Speech Driven Facial Animation

P. Kakumanu R. Gutierrez-Osuna[1] A. Esposito R. Bryll A. Goshtasby O. N. Garcia[2]

Department of Computer Science and Engineering
Wright State University
3640 Colonel Glenn Hwy
Dayton, OH 45435-0001

ABSTRACT

The results reported in this article are an integral part of a larger project aimed at achieving perceptually realistic animations, including the individualized nuances, of three-dimensional human faces driven by speech. The audiovisual system that has been developed for learning the spatio-temporal relationship between speech acoustics and facial animation is described, including video and speech processing, pattern analysis, and MPEG-4 compliant facial animation for a given speaker. In particular, we propose a perceptual transformation of the speech spectral envelope, which is shown to capture the dynamics of articulatory movements. An efficient nearest-neighbor algorithm is used to predict novel articulatory trajectories from the speech dynamics. The results are very promising and suggest a new way to approach the modeling of synthetic lip motion of a given speaker driven by his/her speech. This would also provide clues toward a more general cross-speaker realistic animation.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces And Presentation**]: User Interfaces—*Voice I/O*; I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation*; I.4.8 [**Image Processing And Computer Vision**]: Scene Analysis—*Stereo, Time-varying imagery, Tracking*; I.5 [**Pattern Recognition**]: Applications—*Computer Vision, Signal Processing*.

General Terms

Algorithms, Measurement, Experimentation, Performance, Human Factors

Keywords

Facial animation, lip-syncing, speech processing, computer vision, MPEG-4

PUI'01, November 15-16, 2001, Orlando, FL.

1 INTRODUCTION

Computer simulation of human faces capable of reflecting mouth movements and emotional states has been a flourishing research area for a long time. A number of papers have been published, resulting in a large number of facial models and several animation systems [1, 2, 3, 4, 5, 6, 7]. The interest in this technology has been clearly shown by the inclusion of animated face features and animated 2D meshes using the MPEG-4 standard [8]. Moreover, the usefulness of this technology has been proven by a set of perceptual experiments which showed that facial animation can provide practical and subjective benefits in human-computer interaction, such as cues in understanding noisy spoken text, and positive feelings during waiting times [9, 10].

Despite all this research, current facial models and animation techniques are still inaccurate approximations of natural faces, due to the fact that the dynamics of human facial expressions are not yet well understood, and often not even captured. As a result, the development of appropriate paradigms and tools to animate synthetic faces remains a challenging task. In this general context, realistic synthesis of lip motion is critical to differentiate "talking heads" from facial animations designed for other purposes (e.g., to display the dynamics of emotional expressions). Lip motion is also essential to understand spoken language in noisy environments, to communicate well with hearing-impaired people, and to improve the performance of current speech recognition systems, if viewed as integral part of the speech-recognition, production and communication effort.

As a step towards accomplishing these goals, we propose in this paper a method for learning speech-based lip motions from video. The method operates at a sub-phonetic level, mapping speech acoustic features (e.g. spectral power, F0) onto orofacial movements, thus bypassing problems with phonetic recognition and the many-to-many relationships between phonemes and visemes. Using a tapped delay line to account for coarticulation, we propose a nearest-neighbor approach that compares favorably in the chosen problem with computationally intensive Time-Delay Neural Networks or Hidden Markov Models. This method is employed to drive an MPEG-4 compliant facial animation from a new speech signal using a small audio-visual database from the same speaker. The sample size in this work was sufficiently small that no clustering was necessary in the search space.

[1] To whom correspondence should be addressed: rgutier@cs.wright.edu, (+937) 775 5120.
[2] Principal Investigator.

2 RELATED WORK

Both text and speech have been used as control inputs for animating human faces. In text-driven facial animation, the process generally involves determining a mapping from text (orthographic or phonetic) onto visemes by means of vector quantization [11, 12] or a rule-based system [13, 14]. Facial animation driven by speech can be approached in a similar fashion by deriving the phoneme sequence directly from the speech signal, as done in speech recognition [15, 16, 17]. However, mapping from acoustic frames to phonetic units and then to animation frames involves introducing additional uncertainty that often slights prosodic variables (e.g. pitch, duration or amplitude). It is more appealing, therefore, to construct a direct mapping from speech acoustics (e.g. linear predictive codes) onto facial trajectories, either through control parameters of the facial animation itself [1], three-dimensional coordinates of facial points [18] or articulatory parameters [19].

Frame-to-frame mappings (e.g. regression) based on the spatial, rather than temporal, relationship between speech acoustics and facial animation units are not very accurate since speech can only account for approximately 65% of the facial variance [20]. Several methods have been proposed to capture this complex temporal structure, either by means of tapped delay lines (e.g. a Time-Delay Multilayer Perceptron) or time-series analysis (e.g. a Hidden Markov Model). Lavagetto [19] synthesized mouth articulatory parameters from power-normalized LPC coefficients using Time Delay Multilayer Perceptrons. Brand [18] predicted the trajectories of 3D points in a face from LPC and RASTA-PLP acoustic features using an entropy-minimization algorithm that learned both the structure and the parameters of a Hidden Markov Model. These methods, however, rely on iterative estimation of non-linear models that result in resource consuming and computationally intensive training phases.

In the light of these considerations, our efforts to model lip motion directly from speech acoustics attempt to overcome the drawbacks of phonetic-level methods while providing a computationally effective simpler model based on nearest-neighbors. Only a limited set of acoustic features, carefully selected on the basis of perceptual considerations, have been used in our approach. To model the coarticulatory effects, context is taken into account explicitly. Prosodic elements such as fundamental frequency and energy of the signal are also considered. These preliminary results are very promising and suggest a new way to approach the modeling of synthetic lip and associated facial motions.

3 SYSTEM OVERVIEW

The vision system employed in this research consists of two color cameras (Kodak ES310), and two dedicated personal computers and frame grabbers capable of acquiring 648x484 video to a hard drive at 60 frames per second to capture even very short phonetic phenomena. Speech signals are 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-second of audio between video frames to ensure synchrony. Once the data has been saved to disk, the three-dimensional coordinates of various facial points are tracked using stereo correspondence in color, and a number of acoustic features are extracted from the audio track. These procedures are described in detail in the following subsections.

3.1 Audio Processing

The unprocessed speech waveform contains limited information that can be directly used for the prediction of facial movements. This is mainly due to the fact that speech signals result from a mixture of several sources, including speech content (e.g. phonemes), speaker dependencies and prosody (e.g. speed and intonation). Several analyses in the frequency domain that allow, to some extent, separation of speech content from irrelevant information have been proposed, such as Filter Bank Analysis, Smoothed Spectrum, Cepstral Analysis, Linear Prediction Coding (LPC) [21], Perceptual Linear Prediction (PLP) and Relative Spectrum (RASTA-PLP) [22, 23]. The parameters of these algorithms generally use default values, which are the same for all the phonemes without taking into account the different nature of such segments. This results in articulatory attributes whose robustness depends on the extracted preprocessing parameters. Moreover, all the aforementioned algorithms are based on the assumption that formants and their trajectories are the cues for speech perception. Therefore, they extract features that are based on the unambiguous identification of the 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 shown that neither the overall level nor the exact location and tilt of the patterns in the frequency domain but the changes in the shapes and relative locations of the major spectral features appear to be important for phonetic perception [24]. In the light of these considerations, we decide to preprocess our data through a perceptual based analysis, which uses concepts from psychophysics of hearing in order to obtain an estimate of the auditory spectrum. A possible rationale for the use of perceptually-related transformations of the signal, such as the use of a logarithmic frequency scale, may be justified by 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 the learning process is perceived by the human learner. Therefore, the production of speech is *a fortiori* perceptually biased by the fact that speech is learned through the perceiving auditory mechanism.

The speech signal, sampled at 16 kHz, is processed in blocks of 1/60-second with no overlap between blocks. Each block is pre-emphasized with an FIR filter ($H(z) = 1 - az^{-1};\ a = 0.97$), weighted with a Hamming window to avoid spectral distortions [25] and then passed through a critical-band resolution (Bark-scale transformation) of the Fourier spectrum. This crude perceptually modified spectrum, shown in Figure 1, is used to capture the dynamic aspects of the spectral envelope patterns, resulting in a vector of Perceptual Critical Band Features (PCBF) [26]. Note that PLP and Rasta-PLP also contain a perceptual analysis component, but they further preprocess the speech signals by including a predictive modeling (PLP) and a temporal filtering process (RASTA-PLP) to make the extracted features closer to the formant trajectory. In this context, our PCBFs can be interpreted as the first stage of PLP or Rasta-PLP analysis. How our perceptual acoustic features encode information that relates to facial movements, as well as their relative performance compared to other acoustic features (e.g. filter banks, linear prediction coding, line spectrum pairs) will be the subject of another report [in preparation].

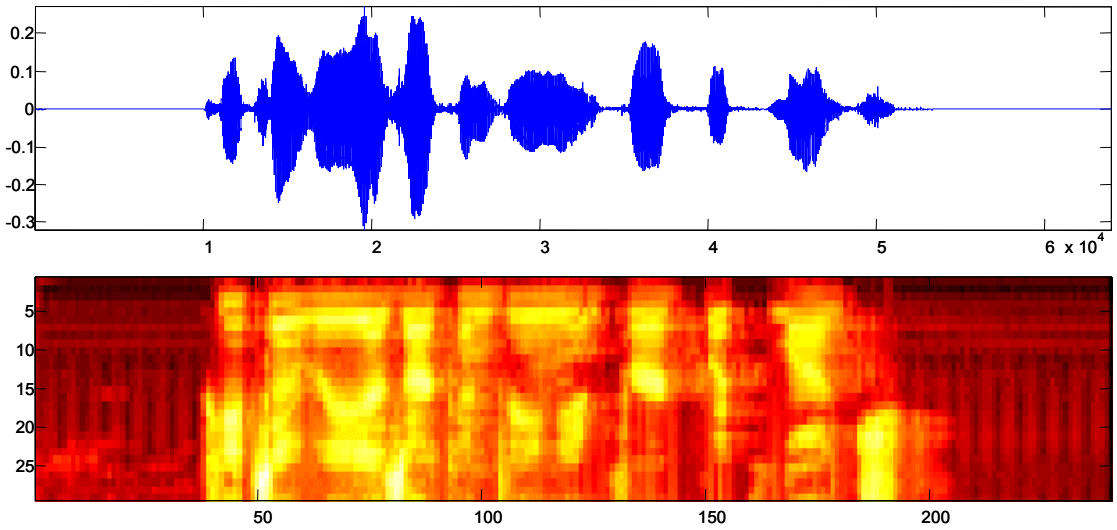


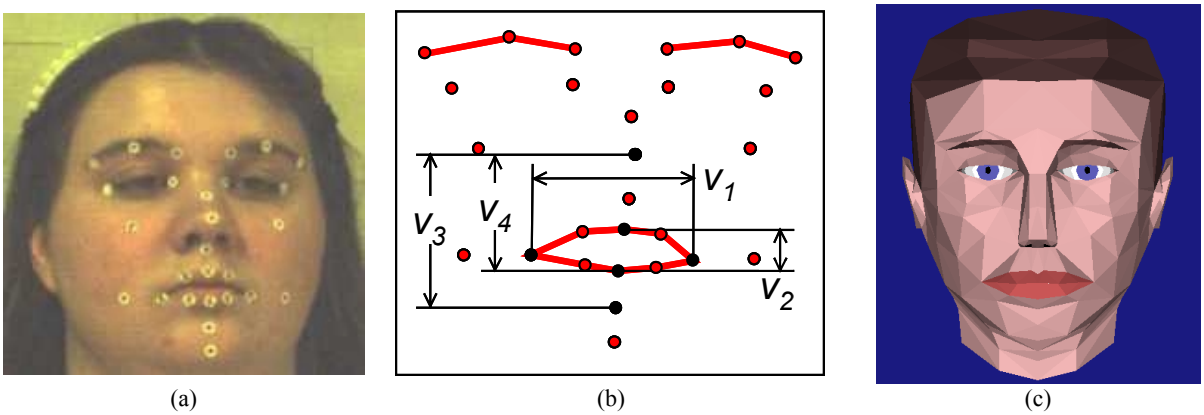
Figure 1. Audio waveform and Perceptual Critical Band Features for the sentence "Pizzerias are convenient for a quick lunch"

Furthermore, we used a narrow band analysis to identify tonal components in the spectral representations. This analysis serves as basis for deciding whether the waveform is periodic, aperiodic, or a mixture of the two and to derive the fundamental frequency, since changes in the intonation and the speaking rate also affect facial movements. A total of 29 PCBFs are extracted from the speech signal which, along with the F0 and frame energy, form a 31-dimensional vector of acoustic inputs to our animation system.

3.2 Video Processing

To facilitate accurate 3D tracking of the facial dynamics using a stereo camera pair, 27 markers were placed in the face of the subject at various MPEG-4 feature points [27], as shown in Figure 2(a). The cameras were calibrated using Tsai's stereo calibration algorithm that takes radial lens distortion into consideration [28]. We applied a standard calibration procedure comprising a prism with fiduciary markers of known configuration and a calibration tool to establish the calibration correspondences and calibration matrix [29, 30]. The position of each marker was independently tracked by finding the maximum cross-correlation in a local region of interest centered on the marker's position in the preceding frame. The initial position of each marker was manually entered by means of a graphical user interface. This process was performed independently on each of the two stereo sequences.

In the present work four articulatory lip descriptors, v_1 through v_4, were extracted from the 3D coordinates of the markers. Shown in Figure 2(b), these articulatory parameters are the horizontal and vertical distances between the lip corners, as well as the vertical distance from the nose to the chin and to the lower lip.



(a) (b) (c)

Figure 2. Neutral face of the subject with visual markers (a), reconstructed marker coordinates and articulatory parameters (b), and Facial Animation Engine (c).

3.3 Mapping Acoustic Dynamics to Visual Articulators

As a result of sub-phonemic dynamics as well as coarticulation effects, it is not feasible to perform a frame-to-frame prediction of video configurations *v(k)* from the corresponding acoustic features *a(k)*, where *k* denotes the frame index. For this reason, we explicitly encode context by associating each video frame *v(k)* with the acoustic features from past and future audio frames $a_n(k) = [a(k-n), \cdots a(k), \cdots a(k+n)]$, forming a lookup table of audio-visual dynamic pairs $[a_n, v](k)$ from training data. This data can be used to estimate a tapped delay line model (e.g. a time-delay neural network). In this work, however, a simple nearest-neighbor procedure is employed. Given a new audio window $\hat{a}_n$, we find the video configuration $\hat{v}$ that corresponds to the closest (Euclidean distance) audio trajectory $a_n(k)$ in the training set:

$$\hat{v} = \left\{ v(k);\ k = \arg\min_j \left\| \hat{a}_n - a_n(j) \right\| \right\}$$

We use a value of *n=5*, which corresponds to 83-millisecond windows into the past and the future. To reduce dimensionality, the acoustic dynamics are sub-sampled 5:1, resulting in a 93-dimensional acoustic vector $a_5(k) = [a(k-5), a(k), a(k+5)]$. Our experiments indicate that using a longer window or a finer sub-sampling does not improve performance significantly.

3.4 Animation

To verify the accuracy of our trackers and the resulting predictions, we employed the Facial Animation Engine (FAE) developed at the University of Genova [27]. The FAE is a high level interface capable of animating MPEG-4 compliant faces at high frame rates in synchrony with an audio track. The FAE has a straightforward interface that requires a facial model ('Mike' in our shareware version, see Figure 2(c)), an ASCII file with the Facial Animation Parameters (FAP) for each frame, and an accompanying *.wav file with the audio track. In our interface, MPEG-4 FAPs were generated either from the 3D coordinates of the markers (for debugging purposes) or from the mouth articulatory parameters.

4 EXPERIMENTAL RESULTS

A dataset consisting of five sentences from TIMIT [31] was collected on a female speaker. This group of sentences was repeated in random order five times, for a total of 25 sentences. In order to minimize head motions, the subject was requested to use a small head-rest that did not constrain articulatory motion. To further reduce intra-speaker variations, each sentence and repetition was recorded starting from a neutral facial expression, such as that of Figure 2(a), which served as a baseline. The video parameters of this neutral face were subtracted from the subsequent frames. Four out of five repetitions of each sentence were used to form the training lookup table, and the remaining repetition was used for testing purposes.

Table 1. Dataset of five TIMIT sentences used in this study

Sentences
Pizzerias are convenient for a quick lunch
The bungalow was pleasantly situated near the shore
The clumsy customer spilled some expensive perfume
Too much curiosity can get you into trouble
December and January are nice months to spend in Miami

The prediction results for the four articulatory parameters on the five test sentences are shown in Figure 3, where the predicted trajectories (thin line) have been mean-filtered with a 50-millisecond window to reduce jitter. The reported data demonstrates the good performance of the proposed method and the reliability of the suggested acoustic features for predicting lip movements. Table 2 shows the average prediction results on the five test sentences in terms of the normalized mean-squared error ε and correlation coefficient ρ between the predictions and the true trajectories, defined by:

$$\varepsilon = \frac{1}{\sigma_v^2} \frac{1}{N} \sum_{k=1}^{N} \left(\hat{v}(k) - v(k) \right)^2$$

$$\rho = \frac{1}{N} \sum_{k=1}^{N} \frac{\left(\hat{v}(k) - \mu_{\hat{v}} \right)\left(v(k) - \mu_v \right)}{\sigma_{\hat{v}} \sigma_v}$$

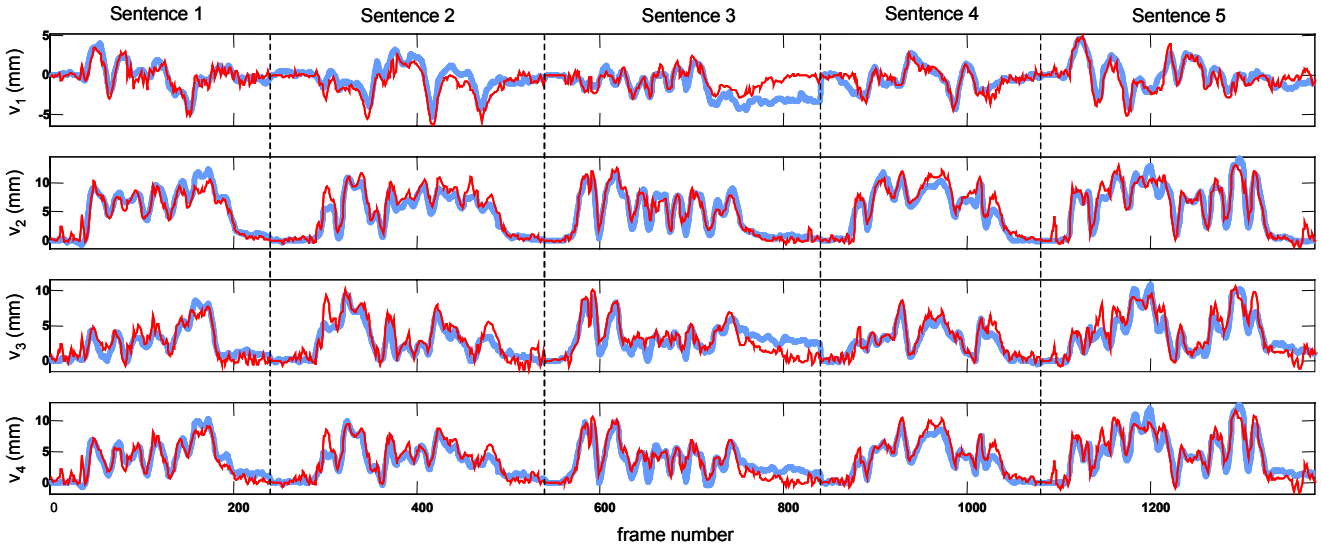


Figure 3. Predicted (thin line) versus actual (thick) trajectories for the four articulatory parameters on the five test sentences

where $v(k)$ is the true articulatory parameter at frame k, $\hat{v}(k)$ is its nearest-neighbor prediction, N is the total number of frames in the dataset, and μ and σ are their sample mean and standard deviation, respectively. For comparison purposes, Table 2 reports the prediction results that were obtained using the coefficients of an LPC analysis. The PCBFs clearly outperform LPC, providing significantly higher correlation coefficients and lower prediction errors on the four articulatory parameters.

Table 2. Normalized MSE (ε) and correlation coefficient (ρ) of the predictions for LPC and PCBF

Articulatory parameter	LPC		4.4.0.0.1 PCBF	
	ε	ρ	ε	ρ
v_1	0.6360	0.6246	0.3929	0.8031
v_2	0.2738	0.8551	0.0995	0.9529
v_3	0.4203	0.7954	0.2525	0.8993
v_4	0.4045	0.7888	0.1701	0.9255

5 CONCLUSIONS

In the present work we have proposed a new approach to model the lip movement portion of facial animation directly from speech. The method relies on a small set of acoustic features, which has been proven to capture the dynamics of articulatory movements. The mapping from acoustics to visual articulatory parameters has been performed using a simple nearest-neighbor algorithm. Our method uses a tapped delay line to exploit contextual information and take co-articulation effects into consideration. It also uses F0 and energy as prosodic cues of speaking rate and intonation. Therefore, given a newly spoken sentence, the system accurately generates the trajectories of the lip articulatory parameters, preserving the realism of the movements. It is expected that the nearest-neighbor procedure will become impractical (in terms of search time and storage requirements) as the size of the audio-visual database increases. For this reason, we are currently investigating vector-quantization procedures to accommodate for an audio-visual dataset with a large number of TIMIT sentences. This AV dataset, being collected at the time of this writing, will also allow us to evaluate the performance of our approach when predicting phonetic sequences not included in the training set.

6 ACKNOWLEDGEMENTS

Nishant Balan and Jessica Garringer are acknowledged for their devoted effort in data collection and video tracking. We would also like to thank Roberto Pockaj (University of Genova) for providing the Facial Animation Engine and for many helpful e-mail discussions. This research was supported by awards from NSF/CISE 9906340, NSF/CAREER 9984426 and WSU/OBR 00-509-10.

7 REFERENCES

[1] M. Cohen and D. Massaro, 1993, "Modeling coarticulation in synthetic visual speech," in N.M. Thalmann and D. Thalmann, editors, Models and Techniques in Computer Animation, pp. 141-155. Springer Verlag, Tokyo.

[2] F. I. Parke, "Parameterized models for facial animation" IEEE Computer Graphics and Applications, vol. 2, no. 9, pp. 61-68, November 1982.

[3] P. Kalra, 1993, "An interactive multimodal facial interaction", Ph.D Dissertation No. 1183, Ecole polytechnique fédérale de Lausanne, Switzerland.

[4] J. Fischl, B. Miller and J. Robinson, 1993, "Parameter tracking in muscle-based analysis-synthesis system," in Proceedings of Picture Coding Symposium (PCS93), Lausanne, Switzerland.

[5] J. Ostermann and E. Haratsch, 1997, "An animation definition interface – rapid design of MPEG-4 compliant animated faces and bodies," in Proceedings of the International Workshop on Synthetic-Natural Hybrid Coding and 3D Imaging, Rhodes, Greece, September 5-9 1997.

[6] E. Cosatto and H. P. Graf, 1998, "Sample-based synthesis of photo-realistic talking heads," in Computer Animation, pp. 103-110, Philadelphia, Pennsylvania, June 8-10,1998.

[7] J. Ostermann, 1998, "Animation of synthetic face in MPEG-4," in proceedings of Computer Animation, Philadelphia, PA.

[8] A. M. Tekalp and J. Ostermann, 2000, "Face and 2-D mesh animation in MPEG-4", in Signal Processing: Image Communication 15, pp. 387-421.

[9] I. S. Pandzic, J. Ostermann and D. Millen, 1999, "User evaluation: synthetic talking faces for interactive services", Visual Computer 15, pp. 330-340.

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

[11] S. Morishima and H. Harashima, 1991, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface," in IEEE Journal on Selected Areas in Communications 9(4), 594-600.

[12] K. Waters and T. M. Levergood, 1993, "DECface: an automatic lip synchronization algorithm for synthetic faces," Technical Report CRL 93/4, DEC Cambridge Research Laboratory, Cambridge, MA.

[13] C. Pelachaud, N. I. Badler and M. Steedman, 1996, "Generating facial Expressions for Speech," in Cognitive Science 20, pp. 1-46.

[14] J. Beskow, 1995, "Rule-based visual speech synthesis," in ESCA EUROSPEECH '95, 4th European Conference on Speech Communication and Technology, Madrid, Spain.

[15] L. M. Arslan and D. Talkin, 1999, "Codebook Based Face Point Trajectory Synthesis Algorithm using Speech Input," in Speech Communication 27, pp. 81-93.

[16] T. Ohman, 1998, "An audio-visual speech database and automatic measurements of visual speech," in Quarterly Progress and Status Report, Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden, Stockholm, Sweden.

[17] E. Yamamoto, S. Nakamura and K. Shikano, 1998, "Lip movement synthesis from speech based on Hidden Markov models," in Speech Communication 28, pp. 105-115.

[18] M. Brand, 1999, "Voice Puppetry," in Proceedings of SIGGRAPH'99 Computer Graphics, Annual Conference Series, pp. 21-28.

[19] F. Lavagetto, 1995, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," in IEEE Transactions on Rehabilitation Engineering 3(1), pp. 90-102.

[20] H. Yehia, P. Rubin and E. Vatikiotis-Bateson, 1998, "Quantitative Association of Vocal-tract and Facial Behavior," in Speech Communications 26, pp. 23-43.

[21] L.R. Rabiner and B.H. Juang, 1993, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[22] H. Hermansky, 1990, "Perceptual linear predictive (PLP) analysis of speech," in Journal of Acoustic Society of America, vol. 87(4), pp. 1738-1792.

[23] H. Hermansky and N. Morgan, 1994, "RASTA Processing of Speech," in IEEE Transactions on Speech and Audio Processing 2(4), pp.578-589.

[24] D. H. Klatt, 1982, "Prediction of perceived phonetic distance from critical band spectra: a first step," in Proceedings of the International Congress on Acoustics, Speech, Signal Processing, Paris, IEEE Press, pp.1278-1281.

[25] L. R. Rabiner and R. W. Schafer, 1978, Digital processing of speech signals, Prentice-hall, 1978.

[26] G. Aversano, A. Esposito and M. Marinaro, 2001, "A new text-independent method for phoneme segmentation," to appear in the Proceedings of IEEE Midwest Symposium on Circuits and Systems, Dayton 14-17 August 2001.

[27] F. Lavagetto and R. Pockaj, 1999, "The Facial Animation Engine: towards a high-level interface for the design of MPEG-4 compliant animated faces", in IEEE Transactions on Circuits and Systems for Video Technology 9(2), pp.277-289.

[28] R. Y. Tsai, 1987, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," in IEEE Journal of Robotics and Automation 3, pp. 323-344.

[29] R. Bryll, X. Ma and F. Quek, 1999, "Camera Calibration Utility Description," Technical Report VISLab-01-15, Vision Interfaces and Systems Laboratory, Wright State University. http://vislab.cs.wright.edu/Publications/technical-reports/BryMQ01.html

[30] F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K. McCullough, N. Furuyama and R. Ansari, 2000, "Gesture, Speech and Gaze Cues for Discourse Segmentation," in Proceedings of CVPR 2000, Hilton Head Island, South Carolina, June 13-15, 2000.

[31] J. Garofolo et al, 1998, DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database, National Institute of Standards and Technology, Gaithersburg, MD.