phonetic level. For this reason, the method can potentially be extended to multi-lingual applications because sub-units of the speech signal are not as language-specific as the phonemes–visemes association. Moreover, the method yields a realistic facial animation and recognizes the importance of the data representation for improving the prediction performance of any speech-driven facial animation system.

Acknowledgements

This research was supported by awards from NSF/CISE 9906340, NSF/CAREER 9984426, by WSU/OBR 00-509-10, and by NSF/KDI 9980054. Nishant Balan, Jessica Garringer and Tamanna Husain are gratefully acknowledged for their devoted effort in data collection and video tracking. We would also like acknowledge valuable comments from two anonymous reviewers, which greatly helped improve the quality of the manuscript.

References

Abry, C., Boe, L.J., Schwartz, J.L., 1989. Plateaus, catastrophes and the structuring of vowel systems. J. Phonet. 17, 47–54.

Arslan, L.M., Talkin, D., 1999. Codebook based face point trajectory synthesis algorithm using speech input. Speech Commun. 27, 81–93.

Arun, K.S., Huang, T.S., Blostein, S.D., 1987. Least-square fitting of two 3-d point sets. IEEE Trans. PAMI 9 (5), 698–700.

Aversano, G., Esposito, A., Esposito, A., Marinaro, M., 2001. A new text-independent method for phoneme segmentation. In: Proc. IEEE-MWSCAS Conference, Dayton, OH, pp. 516–519.

Balan, N., 2003. Analysis and Evaluation of Factors Affecting Speech Driven Facial Animation, MS Thesis, Dept. of Computer Science and Engineering, Wright State University.

Benoit, C., Le Goff, B., 1998. audio–visual speech synthesis from French text: eight years of models, designs, and evaluation at the ICP, Speech Commun. 26, 117–129.

Bernstein, L.E., Benoit, C., 1996. For speech perception by humans or machines, three senses are better than one. In: Proc. ICSLP, Philadelphia 3, pp. 1477–1480.

Beskow, J., 1995. Rule-based visual speech synthesis, Proc. EUROSPEECH, Madrid, Spain 1, pp. 299–302.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. ASSP 27 (2), 112–113.

Brand, M., 1999. Voice puppetry, Proc. SIGGRAPH, LA, California, pp. 21–28.

Bregler, C., Omohundro, S., 1995. Nonlinear image interpolation using manifold learning. In: Tesauro, G., Touretzky, D., Leen, T. (Eds.), Advances in Neural Information Processing Systems 7, MIT press, Cambridge, pp. 401–408.

Bryll, R., Ma, X., Quek, F., 1999. Camera calibration utility description, VisLab Tech. Rep., University of Illinois at Chicago.

Caldognetto, E.M., Vagges, K., Borghese, N.A., Ferrigno, G., 1989. Automatic analysis of lip and jaw kinematics in VCV sequences. In: Proc. of EUROSPEECH, Paris 2, pp. 453–456.

Cohen, M., Massaro, D.W., 1993. Modeling coarticulation in synthetic visual speech. In: Thalmann, N.M., Thalmann, D. (Eds.), Models and Techniques in Computer Animation. Springer, pp. 141–155.

Coianiz, T., Torresani, L., Caprile, L., 1995. 2D deformable models for visual speech analysis. In: Stork, D., Hennecke, M. (Eds.), Speech Reading by Man and Machine. Springer, pp. 391–398.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley, New York.

Duttweiler, D., Messerschmitt, D., 1976. Nearly instantaneous companding for nonuniformly quantized PCM. In: IEEE Trans. on Comm., COM-24, pp. 864–873.

Essa, I., 1995. Analysis, interpretation, and synthesis of facial expression, Ph.D. thesis, MIT Media Arts and Sciences, Cambridge, MA.

Ezzat, T., Poggio, T., 2000. Visual speech synthesis by morphing visemes. J. Comput. Vis. 38 (1), 45–57.

Finn, K., 1986. An investigation of visible lip information to be used in automatic speech recognition, Ph.D. dissertation, Dept. CS, Georgetown University, Washington, DC.

Fu, S., 2002. Visual Mapping Based on Hidden Markov Models, MS Thesis, Dept. of Computer Science and Engineering, Wright State University.

Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P.K., Garcia, O.N., 2005. Audio/Visual Mapping with Cross-Modal Hidden Markov Models. IEEE Transactions on Multimedia 7, No. 2, April.

Garofolo, J., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pellet, D.S., Dahlgren, N.L, 1988. The DARPA TIMIT CDROM. Available from LDC: <http://www.ldc.upenn.edu/cgibin/aesl/aesl>.

Goldschen, A.J., 1993. Continuous automatic speech recognition by lipreading, Ph.D. thesis, George Washington University.

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., 2003. Temporal properties of spontaneous speech—a syllable-centric perspective. J. Phonet. 31 (3–4), 465–485.

Gutierrez-Osuna, R., Kakumanu, P., Esposito, A., Garcia, O.N., Bojorquez, A., Castillo, J., Rudomin, I., 2002. WSU Technical.Report CS-WSU-02-03, Dayton, OH.

Gutierrez-Osuna, R., Kakumanu, P.K., Esposito, A., Garcia, O.N., Bojorquez, A., Castillo, J.L., Rudomin, I.J., 2005. Speech-driven Facial Animation with Realistic Dynamics. IEEE Trans. Multimedia 7 (1), 33–42.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. SAP 2 (4), 578–589.

Hong, P., Wen, Z., Huang, T.S., 2002. Real-time speech-driven face animation with expressions using neural networks. IEEE Trans. Neural Networks 13 (4), 916–927.

Itakura, F., 1975. Line spectrum representation of linear prediction coefficients of speech signal, JASA57, pp. 535 (abstract).

Jourlin, P., Luettin, J., Genoud, D., Wassner, H., 1997. Acoustic-labial speaker verification. Patt. Rec. Lett. 18, 853–858.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: active contour models. Int. J. Comput. Vis. 1 (4), 321–331.

Kim, H., Lee, H., 1999. Interlacing properties of line spectrum pair frequencies. IEEE Trans. SAP 7, 87–91.

Klatt, D.H., 1982. Prediction of perceived phonetic distance from critical band spectra: a first step. In: Proc. of ICASSP, Paris, pp. 1278–1281.

Kühnert, B., Nolan, F., 1999. The origin of coarticulation, in Coarticulation: theory, data, and techniques. In: Harcastle, W., Helwett, N. (Eds.), Cambridge University Press, pp. 7–29.

Lavagetto, F., 1995. Converting speech into lip movements: a multimedia telephone for hard of hearing people. IEEE Trans. Rehab. Eng. 3 (1), 90–102.

Lee, Y., Terzopoulos, D., Waters, K., 1995. Realisitc modeling for facial animation. In: Proc. of SIGGRAPH, LA, California, pp. 55–62.

Leps¢y, S., Curinga, S., 1998. Conversion of articulatory parameters into active shape model coefficients for lip motion representation and synthesis. Signal Process. Image Commun. 13, 209–225.

Luttin, J., Thacher, N.A., Beet, S.W., 1996. Active shape models for visual speech feature extraction. In: Stork, D., Hennecke, M. (Eds.), Speech-Reading by Man and Machine, vol. 150. Springer, pp. 383–390.

Markel, J., Gray, A., 1976. Linear Prediction of Speech. Springer.

Massaro, D.W., 1997. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press.

Massaro, D.W., Beskow, J., Cohen, M.M., Fry, C.L., Rodriquez, T., 1999. Picture my voice: audio to visual speech synthesis using Artificial Neural Networks. In: Proc. AVSP, Santa Cruz, CA, pp. 133–138.

McAllister, D.F., Rodman, R.D., Bitzer, D.K., Freeman, A.S., 1998. Speaker independence in automated lip-sync for audio–video communication. Comput. Networks ISDN Syst. 30, 1975–1980.

Montgomery, A., Jackson, P., 1983. Physical characteristics of lips underlying vowel lipreading performances. JASA 73 (6), 2134–2144.

Morishima, S., Harashima, H., 1991. A Media conversion from speech to facial image for man–machine interface. IEEE J. Selected Areas Commun. 9 (4), 594–600.

Nakamura, S., Yamamoto, E., 2001. Speech-to-lip movements synthesis by maximizing audio–visual joint probability. J. VLSI Signal Proc. 27, 119–126.

Parke, F.I., 1982. Parameterized models for facial animation. IEEE Comput. Graph. Appl. 2 (9), 61–68.

Parsons, T.W., 1986. Voice and Speech Processing. McGraw Hill (Chapter 3).

Pelachaud, C., Badler, N.I., Steedman, M., 1996. Generating facial expressions for speech. Cognit. Sci. 20, 1–46.

Picone, J.W., 1993. Signal modeling techniques in speech recognition. Proc. IEEE 81 (9), 1215–1247.

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice-Hall.

Rogozan, A., Deléglise, P., 1998. Adaptive fusion of acoustic and visual sources for automatic speech recognition. Speech Commun. 26, 149–161.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. J. Phonetics 16 (1), 55–76.

Sharma, S., Vermeulen, P., Hermansky, H., 1998. Combining information from multiple classifiers to speaker verification. In: Proc. RL2C, France, pp. 115–119.

Soong, F.K., Juang, B.H., 1993. Optimal quantization of line LSP parameters. IEEE Trans. SAP 1, 15–24.

Summerfield, Q., 1979. Use of visual information for phonetic perception. Phonetics 36, 314–331.

Tekalp, A.M., Ostermann, J., 2000. Face and2-D mesh animation in MPEG-4. In: Sig. Processing: Image Comm. 15, pp. 387–421.

Tibrewala, S., Hermansky, H., 1997. Sub-band based recognition of noisy speech. In: Proc. of ICASSP, Munich, Germany, pp. 1255–1258.

Tsai, R.Y., 1987. A versatile camera calibration technique for high-accuracy 3D machine vision metrology. IEEE J. Robot. Automat. 3, 323–344.

Waters, K., Frisbie, J., 1995. A coordinated muscle model for speech animation. In: Proc. of Graphics Interface, Ontario, pp. 163–170.

Waters, K., Levergood, T., 1993. DECface: an automatic lip synchronization algorithm for synthetic faces, RLE, Cambridge, MA Tech. Rep. CRL 93/4.

Wilson, D.R., Martinez, T.R., 2000. Reduction techniques for exemplar-based learning algorithms. Mach. Learning 38 (3), 257–286.

Yamamoto, E., Nakamura, S., Shikano, K., 1998. Lip movement synthesis from speech based on Hidden Markov models. Speech Commun. 28, 105–115.