Czech Audio-Visual Speech Synthesis with an
HMM-trained Speech Database and Enhanced Coarticulation

MILOŠ ŽELEZNÝ and ZDENĚK KRŇOUL
Department of Cybernetics
Faculty of Applied Sciences
University of West Bohemia
Univerzitní 8, 306 14 Plzeň
CZECH REPUBLIC

Abstract: - The task of visual speech synthesis is usually solved by concatenation of basic speech units selected from a visual speech database. Acoustical part is carried out separately using similar method. There are two main problems in this process. The first problem is a design of a database, that means estimation of the database parameters for all basic speech units. Second problem is a way how to concatenate selected basic phonetic units so as to eliminate the coarticulation effect. Both problems are aimed in our work, resulting in the Czech audio-visual speech synthesizer. We use HMM training process instead of some form of averaging for obtaining statistically best parameters for all basic phonetic units. For solution of a coarticulation effect we use the method of dominance functions. This paper presents the new Czech audio-visual synthesis. The designed talking head provides now intelligible speech and is ready for further work on the adopting real head look, which will incorporate model adaptation to a real head shapes and applying a texture.

Key-Words: - audio-visual speech synthesis, hidden Markov models, coarticulation.

1 Introduction

Visual information is important for perception of a speech, especially in noisy environment or for hearing impaired people. There have been proposed many approaches to incorporate visual information into an originally speech-based communication between human and computer. Besides other tasks one of the most common one remains the audio-visual speech synthesis, often called talking head, which supplies the speech synthesizer (text-to-speech system) with a visual output. This task is being solved on various levels, starting from replaying stored image sequences up to video realistic fully 3D parametrically controllable talking heads.

Decided for the last choice, we started with the collection of an audio-visual speech corpus and definition of an appropriate parameterization [1]. We designed artificial 3D head as a triangular model. Lips are controllable using defined parameterization. To be able to control motion of the part of the triangular model, we map 16 points describing lip shape (control points) to 16 vertices of the triangular model. The motion of the mouth area according to these 16 points is performed by a smoothing function, which describes the motion of the vertex in dependence on the distance from the control point.

Motion of a chin is carried out separately. Controlled by 17th control point a chin adopts a rotational motion with a center of rotation given by the physical properties of the lower jaw.

Motion and a shape of lips is derived from real lips recorded by a camera with system of mirrors for stereovision. First version of a talking head used simple speech database created by selecting representative basic speech units from the corpus. This primitive approach was successful in confirmation of the functionality of controlling the part of the 3D model, but the lip motions are not smooth and are quite uncomfortable for viewing. Some enhancements should be adopted. First of them is the way of generalizing the speech unit parameters. Second problem is the solution of the coarticulation effect at the crossings between speech units, which occurs during concatenation of the speech units.

Both problems are discussed in this paper. In the very first version of our audio-visual speech synthesis we selected one representative speech unit and used its parameters as the phoneme representation in the database. This process is very sensitive to the selection of the representative instances of speech units. Some method of averaging should be used. We chose a method of
HMM training for this task. Instead of the parameters of a selected speech unit, the parameters of hidden Markov model trained on a whole speech corpus are used in the database. Although generally increase in the intelligibility of a perceived speech is expected when supplying the acoustical form of speech with visual perception, it can be decreased when the lip motion is contradictory to the perceived speech. Therefore, good interpolation of lip parameters is crucial for the overall performance of the audio-visual speech synthesizer, which can be measured by intelligibility of perceived audio-visual speech. This is more important for the interpolation on boundaries between basic phonetic units, solving the problem of coarticulation.

2 System Properties
The system of audio-visual speech synthesis is based on a method of concatenation of basic speech units, or better said, on concatenation of their parameter vectors. Movable parts of the resulting 3D model are then controlled by this sequence of parameter vectors, from whose the model is rendered for each output frame.

![Control Points](image)

Fig. 1 – An illustration of the layout of the control points, a) the original layout, b) detail of the newly added points

Used parameterization consists of 17 points placed on lip edges, a nose and a chin. Comparing to the original parameterization described in [1], 6 points were added, placed on an inner edge of the lip, see Figure 1. These parameters are obtained by an image segmentation from input stereo videos and by computing the 3D coordinates of points. For better modelling of the basic speech units we collected a corpus of about one hour of audio-visual speech. This modelling is based on HMM training with the aim of describing the statistical parameters of the basic speech units. The parameters of the trained HMMs then comprise a visual speech database.

3 Model Training
Design of a proper model is a task consisting of several decisions. Firstly, we have to choose an appropriate parameterization and design a method of extracting this parameterization from the input visual data. Once we have the sequence of the parameter vectors and relating aligned text, we have to decide how to make the speech database items based on these data. Last decision is how to concatenate basic speech units from the database and how to solve interpolation of parameter values on the inter-phoneme boundaries.

![Block Diagram](image)

Fig. 2 – Block diagram of HMM parameters training

Let us briefly explain the terminology used. By audio-visual speech corpus we mean collected audio-visual data, that means the audio files containing the acoustical part, video files containing the visual part and text files containing the phonetic transcription. Under the term audio-visual speech database we understand the result of the training process and the source from which we select basic speech units when concatenating them during synthesis. Let us note that in our case it is practically only visual speech database, since the acoustical synthesis is maintained separately by another application.
In our case, parameterization is defined as above and thus consists of 17 points, each of 3 coordinates. Total length of the parameter vector is then 51. The training data set can be segmented and divided into basic speech units. Data for these units are in the form of sequences of parameter vectors. In the very first version of the visual speech synthesis we created the speech database by selecting the representative speech unit. For a generalization of this process we now utilize the HMM training framework. The process of visual speech database training is illustrated on Figure 2.

Another important issue is the choice of the type of the basic speech unit. Different approaches can be adopted. For example, for the Japanese language the efficient choice of basic speech unit is a syllable [2,3]. For the simplicity of synchronization with acoustical synthesis the phoneme is often used [4,5]. The main disadvantage of phoneme as basic speech unit is that it is independent of the context. This causes the problem of coarticulation later during concatenation. Possible solution is the use of triphone as a basic speech unit. The main disadvantage of triphone is that number of triphones is much higher than phonemes, which puts higher demand on the training set and the modelling process. It is obvious that some compromise should be chosen.

For our task we have to take into account also the properties of the Czech language. A syllable is not usable here, since in the Czech language there are much more syllables and they also appear in various structures (for example many syllables are in the form "CVC", and not only "CV", as reported for Japanese in [2]. We can also mention special syllables like "CCCCC"="čtvrť" with syllable forming "r", note that "C" stands for consonant and "V" for vowel). Considering the very high number of triphones we chose the compromise—a phoneme as a basic speech unit. The coarticulation problem have to be solved by some other way.

Training of the HMM models is based on standard Baum-Welch procedure used for training acoustic HMMs for speech. The only difference is in the parameterization, which now describes lip shape and motions instead of the vocal tract description. We use the HMM phoneme model with the topology illustrated on Figure 3. We use 5-state left-to-right HMM with 3 central emitting states. We left out the transition from state 2 to state 4 to force transition through the central state 3. The training process is carried out using the HTK system [6]. For the visual speech database, means of HMM model parameters are used. This provides the generalization of the parameter values.

4 Interpolation and Coarticulation

Using the data stored in the database, we now carry out the concatenation of basic speech units. In our concept, the acoustical and visual syntheses are performed separately. The acoustical synthesis is accomplished by the ARTIC text-to-speech system [7]. This system provides a sound file and a notification file as an output. The notification file contains information about the time boundaries between speech units. We need to use this information for synchronization of the two data streams. The acoustical synthesis thus serves as the master process, while the visual synthesis remains the controlled process.

There are two main problems in this task. Firstly we need to interpolate parameter values at the time instances of output device (window rendering or file output). These time instances are given by the frame rate of the output stream, which can vary. It is obvious that these time instances do not fit into time instances with which the data are stored in the database. Second problem is the coarticulation, effect that appears at the crossings of neighbouring speech units. This effect is caused by the fact that each speech unit stored in the visual speech database is trained using various contexts in the source data. Coarticulation solves this problem by smoothing the parameter values at the crossings between the basic speech units.

4.1 Interpolation

The output of a concatenation method is a sequence of parameter vectors for controlling the motion of lips and a chin on a rendered 3D model. Let us now interpret these data rather as a set of sequences representing the time evolution of the individual parameter values. We need to compute values at the time instances needed by the rendering process.

To solve this problem we need to employ some interpolation method. We decided for using the cubic spline functions. For each parameter value we build cubic spline functions for interpolation of the parameter values. Knowing these functions we can compute the parameter value at any time instant.
4.2 Coarticulation

When the concatenation and interpolation is complete, we need to apply some smoothing function on the resulting visual stream so as to avoid the coarticulation problem. There are many solutions of this task. Our approach is based on the one proposed by Löfqist [8] and modified by Cohen and Massaro [9]. This coarticulation model provides good linguistic, adaptable and complex control of the human face coarticulation. This model assumes the existence of a dominance and blending function. The dominance is the influence on coarticulation of visual segments. The dominance function is a negative exponential function, which falls with running time $t$ from the center of a segment. For each phoneme segment and each parameter we have two dominance functions for the backward and forward coarticulation.

A dominance is the influence on coarticulation of visual segments. The dominance is given by the negative exponential function

$$D = e^{-Ot}$$

This function is falling with running time $t$ from center segment. The falling is driven by parameter $c$ and $O$. The parameter $c$ modifies the steepness and $O$ modifies the rate of the falling. Cohen and Massaro [9] proposed function

$$D_{sp} = \alpha * e^{-Ot}$$

where $D_{sp}$ is the dominance of facial parameter $p$ on segment $s$. The parameter $\alpha$ drives the magnitude of the dominance function and thus drives the magnitude of influence on neighboring speech segments. For each speech segment and each parameter we have two $D_{sp}$ functions for the backward and forward coarticulation.

Some studies reported that coarticulation is different for different languages. As an example, Lubker and Gay [10] reported study on differences between American English and Swedish, Boyce [11] reported study on differences between American English and Turkish. Differences can be modelled by changing of parameters of the dominance function. For a synthesized visual sequence we have to design a smoothing function that is based on dominance functions of neighboring phonemes and that controls the smoothing of lip motion and provides more naturally looking visual speech.

Thus, for the Czech language we estimated values of the dominance function parameters $\alpha$, $O$, and $c$ for all features of all phonemes from our visual speech corpus.

Effect of an interpolation and coarticulation on the generated sequence is illustrated on Figure 4. A curve labelled "original" shows the evolution of a parameter value in the training data, a curve labelled "database" represents output of the HMM training process and corresponds to the values in visual speech database. A curve labelled "interpolated" shows the result of the interpolation by cubic spline functions and a last one labelled "smoothed" represents the results of smoothing using proposed method of dominance functions.

5 Experiments

For evaluation experiments we used two versions of the audio-visual speech synthesizer. Let us label them as version A and B. Version A uses the original approach, uses phoneme as basic phonetic unit, visual speech database is created by selecting representative speech units from the training data. Version A uses bicubic interpolation of parameters, but no smoothing function. Version B is the implementation of the approach presented in this paper. Version B uses phoneme as a basic phonetic unit, visual speech database is created by training HMMs, as parameters of speech units stored in database are used means of trained HMMs. Version B uses bicubic interpolation and smoothing function based on dominance functions. Dominance function parameters were optimized for the Czech language. We used the 3D model of a head depicted in Figure 5.

To compare these two versions and also for comparison of audio-visual synthesis with acoustical-only synthesis we performed a listening test. Listening test consists of 2 sets: acoustical only, and audio-visual synthesis. Audio-visual part of the synthesis contained mixed version A and version B outputs. Listeners had to evaluate the intelligibility of the perceived speech. They did not know which sentence is generated by version A and which by
version B. They first listened (and watched) the output of synthesis, then they read the text and evaluate the intelligibility by a mark from 0 (not intelligible at all) to 10 (perfectly intelligible).

Fig. 5 – Shaded view of the 3D talking head model

Let us note that both version A and B use the same version of the acoustical synthesizer (ARTIC). This version of acoustical synthesizer uses worse speech database than that presented in [7]. This database was created from the same data recorded for the audio-visual database, that means about 1 hour of speech, mostly short sentences. Due to recording by camera simultaneously this database contains also more noise than the original ARTIC database. Using this simpler version of ARTIC was intentional to prove increase in intelligibility of audio-visual combination, especially in the case when the speech itself is not very intelligible.

Table 1 shows the result of the listening test. Numbers in the right column are averages of intelligibility factors over all sentences over all listeners. Result shows that there is really increase in intelligibility of speech when perceiving the auditory information in combination with visual information. But it also shows that without good interpolation and coarticulation the increase in intelligibility is very slight. The results in this table are not very detailed. Let us describe one phenomenon that is not visible straightly from the averages. When perceiving audio + visual ver. A, sometimes the intelligibility factor was lower than for audio only. It shows that contradictory or confusing visual perception can worsen the intelligibility of perceived speech.

<table>
<thead>
<tr>
<th>synthesis version</th>
<th>overall intelligibility index</th>
</tr>
</thead>
<tbody>
<tr>
<td>acoustical only</td>
<td>6.1</td>
</tr>
<tr>
<td>audio-visual ver. A</td>
<td>6.7</td>
</tr>
<tr>
<td>audio-visual ver. B</td>
<td>7.8</td>
</tr>
</tbody>
</table>

6 Conclusion
We have proposed a solution to the generalization of visual speech database for the Czech audio-visual synthesis. The approach is based on a combination of two techniques. A phoneme HMM training more reflects properties of basic speech units throughout the whole corpus. Context independency of the phoneme is then solved by interpolation by cubic splines and smoothing by dominance-based function. Performed experiment proves the increase in intelligibility of combined audio-visual perception of a speech.

Our future work will be directed to the 3D head model adaptation to a shape of a real head and texture application. Our ultimate goal is to be able to change the shape and texture of the talking head according to a head of any real person.

7 Acknowledgements
This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/03/0650 and by the Ministry of Education of the Czech Republic, project No. MSM235200004.

References:


