# Lip Synchronization Using Linear Predictive Analysis

Sumedha Kshirsagar, Nadia Magnenat-Thalmann
MIRALAB, CUI, University of Geneva
24 rue de General Dufour
CH-1211 Geneve, SWITZERLAND
+41-22-7057769
{sumedha.kshirsagar,nadia.thalmann}@cui.unige.ch

## ABSTRACT

Linear Predictive analysis is a widely used technique for speech analysis and encoding. In this paper, we discuss the issues involved in its application to phoneme extraction and lip synchronization. The LP analysis results in a set of *reflection coefficients* that are closely related to the vocal tract shape. Since the vocal tract shape can be correlated with the phoneme being spoken, LP analysis can be directly applied to phoneme extraction. We use neural networks to train and classify the reflection coefficients into a set of vowels. In addition, average energy is used to take care of vowel-vowel and vowel-consonant transitions, whereas the zero crossing information is used to detect the presence of fricatives. We directly apply the extracted phoneme information to our synthetic 3D face model. The proposed method is fast, easy to implement, and adequate for real time speech animation. As the method does not rely on language structure or speech recognition, it is language independent. Moreover, the method is speaker independent. It can be applied to lip synchronization for entertainment applications and *avatar* animation in virtual environments.

## Keywords

LP analysis, lip synchronization, real-time speech animation

## 1. INTRODUCTION

In today's multi-modal user interactive systems, talking heads form an important and essential part. For virtual presenters, storytellers, and *avatars* in virtual environments, synthetic faces talking in natural voice are gaining more potential. The advances in speech synthesis technologies are resulting in better quality computer generated voices. Nevertheless, using natural voice for the animation of synthetic faces remains a challenging area of research in computer animation. The problem can be easily divided into two parts; *viz.* extracting the mouth shape information from speech signal and then applying it to a synthetic 3D face model with synchronization for realistic animation. We concentrate on the former part in this paper, and briefly discuss the issues involved in the later.

The goal is to extract the parameters from speech signal which are directly or indirectly related to the mouth/lip movements. McAllister *et al* [1] used mouth shape descriptors called *moments* computed from the FFT coefficients of the speech signal as these parameters. LPC derived cepstral coefficients were used by Curinga, Lavagetto, and Vignoli [2]. They trained *Time Delay Neural Network* to take care of the co-articulation. Yamamoto, Nakamura, Shikano [3] and Tamura *et al* [4] used HMM techniques for the synthesis of lip movements from the speech signal. Morishima [5] described a real time voice driven talking head and its application to entertainment. He used LPC derived cepstral coefficients to extract mouth shape parameters using neural networks.

Most of the above mentioned researchers used the mouth shape parameters like width, height, lip-to-lip distance or the control point locations around the lips. These parameters were extracted from the video sequences associated with the speech recordings, and were then used for training. We propose to take a different approach that will allow us to apply the results to any generalized 3D head. The straightforward alternative is to extract the phonemes or *visemes* (visual counterparts of phonemes) directly from the speech signal. We choose LP analysis to extract the parameters from the speech signal. This technique, as explained in the subsequent sections, is inadequate for the consonants. Thus, it has limited use for the accurate phoneme extraction. We partly overcome the limitation by augmenting the results of vowel recognition with the energy envelope modulation. We also use zero crossing rate to recognize unvoiced fricatives. Our face model can directly animate such modulated vowels, making the animation process easy.

The next section explains the proposed system in brief. Section 3 explains the use of the LP analysis and use of neural networks, energy criterion and zero crossing rate for phoneme extraction. The issues related to 3D synthetic faces for speech animation are discussed in section 4. Finally, we give conclusions and discuss future work.

## 2. SYSTEM OVERVIEW

Figure 1 shows the overall block diagram of the system. Input speech is sampled at 10 kHz with a frame size of 20 ms. Preprocessing includes pre-emphasis and hamming windowing of the signal. Currently, no filtering is done for noise reduction. 12 reflection coefficients are calculated as a result of LP analysis. The coefficients are obtained from sustained vowel data and are used to train the neural

network. As a result, one of the 5 chosen vowels (/a/, /e/, /i/, /o/, /u/) is obtained for the frame. We have chosen these vowels since we notice that the vowels in many languages can be roughly classified into these basic sounds or their combinations/variations. We use median filtering to smooth the resulting recognized vowels. The average energy of the signal is calculated as the zeroth auto-correlation coefficient over the frame and is used to decide the intensity of the detected vowel. Zero crossings are calculated to decide the presence of the unvoiced fricatives and affricates (/sh/, /ch/, /zh/ *etc.*). Finally, the Facial Animation Module generates the *Facial Animation Parameters* (FAP) as supported by MPEG-4 standard depending upon the phoneme input. Note that any 3D parameterized facial model can be used here. The speech animation is then easy using the pre-defined parameters for these extracted phonemes.

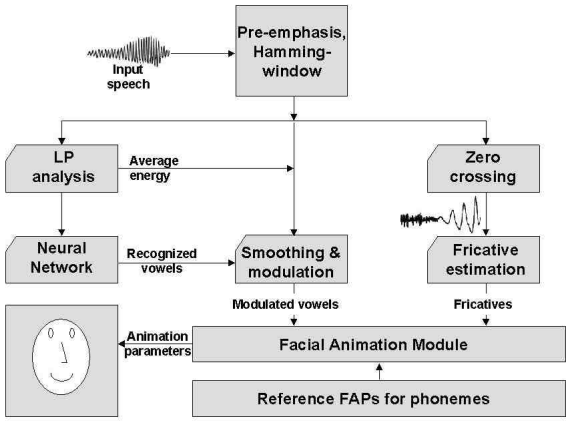


**Figure 1:** System overview

## 3. SPEECH ANLYSIS

As previously mentioned, several parameters can be used to correlate the speech signal to the mouth shape. Lewis and Parke [6] suggested the use of linear prediction for lip synchronization. However, they used Fourier transform of the negated zero extended LP coefficients for analysis. An analyzed speech frame was then classified using the *Euclidean distance norm* after comparing it with the spectra of reference phonemes. We use the LP derived *reflection coefficients*, the average energy in the speech signal and the zero crossing rate. This section explains the choice of these parameters in details.

### 3.1 Speech Production and LP Analysis

The human speech production system can be easily divided into the glottis or vocal cords and the vocal tract (mouth, tongue and lips). The glottal excitation acts as the source signal. The vocal tract, acting as a filter, then shapes it to generate the output speech. The phonemes can be characterized together by the excitation and the vocal tract shape. We concentrate on the vocal tract shape here. This subsection focuses on extraction of vowels, whereas issues involved in the extraction of some of the consonants are discussed in the subsequent subsections. For the production of vowels, the vocal tract shape is constant with time and uniform (without constrictions), with the sustained vibrations of the vocal cords.

Thus for the vowels, the vocal tract can be approximately modeled as a concatenation of a number of cylindrical tubes of uniform cross-sectional area [7]. Figure 2 shows a simple approximation of the model consisting of *m* acoustic tubes. The tubes have cross-sectional areas $A_1$ to $A_m$. Though these values have great variation from person to person, the relative distribution is similar for a given vowel. We are interested in extracting this vocal tract shape information, which will be directly useful for speech animation.

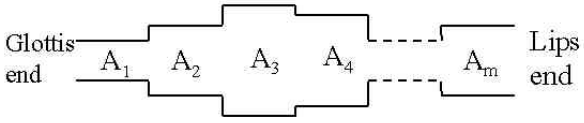


**Figure 2:** Schematic representation of vocal system and concatenated tube approximation

Wakita [8] compared the above acoustic filter model represented by the concatenated tubes, with the speech production model suggested by LP analysis. The details are beyond the scope of discussion here, but we state the result. The comparison between the acoustic tube model, and the LP derived model led to the following conclusion. The reflection coefficients $r_l$, computed as a by-product of the recursive LP algorithm, are directly related to the vocal tract area as per the concatenated tube model by the following equation.

$$r_l = \frac{A_{l-1} - A_l}{A_{l-1} + A_l}$$

As clearly depicted by the equation, the reflection coefficients are directly related to the variation of the vocal tract area for sustained vowels. A definite pattern observed in these coefficients for a particular vowel suggests the use of neural networks for classification.

### 3.2 Use of Neural Network

With the background given in the last subsection, the problem of recognizing the vowels reduces to a classification problem. A three-layer back-propagation neural network is widely used for a variety of pattern recognition and classification problems [9]. We use the same configuration to classify the reflection coefficients. There are 12 input nodes for the coefficients, 10 hidden nodes and 5 output nodes for the vowels. These parameters were tuned by running the training sessions several times on

the data and studying the classification result. We train the network in five repeated cycles, every time using the data in a different random order. We use reflection coefficients from sustained vowel data and also short vowel segments extracted from continuous speech. The speech data was recorded from 12 male and 5 female speakers. The following table shows the classification results on the test data set consisting of 4 male and 3 female speakers. The utterances were chosen from sustained vowels and the frames were chosen randomly. Note the mis-recognition between /e/ and /i/, and /o/ and /u/. The mouth shapes for these pairs of vowels are also similar.

| | | Recognized | | | | |
|---|---|---|---|---|---|---|
| | | /a/ | /e/ | /i/ | /o/ | /u/ |
| Expected | /a/ | 241 | 2 | 15 | 11 | 0 |
| | /e/ | 0 | 177 | 89 | 0 | 5 |
| | /i/ | 0 | 3 | 301 | 0 | 2 |
| | /o/ | 10 | 0 | 0 | 224 | 36 |
| | /u/ | 4 | 12 | 0 | 88 | 143 |

**Table 1:** Results of neural network classification

### 3.3 Energy Analysis

In the previous subsections, we have explained how the vowels can be extracted directly from the speech signal using LP analysis and neural networks. However, we are aware that the application of the vowels alone for speech animation is not sufficient. The vowel-to-vowel transition and the consonant information are missing, which are very important for realistic speech animation. The consonants are typically produced by creating a constriction at some place along the length of the vocal tract. During such constrictions/closures, the energy in the speech signal diminishes. Hence, we use the average energy in a speech frame to modulate the recognized vowel. It is calculated as the zeroth autocorrelation coefficient of the frame, and has already been computed during the LP analysis phase. Thus, the calculation of energy does not cause any additional computational overhead.

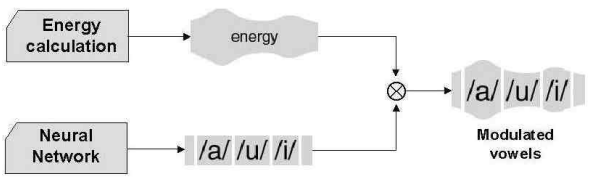


**Figure 3:** Modulating vowels with energy envelope

As an initialization process, we record background noise in the room to set the energy threshold for silence. Also, we record sustained vowel /a/ from the user asking her to say the utterance with maximum volume expected in normal speech. This enables us to compute the maximum energy. This value is used to get a normalized weighting factor for the vowels during normal speech.

As explained in Section 4, the parameterized face model (MPEG4 model in our case) enables us to animate these modulated vowels. The normalized weighting factor directly proportional to the energy in the speech signal is used to scale the parameters for the corresponding vowel. Figure 3 pictorially depicts the idea behind modulating vowels with energy envelope.

### 3.4 Zero Crossing

Using the energy content of the signal may result in false closure of mouth, especially in the case of affricates and unvoiced fricatives. For such cases, we can use the average zero crossing rate in the speech signal for each frame. The mean short time average zero crossing rate is 49 per 10 msec for unvoiced, and 14 per 10 msec for voiced speech [7]. This criterion is useful in making a distinction and is sufficient for our purpose. A short segment of the utterance /sh/ (as in sharp) shown in figure 4 highlights this criterion. In case of the presence of low energy in the speech frame, the zero crossing criterion decides the phoneme.

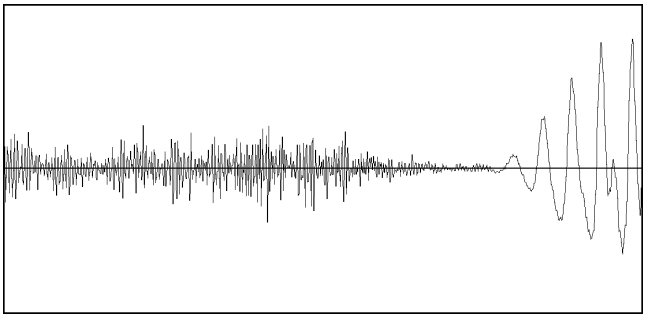

**Figure 4:** High zero crossing rate for fricative /sh/

## 4. FACIAL ANIMATION

So far, we have focused our attention on the extraction of phonemes directly from speech signal. In this section, we briefly consider the issues involved in the speech animation using this extracted information.

The phonemes extracted using the method described so far can be applied to any parameterized face model for speech animation. It is necessary to define the mouth shapes for the static phonemes in terms of these parameters. Depending upon the phoneme intensities, the corresponding parameter intensities can be set to achieve the speech animation. MPEG-4 standard provides such a way. The *Facial Animation Parameters* defined in the standard enable the user to define any facial expression in terms of these parameters with corresponding intensities. Moreover, since these parameters are normalized with respect to the distance

between certain key feature points on the face, the animation results are consistent when applied to any face model. For more detail discussion on the MPEG-4 standard and MPEG-4 compatible 3D faces, refer to [10][11].

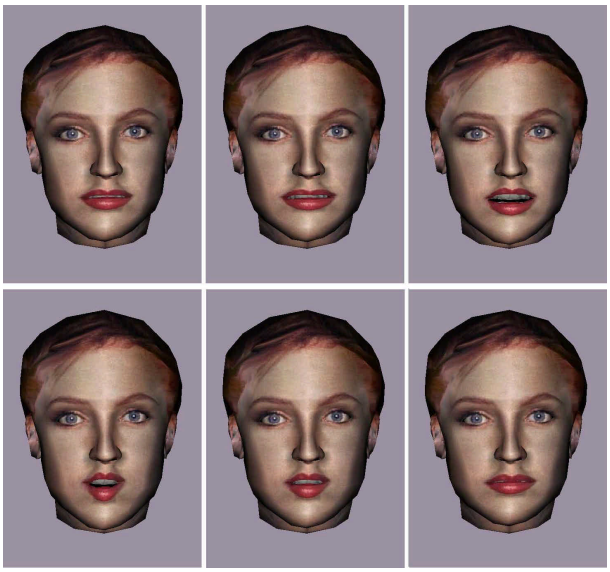

**Figure 5:** Frames of speech animation "hello"

In a networked virtual environment, an avatar of an individual can be truly represented by the individualized 3D face model and the voice of that individual using the lip synchronization method proposed here. An MPEG-4 compatible individualized 3D face of a person can be generated from two orthogonal photographs. We have used the 3D face models developed by [12]. We define the FAPs corresponding to the phonemes under consideration, and according to the modulation, the intensities of these FAPs are set to generate the facial animation in real time. Figure 5 shows successive frames for speech animation for the word "hello" which involves modulated vowels /a/, /e/, and /o/.

## 5. CONCLUSIONS AND FUTURE WORK

We have proposed a simple and fast method for realistic speech animation. As we are extracting higher level information (phonemes) directly from the speech signal, the results can be easily applied to any parameterized face model. We have used MPEG-4 compatible face model. The results of the speech animation using the recorded speech of different speakers can be seen at the following website: http://www.miralab.unige.ch/~sumedha/lipsynchronization. Note that the method is language as well as speaker independent. The score of the recognition given by the neural network can be used to combine two vowels generating a mouth shape that will represent the transition between them. We are aware that this method does not give accurate results as far as phoneme recognition is concerned. However, in the context of talking heads used for real time interactive system, the animation is satisfactory.

## 6. ACKNOWLEDGEMENTS

This work is supported by the EU ACTS VPARK project. We are thankful to the staff of the MIRALab for their valuable help in various matters.

## 7. REFERENCES

[1] D. V. McAllister, R. D. Rodman, D. L. Bitzer, A. S. Freeman, "Lip synchronization for Animation", *Proc. SIGGRAPH 97*, Los Angeles, CA, August 1997.

[2] Sergio Curinga, Fabio Lavagetto, Fabio Vignoli, "Lip movements synthesis using time delay neural networks", *Proc. EUSIPCO 96*, Sep. 1996.

[3] E. Yamamoto, S. Nakamura, K. Shikano, "Lip movement synthesis from speech based on Hidden Markov Models", *Speech Communication*, Elsevier Science, (26)1-2 (1998) pp. 105-115.

[4] M. Tamura, T. Masuko, T. Kobayashi, K. Tokuda, "Visual speech synthesis based in parameter generation from HMM : Speech driven and text-and-speech driven approaches", *Proc. AVSP 98*, International Conference on Auditory-Visual Speech Processing.

[5] S. Morishima, "Real-time talking head driven by voice and its application to communication and entertainment", *Proc. AVSP 98*, International Conference on Auditory-Visual Speech Processing.

[6] J. P. Lewis, F. I. Parke, "Automated lip-synch and speech synthesis for character animation", SIGGRAPH 1990, *Course Notes*, August 1990, pp.83-87.

[7] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signal*, Englewood Cliffs, New Jercy : Prentice Hall, 1978.

[8] Hisashi Wakita, "Direct estimation of the vocal tract shape by inverse filtering of the acoustic speech waveforms", *IEEE Trans. Audio & Electrocoustics*, Vol. 21, October 1973, pp. 417-427.

[9] J. A. Freeman, D. M. Skapura, "Neural Networks Algorithms, Applications, and Programming Techniques" Addison-Wesley, 1991.

[10] P. Doenges, F. Lavagetto, J. Ostermann, I. Pandzic, E. Petajan, "MPEG-4: Audio/Video and Synthetic Graphics/Audio for Mixed Media", Image Communications Journal, Vol.5, No.4, 1997, pp.433-463.

[11] M. Escher, I. Pandzic, N. M.-Thalmann, "Facial deformation from MPEG-4", *Proc. Computer Animation 98*, IEEE Computer Society, pp. 56-62.

[12] W. S. Lee, M. Escher, G. Sannier, N. M.-Thalmann, "MPEG-4 compatible faces from orthogonal photos", *Proc. Computer Animation 99*, IEEE Computer Society, pp. 186-194.