Real-time Face Animation Driven by Human Voice

Goranka Zorić
Department of Telecommunications
Faculty of Electrical Engineering and Computing, University of Zagreb
HR-10000, Croatia
goranka.zoric@tel.fer.hr

Abstract--**This paper presents a review of the recent research that examines the problem of generating full facial animation from auditory input speech signal. The major applications of this technique include face animation, human-computer interfaces, computer-aided instruction, video games and multimedia telephony for the hearing impaired. Using human voice for the real-time animation of synthetic faces remains a challenging area of research in computer animation. The key issue of this research is to find a mapping from audio information to visual information. A variety of approaches can be found in literature. We classify and summarize these approaches and propose a plan for an implementation based on a combination of these approaches.**

Keywords—Real-time speech animation, lip-sync

I. INTRODUCTION

Human speech is bimodal in nature [1]. Speech that is perceived by a person depends not only on acoustic cues, but also on visual cues such as lip movements or facial expressions. When noisy environments are encountered, visual component of speech can compensate for a possible loss in speech signal. This combination of auditory and visual speech recognition is more accurate than auditory only or visual only. Use of multiple sources of information (such as auditory information or visual information) generally enhances speech perception and understanding. Consequently, there has been a large amount of research on incorporating bimodality of speech into human-computer interaction interfaces. Real-time speech-driven face animation is one of the research topics in this area.

The goal is to animate the face of a speaking avatar (i.e. a synthetic 3D human face) in such a way that it realistically pronounces the given text, which is based only on the speech input. Especially important component of facial animation is the movement of the lips and tongue during speech [2]. For a realistic result, lip movements must be perfectly synchronized with the audio.

Lip synchronization is the determination of the motion of the mouth and tongue during speech [3]. Speech sound is produced by the vibration of the vocal cords in the case of voiced sounds and air turbulence in the case of whispered sounds [4]. Vocal tract, which consists of the throat, mouth, tongue, teeth, lips and nasal cavity, additionally models the produced sound. Vowels are created by the relatively free passage of breath through the larynx and oral cavity, while consonants are produced by a partial or complete obstruction of the air stream by any of various constrictions of the speech organs. Intonation characteristics are pitch, amplitude and voiced/whispered quality and they are dependent on the sound source, while vocal tract determines the phoneme. A phoneme is the basic unit of acoustic speech. Visual representation of phonemes is called viseme. There are many acoustic sounds that are visually ambiguous. Therefore, there is a many-to-one mapping between phonemes and visemes. To make lip synchronization possible, position of the mouth and tongue must be related to characteristics of the speech sound (basic idea of lip synchronization is shown on Figure 1). Whereas, positions of the mouth and tongue are functions of the phoneme and are independent of intonation characteristics of speech sound.

Various techniques have been proposed to convert human voice into the facial motion and then to drive the facial animation from speech signals [3, 4, 5, 6, 7, 8, 9, 10, 12]. The key issue is audio-to-visual mapping that converts acoustic speech to mouth shape parameters. This problem can be solved at several different levels, as it is explained in the next Section. Section 3 gives an overview of the most used approaches that attempt to extract the mouth shape information from the speech signal, while Section 4 briefly explains our proposed plan for an implementation. The paper closes with a conclusion and a discussion of the future work.

II. AUDIO AND VISUAL SIGNAL REPRESENTATION

The problem of converting a speech signal to the mouth shape information can be solved on several different levels, depending on the speech analysis that is being used [6]. These levels are:

Front end (signal level)

Acoustic model (phoneme level)

Language model (word level)

Each of the three levels can be applied within speech-driven face animation system. However, the choice will depend on a specific application, considering characteristics of the individual solution.

Signal level concentrates on a physical relationship between the shape of the vocal tract and the sound that is produced. Speech signal is segmented into frames.

Consequently, there should exist a mapping from acoustic to visual feature, frame by frame. This method uses a large set of audio-visual parameters to train the mapping. There are many algorithms that can be modified to perform such mapping – Vector Quantization (VQ), the Neural Networks (NN), the Gaussian Mixture Model (GMM), etc.

At the second level, speech is observed as a linguistic entity. The speech is first segmented into a sequence of phonemes. Mapping is then found for each phoneme in the speech signal using a lookup table, which contains one visual feature set for each phoneme. The standard set of visemes is specified in MPEG-4 and contains 14 static visemes that can be easily distinguished [11].

The language model is more concerned about context cues in the speech signals. Speech recognizer must be first used for segmenting the speech into words. Then a Hidden Markov Model (HMM) can be created to represent the acoustic state transition in the word. In the next step, one of the methods used in the first level, can be applied for each state in this model to perform mapping from audio to visual parameters, frame by frame. Because mapping is modeled inside individual words, better results can be achieved using this solution.

The latter two approaches are providing more precise speech analysis. Acoustic speech signal is explored together with the context, so that co-articulations (co-articulation is a process by which one sound effects production of the neighboring sounds) are incorporated. However, higher input signal level requires a more complex system. At the same time, because the motion of the lips, tongue and mouth can be found from the speech signal without previous recognition of phonemes or spoken words, these methods produce a certain amount of computation overhead. Another problem with phoneme level approach is the definition of different phonemes in different languages, so that there is no standard phoneme set [7]. Additionally, speaker's gender, dialect or co-articulation could be an obstacle for obtaining a correct segmentation of the phonemes for individual's speech.

On the other hand, an approach based on the low level acoustic signals is simple, language independent and suitable for real-time implementation, what is not the case in acoustic model where speech engine have to be incorporated in the system in order to obtain a phoneme sequence for a given speech.

III. AUDIO TO VISUAL MAPPING

One key issue in bimodal speech processing is the audio to visual mapping. Many approaches have been proposed in an attempt to solve the problem of extracting the mouth shape information from the speech signal. A Background of the most used techniques is given.

A. Vector Quantization

Vector quantization is a classification-based audio to visual conversion. First, the audio features (training data) are classified into one of a number of classes. For each acoustic class, the corresponding visual codewords are averaged. Each class is then mapped into a corresponding visual output. Therefore, each input acoustic feature would be classified using an optimal acoustic vector quantizer and then mapped to the corresponding visual output. This approach is computationally efficient, but it does not produce a continuous mapping.

B. Neural Networks

Mapping between the acoustic speech and the appropriate visual speech movements can be determined by training a neural network [10]. In the training phase, input patterns and output patterns are presented to the network. Suitable technology, as well as the number of hidden layers and the number of nodes per layer should be determined. Single network can be trained to reproduce all the visual parameters or many networks can be trained so that each network estimates a single visual parameter.

C. Gaussian Mixture Model

In this approach, the Gaussian mixture is used to model the probability distribution of the audio-visual vectors. Joint feature vector is composed from collected training data. Then the best estimation of the visual parameters is derived directly from the composed vector. The Expectation-Maximization (EM) algorithm is used for fitting a mixture model to a set of training data. After GMM is trained, the model is used to map an audio feature to the visual feature. The Gaussian mixture approach produces smoother results than vector quantization.

D. Hidden Markov Model

The Hidden Markov Model (HMM) can be used for audio to visual parameter conversion [8]. In the training phase, N state, left-right HMM is trained for each word in the vocabulary. For each state, the observation probability distribution is modeled by the GMM. Also, acoustic HMM is derived from the joint HMM. For each state in the acoustic HMM, optimal estimator of visual parameter is derived. In the conversion phase, acoustic HMM can be used to segment the sequence of acoustic parameters into the optimal state sequence using the Viterbi algorithm. At each state, the visual parameters can be estimated to given acoustic parameters with the optimal estimator.

E. Different approaches

Hidden Markov Model takes into consideration audio contextual information, which is very important for modeling mouth co-articulation during speech. That is not the case with vector quantization and the Gaussian mixture. Neural networks can be trained for audio to visual mapping so that they take into account the audio contextual information (e.g. time-delay neural networks - TDNN). TDNN is more computationally efficient than HMM, but requires a large number of hidden units, which results in high computational complexity during training phase.

Many approaches use a combination of the different techniques. Huang and Chen implemented a real-time audio to visual mapping using Hidden Markov Model together with Gaussian mixture model [6], while Hong, Wen and Huang [12] train Gaussian mixture models and multilayer neural network. Huang, Ding, Guo and Shum [9] calculate acoustic feature

vectors from the input voice and then find the minimum distance between the vectors and the vocal data of the sequences in the base. If the distance is larger than a threshold, the face is synthesized by HMM-based method. Otherwise, the face in the corresponding sequence is exported. Kshirsagar and Magnenat-Thalmann [5] train three-layer neural network to classify coefficients derived with linear predictive (LP) analysis into the vowels. Also, the average energy in the speech signal is used to modulate vowel-vowel and vowel-consonant lip-shape transition and zero crossing rate is used to detect fricatives.

IV. PROPOSED IMPLEMENTATION

Our intention is to model face animation in the real-time that will be driven by human voice. Furthermore, it should be language and speaker independent.

This may be realized with combination of two different techniques. One will take into account physical side (e.g. Gaussian mixture model) and the other, language side (e.g. Hidden Markov Model) of the speech. Mouth shape information extracted from speech signal will drive MPEG-4 compliant facial animation.

V. CONCLUSION

In this paper we have described the problem of converting speech signal to mouth shape information. Possible solutions have been classified into three different levels. Every level has been presented with its pro and contra. Also, a background of the most used techniques has been given.

Based on a combination of described approaches, we have proposed a plan for an implementation. Our next efforts will be to work out that plan and implementation.

REFERENCES

[1] T. Chen and R. Rao, "Audio-visual integration in multimodal communication", *Proceedings of IEEE, Special Issue on Multimedia Signal Processing*, pp. 837-852, May 1998.

[2] P. Vanroose, G. A. Kalberer, P. Wambacq, L. Gool, "From speech to 3D face animation", *Proceedings of the Benelux Symposium on Information Theory, 2002.*

[3] D. F. McAllister, R. D. Rodman, D. L. Bitzer, A. S. Freeman, "Lip synchronization for Animation", *Proceedings of SIGGRAPH 97, Los Angeles, CA, 1997.*

[4] J. P. Lewis, F. I. Parke, "Automated lip-synch and speech synthesis for character animation", *Proceedings of SIGGRAPH 1990.*

[5] S. Kshirsagar, N. Magnenat-Thalmann, "Lip synchronization using linear predictive analysis", *Proceedings of IEEE International Conference on Multimedia and Expo, New York, 2000.*

[6] F. J. Huang, T. Chen, "Real-time lip-synch face animation driven by human voice", *Proceedings of IEEE Multimedia Signal Processing Workshop, Los Angeles, California, 1998.*

[7] Y. Li, F, Yu, Y. Xu, E. Chang, H. Shum, "Speech-driven cartoon animation with emotions", Proceedings of the ninth ACM international conference on Multimedia, Ottawa, Canada, 2001.

[8] M. Brand, "Voice Puppetry", Proceedings of *SIGGRAPH'99, 1999.*

[9] Y. Huang, X. Ding, B. Guo, H. Shum, "Real-time face synthesis driven by voice", *Proceedings of Computer-Aided Design and Computer Graphics, Kunming, PRC, 2001.*

[10] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks", *Proceedings of AVSP'99, Santa Cruz, California, 1999.*

[11] I. S. Pandžić, R. Forchheimer, Editors, "MPEG-4 Facial Animation - The Standard, Implementation and Applications ", *John Wiley & Sons Ltd, 2002.*

[12] P. Hong, Z. Wen, T. S. Huang, "Real-time speech driven avatar with constant short time delay", *in Proceedings of International Conference on Augmented, Virtual Environments and 3D Imaging, Greece, 2001.*

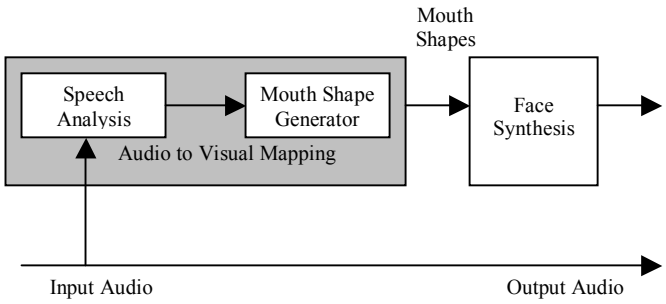

Figure 1. Basic idea of lip synchronization