A REAL-TIME LIP SYNC SYSTEM USING A GENETIC ALGORITHM FOR AUTOMATIC NEURAL NETWORK CONFIGURATION

Goranka Zoric, Igor S. Pandzic

Department of Telecommunications
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, HR-10000 Zagreb, Croatia
{Goranka.Zoric, Igor.Pandzic}@fer.hr

ABSTRACT

In this paper we present a new method for mapping a natural speech to the lip shape animation in the real time. The speech signal, represented by MFCC vectors, is classified into viseme classes using neural networks. The topology of neural networks is automatically configured using genetic algorithms. This eliminates the need for tedious manual neural network design by trial and error and considerably improves the viseme classification results. This method is suitable for real-time and offline applications.

1. INTRODUCTION

A human speech is bimodal in its nature [1]. A 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. In noisy environments, a visual component of the speech can compensate for a possible loss in speech signal. This combination of the auditory and visual speech recognition is more accurate than auditory only or visual only. Use of multiple sources generally enhances a speech perception and understanding. Consequently, there has been a large amount of research on incorporating bimodality of speech into human-computer interaction interfaces. Lip sync is one of the research topics in this area.

The goal is to animate the face of a speaking avatar in such a way that it realistically pronounces the text based on the process called lip synchronization. For a realistic result, lip movements must be perfectly synchronized with the audio. Other than the lip sync, realistic face animation includes face expressions and emotions. However, this is not in the scope of this paper.

For interactive applications it is necessary to perform lip sync in real time, which is a particular challenge not only because of the computational load, but because the low delay requirement reduces the audio frame available for analysis to a minimum.

The next section introduces the problem of the lip synchronization. Section 3 gives an overview of our automatic lip sync system, while Section 4 explains training of the neural networks with genetic algorithms. Implementation of our system is briefly described in Section 5. The paper closes with a conclusion and a discussion of the future work.

2. BACKGROUND

The speech sound is produced by a vibration of the vocal cords and then it is additionally modeled by the vocal tract. A phoneme, defined as basic unit of the acoustic speech, is determined by the vocal tract, while intonation characteristics (pitch, amplitude, voiced/whispered quality) are dependent on the sound source.

Lip synchronization is the determination of the motion of the mouth and tongue during speech [2]. To make lip synchronization possible, position of the mouth and tongue must be related to characteristics of the speech signal. Positions of the mouth and tongue are functions of the phoneme and are independent of intonation characteristics of speech.

There are many sounds that are visually ambiguous when pronounced. Therefore, there is a many-to-one mapping between phonemes and visemes, where viseme is a visual representation of phoneme [3].

The process of the automatic lip sync consists of two main parts (Figure 1). The first one, audio to visual mapping, is a key issue in bimodal speech processing. In this first phase the speech is analyzed and classified into viseme categories. In the second part, calculated visemes are used for animation of virtual character's face.

Audio to visual mapping can be solved on several different levels, depending on the speech analysis that is being used [4]. These levels are: front end or signal level,

0-7803-9332-5/05/$20.00 ©2005 IEEE

acoustic model or phoneme level and language model or word level.

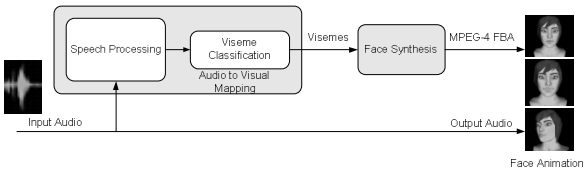

Figure 1: Schematic view of lip sync system

The choice will depend on a specific application, whereas a balance between time needed for signal processing and the quality to be achieved must be found.

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. A mapping is then performed from acoustic to visual feature, frame by frame. 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 [3].

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 [4].

The latter two approaches are providing more precise speech analysis. Speech signal is explored together with the context. However, higher input signal level requires a more complex system. Another problem with phoneme level approach is the definition of different phonemes in different languages [5].

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.

3. OUR LIP SYNC SYSTEM

Our system for automatic lip sync is suitable for real-time and offline applications. It is speaker independent and multilingual. Our system is in between signal and phoneme level, as we use visemes as the main classification target. Speech is first segmented into frames. For each frame most probable viseme is determined. Classification of speech into viseme classes is performed by neural networks. Then MPEG-4 compliant facial animation is produced.

Next sections present the components of the system.

3.1. Phoneme database

As training data, a set of phonemes is collected. For every phoneme, three different samples were recorded by nine test subjects. This gives 27 versions of each phoneme in our database. These phonemes are manually mapped onto MPEG-4 visemes, and in doing so the database is organized in 14 classes, each corresponding to one MPEG-4 viseme. On average, each viseme class is represented by 50 samples in the database.

For fine tuning of animation, phonemes specific for certain language might be added in the database.

Accuracy of the speech classification depends greatly on the quality and the size of the recorded database.

3.2. Audio to Visual Mapping

In order to synchronize the lips of a computer generated face with the speech, speech must be first preprocessed and then classified into visemes.

3.2.1. Speech Analysis

MFCC representation of the speech is chosen as first step in preprocessing the speech.

The Mel-Frequency Cepstrum Coefficients (MFCC) is audio feature extraction technique which extracts parameters from speech similar to ones that are used by humans for hearing speech, while, at the same time, deemphasizes all other information. As MFCCs take into consideration the characteristics of the human auditory system, they are commonly used in the automatic speech recognition systems [6].

Additionally, Fisher linear discriminant transformation (FLDT) is done on MFCC vectors to separate classes. If there is no separation between classes before FLDT, transformation will not enhance separability, whereas if there is only slight distinction between classes, the FLDT will separate them satisfactory.

In order to use MFCCs on the speech signal, frame length and the dimension of the MFCC vectors must be determined. The frame length must be chosen, so that the frame contains enough information [6]. The choice is frame length of 256 samples and 12 dimensional MFCC vector. Overlapping of the frames is used to smooth transition from frame to frame (Figure 2). The phoneme database is now used as a training set in order to train neural network, as it will be described in the next Section.

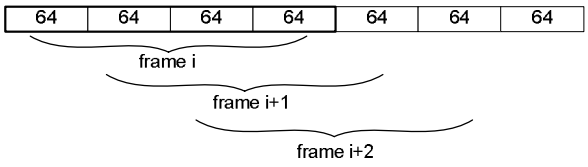

Figure 2: Overlapping of the frames

3.2.2. Lip Shape generator

Every frame of the speech is classified in the correct viseme class. When correct viseme is chosen, it can be sent to animated face model.

To reduce sudden discontinuous facial expressions, the outputs from neural network for four consecutive frames are analyzed [6]. One with the largest output sum is chosen as a correct viseme.

3.3. MPEG-4 Face Animation

Face animation (FA) is supported in MPEG-4 standard [3]. MPEG-4 FA specifies a face model in its neutral state, a number of feature points (FPs) and a set of Facial Animation Parameters (FAPs). Each FAP corresponds to a particular facial action deforming a face model in its neutral state. The first group of FAPs contains high-level parameters, visemes and expressions. Only 14 static visemes are included in the standard set.

Facial animation can be generated for any parameterized face model for speech animation if the visemes are known.

4. TRAINING NEURAL NETWORKS FOR AV MAPPING USING GA

Neural networks (NNs) are widely used for mapping between the acoustic speech and the appropriate visual speech movements [7]. Many parameters, such as weights, topology, learning algorithm, training data, transfer function and others can be controlled in neural network. A major unanswered question in NN research is how best to set a series of configuration parameters so as to maximize the network's performance.

As training neural network is an optimization process where the error function of a network is minimized, genetic algorithms can be used to search optimal combination of parameters.

Genetic algorithms (GA) are a method for solving optimization or search problems inspired by biological processes of inheritance, mutation, natural selection and genetic crossover. A conventional GA consists of coding of the optimization problem and set of the operators applied on the set of possible solutions [8]. The algorithm's three main operators are selection, crossover and mutation. Only individuals that are good enough get the possibility to survive.

GAs might be used to help design neural networks by determining [9]:

- *The weights*. Algorithms for setting the weights by learning from presented input/output examples with given fixed topology often get stuck in local minima. GAs avoid this by considering many points in the search space simultaneously.
- *Topology* (number of hidden layers, number of nodes in each layer and connectivity). Determining a optimal topology is even more difficult – most often, an appropriate structure is created by intuition and time consuming trial and error.
- *A suitable learning rule*.

However, they are generally not used in all three problems at the same time, since they are computationally very expensive.

4.1. Training NNs

In our approach, we use multilayer feedforward networks to map speech to lip movements. These kind of neural networks are widely used and operate in a way that an input layer of nodes is projected onto output layer through a number of hidden layers. Backpropagation algorithm is used as training algorithm for adjusting weights.

After experimenting with different number of networks, we have decided to train 15 networks, one per viseme class. The reason is the following. A single network tested on the entire training data gave poor results, computation time was long and recognition was not satisfactory [6]. Then the number of networks was chosen to be the same as the number of phonemes. The results were satisfactory, but it seemed like a good idea to reduce number of networks, if possible, since training of every network requested extra time.

Therefore, we took advantage of idea that phonemes that are visual ambiguous, do not need to be separated, since it does not influence animation.

The 12-dimensional MFCC vectors are used as inputs to 15 different networks. For each viseme class, a NN with 12 inputs, a number of hidden nodes and 1 output is trained. The number of hidden layers and the number of nodes per each layer should have been determined for each network. This is laborious and time consuming work since the training session must be run until the result is satisfactory. In order to avoid time consuming trial and error method, we introduced simple genetic algorithm to help find suitable topology for our NNs.

4.2. GA and NNs in our approach

Since the design of neural network is optimized for a specific application, we had to find suitable network for our lip sync application. As determining a good or optimal topology is even the most difficult task in design of NN, we tried to solve this problem with GAs.

In our example, given the learning rule (Levenberg-Marquardt), we used GA for training a backpropagation feedforward network to determine near optimal network topology, including the number of hidden layers and the number of units within each layer.

We use simple genetic algorithm, where number of genes specify the number of hidden layers (*n*). Gene maximum and minimum values are defined in the range from zero to *m*, determining the number of nodes per layer. If a value of the single gene is set to zero, the number hidden layers is decreased, so practically it ranges from zero to *n*. Other parameters that have to be specified are population size, maximum number of generation and mutation rate.

Experiments have shown that it is not necessary to have more than three hidden layers (n = 3). From the same reason, maximum number of nodes per layer is set to 30 (m = 30). Such configuration of GA seems suitable since larger network increases computation time, but does not give better results.

By using genetic algorithms, the process of designing neural network is automated. Once the problem to be solved is coded and GA parameters are determined, the whole process is automated. Although it is still a time consuming work, much time is saved by making the process automatic.

5. IMPLEMENTATION

Database construction and creation of 15 neural networks have to be done only once. In the training process, network's biases and weights are extracted and saved for later use. Together with Fisher matrix (obtained by calculating FLDT), biases and weights matrix are loaded in the application.

Application captures speech from the microphone and segments it into frames of 256 samples. When a frame has been captured, data is stored and calculations are performed during capturing of the next frame. These calculations consist of MFCC extraction and simulation of 15 networks. The outputs are added to outputs from the previous frame. Every fourth frame, the viseme class that has the largest sum of output values from NNs is presented on the screen [6]. It is important that calculation time does not exceed time needed for recording of a frame (in case of 16 kHz, 16 bit coding it is 16 ms).

6. CONCLUSION AND FUTURE WORK

In this paper we have described our approach for lip sync system by speech signal analysis. Speech is classified into viseme classes by neural networks and GA is used for obtaining optimal NN topology. By introducing segmentation of the speech directly into viseme classes instead of phoneme classes, computation overhead is reduced, since only visemes are used for facial animation. Automatic design of neural networks with genetic algorithms saves much time in the training process. Moreover, better results are achieved than with manual search of network configuration.

Our next step will be to extract face expressions in addition to lip movements from the speech signal.

7. ACKNOWLEDGMENT

The initial version of this lip sync system has been implemented by A.Axelsson and E. Björhall as part of their master thesis of Linköping University [6] and in collaboration with Visage Technologies AB, Linköping, Sweden. This work is also partly supported by Visage Technologies.

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

[3] I.S. Pandzic, R. Forchheimer, Editors, “MPEG-4 Facial Animation - The Standard, Implementation and Applications“, John Wiley & Sons Ltd, ISBN 0-470-84465-5, 2002.

[4] 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.

[5] 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.

[6] A.Axelsson and E. Bjorhall, "Real time speech driven face animation", Master Thesis at The Image Coding Group, Dept. of Electrical Engineering at Linkoping University, Linkoping 2003.

[7] J.J.Davila, "Genetic optimization of neural networks for the task of natural language processing", dissertation, New York, 1999.

[8] R.Rojas, "Neural networks, A Systematic Introduction", Springer-Verlag Berlin Heidelberg, 1996.

[9] A.J.Jones, "Genetic algorithms and their applications to the design of neural networks", Neural Computing & Applications, 1(1):32-45, 1993.