MOUTH MODELING BY LOCAL PCA FOR AUDIO VISUAL SYNCHRONIZATION

Stanisław Badura, Mariusz Leszczyński, Władysław Skarbek

Multimedia Laboratory, Institute of Radioelectronics, Warsaw University of Technology

ABSTRACT

This paper describes a local PCA approach to shape and appearance modeling for animating talking heads. In the basic scenario the only input is optionally annotated text from which the system generates speech together with mouth views (visemes). It uses the PCA models, separate for each viseme class, in order to generate their flexible representatives assigned to spoken diphones at the given diphone context. In the second scenario beside the text, the video image of its reader is delivered. Then, the local PCA classifier of visemes performs labeling of audio input synchronizing the animated mouth with natural speech. The final labeling is obtained by combination of three diphone streams: one generated from text, the second obtained by inverse mapping from visemes to diphones, and the third one produced by acoustic feature diphone analysis.

1. INTRODUCTION

Talking heads are playing an increasingly important role in computer interfaces. An animated talking head attracts immediately the attention of a user. Seeing a face makes many people feel more comfortable interacting with computer. Generating animated taking heads that look like real people is a difficult task. To be considered natural, a face has to be not just realistic in appearance, but it must also exhibit proper plastic deformations of the lips synchronized with speech, realistic head movements and emotional expressions.

In computer graphics, many significant and different techniques exist for modeling talking head, achieving various degrees of realism and flexibility [1],[2],[3],[4],[5]. All synthesized heads are still far from reaching a perfect animation of lip synchronized with speech.

In this work we consider the audiovisual synchronization as a diphone – viseme (DV) correspondence. This correspondence is not unique if diphone contexts are ignored. It means that one diphone may have significantly different visual appearances, and one viseme can correspond to many diphones.

The DV correspondence can be built using the following three approaches:

- audio track is leading – diphones are generated by text to speech synthesizer or extracted by speech analyzer; one viseme class is assigned to one diphone class when the diphone context is fixed;
- visual track is leading – visemes are extracted from the visual input by viseme classifier; audio track is labeled by viseme class labels; diphone instances are simply recorded without timeline segmentation;
- audio and visual tracks are equally important – visemes are extracted from the visual input by viseme classifier; diphone class labels are extracted from three sources: viseme stream by inverse mapping, textual stream, and speech analyzer output stream.

In this paper we extract diphones using only the textual input. The speech analyzer which is language and dictionary independent, will be considered in further research.

2. DIPHONE MODELING

As offline preprocessing we have captured a set of words and sentences. The words represents all the possible phonemes in spoken language. For Polish language we have used CORPORA database [6]. For other languages we can use equivalent set of words, e.g. for American English we can use TIMIT database [7].

The speaker have uttered these phonemes with markers attached to his lips.

To get phonetic notation, we have made transcription of given text from CORPORA database.

Polish language has 36 phonemes. All the phonemes has representation in recorded speech, and using HTK tools [8] a time labels were determined for them. The labeling is required for making a phones library for concatenative speech synthesizer [9] and for selecting visemes for subsequent modeling.

The speech synthesizer is used for creating synchronous animation visual speech. Our synthesizer is using concatenative diphones method.

Diphone is defined as transition from middle of one phoneme to middle of next one. The phones library

consists of set of diphones. Every unit of the diphones library is chosen from the group of the various instances of the same diphone given, it must have the properties closest to the group average.

For the case when we need diphones which are not included in the library we, have added to the library all available instances of half-diphones. Half-diphone is equivalent of half-phoneme. From all halves of phonemes we can synthesize any required diphone.

3. VISEME MODELING

Based on phonetic time sections we could observe any viseme as visual equivalent of phonemes. Each viseme has multiple instances, but not all of them can be clear-cut classified. Therefore for the modeling viseme sets of instances were selected manually. This selection allows creating good models, but it still requires quite a lot of selected instances L.

Shape of mouth for some different phonemes look similarly, so visemes are grouped into class of visemes. Previous classification distinguishes classes of visemes by grouping images for similar uttering phonemes. We are proposing new classification method.

Firstly we distinguish six classes of shape and appearance of mouth:

1. opened mouth with upper teeth, lower teeth, and tongue visible;
2. opened mouth with upper teeth, and lower teeth visible, and tongue invisible;
3. opened mouth with upper teeth, and tongue visible, lower teeth invisible;
4. opened mouth with upper teeth visible, tongue and lower teeth invisible;
5. opened mouth visible, upper teeth, tongue, and lower teeth invisible;
6. closed mouth.

Each of the above classes with opened mouth has three subclasses referring to the degree of mouth openness: wide, medium, narrow.

Each pronounced phoneme is classified to one of the above classes depending on the uttering context in the uttered word. During pronouncing of the given phoneme the shape and the appearance of the mouth depends also on the previous and next phonemes.

The main requirement for the modeling is to keep in the model the essential information on shape and appearance of class visemes. Below there are described the steps operations which are performed for each viseme class separately:

1. Each selected instance of the classes is geometrically normalized to obtain this same size and orientation of mouth for different instances of viseme.
2. After normalization for each sample the information about lips shape is collected into shape vector data.
3. Each sample is triangulated according to example in Figure 1
4. Triangulation is used for the transformation of viseme content using barycentric coordinations in the reference viseme instance.
5. In barycentric coordinations appearance data are collected into appearance vector data. The data is collected only from triangles area.

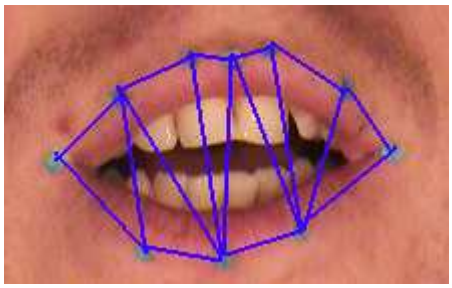

Figure 1 Triangulation order

In this way we have collected all data necessary to built any model of visemes. For the shape, the a model is build as follows:

$$X = [X_1, ..., X_L]; \quad X_i \in R^N$$

where:

N – number of markers point on lips

Average: $$\overline{X} = \sum_{1}^{L} x_i / L$$

Normalization:

$$X = [X_1 - \overline{X}, ..., X_L - \overline{X}]$$

PCA step:

$$[U, \Sigma] = svd(X);$$

$$U \in R^{NxN}; \sum = diag(\delta_1, ..., \delta_L);$$

Number of coefficients K is found from the condition:

$$\frac{\sum_{i=1}^{K} \delta_i^2}{\sum_{i=1}^{N} \delta_i^2} \geq \tau;$$

where: τ - threshold of good model $< 0,1 >$

From the designed model a viseme can be reconstructed:

$$\widetilde{X} = \overline{X} + U_K[U_K^t(X - \overline{X})]$$
$$Y = U_K^t(X - \overline{X})$$
$$\widetilde{X} = \overline{X} + U_K Y$$

In the similar way the appearance model is processed.

The PCA models can be used for animation. In order to get a higher realism we randomly disturb PCA coefficients nearby of the previous instance of the viseme class and next display the closest stored instance in the same viseme class.

4. SYNTHESIS AND ANIMATION

The main goal of our efforts is the real-time generating animated face synchronized with synthesized or real speech.

In case when the input of the synthesis and animation is only text, the text has to be preprocessed in order to expand numerals, abbreviations and acronyms to the full text form.

To get a list of diphones, the text after preprocessing undergoes phonetic and prosodic transcriptions.

Now the system has all information to select essential diphones from the diphones database for speech synthesizer and essential visemes from visemes database for mouth animation.

Consecutive diphones are concatenated to form words and accent is added by the synthesizer for every generated word.

At the end of speech synthesis generated sound is scaled in time depending on given speed of speech which is the parameter of the application.

The animation is created in parallel to the speech synthesis. The animation relies on results of phonetic and prosodic transcriptions, which is necessary for selecting appropriate visemes from database.

In the next step, the consecutive visemes are processed to animate given 3D head model.

In case when we have speech as sound track the system needs make an analysis to get a list of consecutive phonemes. The phoneme which are context sensitive transcoded into viseme symbols.

In the animation process the selected visemes are integrated with face model. For transition between visemes we used a shape morphing methods. In the intermediate frames we took advantage of textures from subclasses of currently animated consecutive visemes. At the end the system applies smoothing filters to remove any distortion arisen.

5. CONCLUSIONS

We have shown that PCA used for visual modeling new viseme classes gives robust representation of them with respect of realism in talking head animation.

On the other hand in audio modeling the proposed diphone based extraction of aural elements results in better synchronization of visual speech.

6. REFERENCES

[1] V. Blanz, C. Basso, T. Poggio, T. Vetter: Reanimating Faces in Images and Video, Eurographics 2003

[2] T. Ezzat, T. Poggio: Visual Speech Synthesis by Morphing Visemes, A.I. Memo No. 1658, C.B.C.L. Paper No. 173, May 1999

[3] Ch. Bregler, M. Covell, M. Slaney: Video Rewrite: Visual Speech Synthesis from Video.

[4] E. Cosatto, H.P. Graf: Photo-Realistic Talking-Heads from Images Samples, IEEE Transactions on Multimedia, Vol. 2, No. 3, September 2000

[5] T. Ezzat, G. Geiger, T. Poggio: Trainable Videorealistic Speech Animation, San Antonio, SIGGRAPH 2002

[6] Stefan Grocholewski CORPORA - Speech Database for Polish Diphones, 5th European Conference on Speech Communication and Technology EUROSPEECH '97 Rhodes, Greece, September 22-25, 1997

[7] "TIMIT Acoustic-Phonetic Continuous Speech Corpus", *National Institute of Standards and Technology Speech Disc 1-1.1,* NTIS Order No. PB91-5050651996, October 1990

[8] The Hidden Markov Model Toolkit (HTK) http://htk.eng.cam.ac.uk/

[9] A Short Introduction to Text-to-Speech Synthesis http://tcts.fpms.ac.be/synthesis/introtts.html