UACEE International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 3 : Issue 2 [ISSN 2250 – 3765]
Publication Date : 05 June 2013

Interactive animation for user verification with the help of audio-visual parameters for enhanced Human Computer Interaction (HCI)

Anurag Pujari
Abhishek Tyagi

***Abstract*— Animation is a broadly explored research field. In spite of it, the field has not been quested much for enhancing the real time user interaction. Research on motion detection has been done for identifying an interacting user's identity, though it holds certain limitations. Our research presents an expressive facial animation model which identifies the interacting user as well as authenticates. The model animates the face in such a way that it is realistically pronouncing the given text, which is based on the speech input. The input speech is the vocabulary words that have been chosen by us specifically for this model. This intelligent animated model can identify an interacting user and works for that particular user only.**

***Keywords*--SpeakerRecognition,Lipsynchronization, Performance driven facial animation(PDFA)**

I. *Introduction*

Human interaction to understand each other happens with the help of communication with each other; the same happens in Human Computer Interaction (HCI). According to Ren.C.Luo, Shu-Ruei Chan, Chien-Chieh Huang, Yee-Pien Yang in Human Robot Interaction Using Speech Synthesis and Recognition with Lip Synchronization [1], that auditory and visual speech synthesis and recognition are crucial abilities for auditory signal. Mimic lip synchronization also contributes improvement to the human robot interaction. In our approach, we are trying to enhance HCI by substituting the robot by a performance driven facial animation (PDFA).A facial animation which can synchronize the speech of a user with lip synchronization can effect highly in the interaction of human and computer. The lip sync focuses on the words that have been uttered by the user. The animation is further integrated for roll playing environment. In face tracking as an augmented input in video games: Enhancing presence, roll playing and control by Shuo Wang and colleagues[2] for Microsoft Research Asia, they had a flaw that there system was not holding unique identity for each player. Therefore, this paper is going develop a model for better Human Computer Interaction (HCI) by speech synchronization, speech recognition, synthesized speech and Performance driven facial animation (PDFA). A unique identity for players enhances the presence of the player in the game. We are proposing a model where each animation (which is assigned to each user) will have an identity according to their speech. Hypothesis of this work is providing a delicate human computer interaction (HCI) with the help of animation as well as to provide more security in the roll-playing games (RPG) in case of providing unique identity in multiplayer gaming enviroment. In this work, the system needs the following:

- User
- A computer system
- Peripheral devices such microphone

The steps which has to be follow to deploy our work are as follows:

- User provides his voice input to the system through a peripheral for voice recognition
- If the previously stored parameters for the user matches , user is assigned the particular animation
- As this animation includes his identity, the user can now use his identity to work on any roll-playing application or system
- The animation assigned to the user now work for him for his voice input; i.e. it provides a lip synchronozed speech for the voice input of that particular user.

The goal of this thesis is to propose as well as implement a system for lip synchronization that will be suitable for realtime and offline applications.The thesis is focused on signal processing of audio signal and synchronizing the audio visual mapping.

II. *A. Work on Lip Synchronization*

In 2011, Ren.C.Luo, Shu-Ruei Chang, Chien-Chieh Huang, Yee-Pien Yang [1] proposed a speech driven model for lip synchronization as well as for enhanced human robot interaction. In their proposed model, they have designed we a lip synchronization system for their humanoid robot using Microsoft Speech API (SAPI). With thirty degrees of freedoms (twelve for mouth) they have built sixteen lip shapes to perform all the visemes. The precise and mimic lip synchronization can let the users to have favorable impression, and gains the closeness to the people. At last the whole system is implemented on their humanoid robot head to demonstrate the success of the system. Their model consists of five parts:

- The architecture of lip synchronization and human robot interaction system,
- The speech synthesis and recognition system,

- The lip shape design
- Lip synchronization control system
- Robot neck control

Their system deals with text input and wave input because it needs not only speak directly by text input but also realizes the auditory input of meaningful sentences. The text input is processed by speech synthesis engine, and synthesized waves and viseme events are produced. After that the speaker plays the artificial speaking and the robot mouth synchronizes simultaneously. Relatively, the wave input is processed by speech recognition engine. It converts the recognized spoken waves into text strings. Once the main system receives the text strings converted by speech recognition engine, it will give responses according to those text strings by speech synthesis process and lip synchronization process. All these processes are automatic and real-time. In our proposed model, we have replaced the robot by a performance driven facial animation where the robot will identify a user from its voice input.

Authors name: Anurag Pujari
University name: Lovely Professional University
Country: India
Email: anurag.pujari123@gmail.com

Authors name: Abhishek Tyagi
University Name: Lovely Professional University
Country: India
Email: abhishektyagi43@gmail.com

A. *Work on presence of a user in a virtual world*

In 2006, Shuo Wang, Xiaocao Xiong, Yan Xu, Chao Wang, Weiwei Zhang, Xiaofeng Dai and Dongmei Zhang [2] has designed as well as implemented two game prototypes. These two prototypes applied real time face positioning information as built-in element of gameplay to enhance the experience of gaming. The first prototype has amplified the typical motion detection based game. This model is designed to increase the sense of presence of a user in a roll-playing game. In the second prototype, face tracking is applied as a control. The result of this paper demonstrated that after comparing face tracking and non face tracking camera based video games, tracked information of face can effectively enhance the presence of roll playing. Unlike motion detection, information used in face tracking is constant and doesn't require constant motion from the user. The future work of this paper was to study the circumstance of multiple side by side players where each one is holding a unique identity. In our current work, the identity is the field which we are trying to achieve.

C. *Work on facial animation*

In 2005, Scott A. King, and Richard E. Parent [3] demonstrated in their research a facial model that support facial animation. Their model used a muscle based parameterization. This model allowed an easier integration between speech synchrony and facial expression. Whereas, our research focuses on the same field but with more standard visems.

III. *Architecture Of our proposed model*

The architecture of our system has two phases which have been figured on figure 1(a) and figure 1(b) as below:

Figure 1(a)

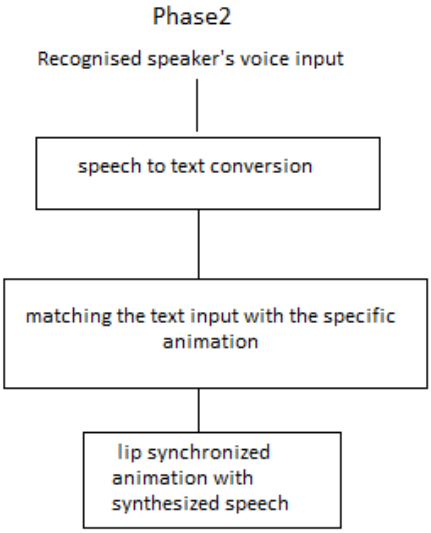
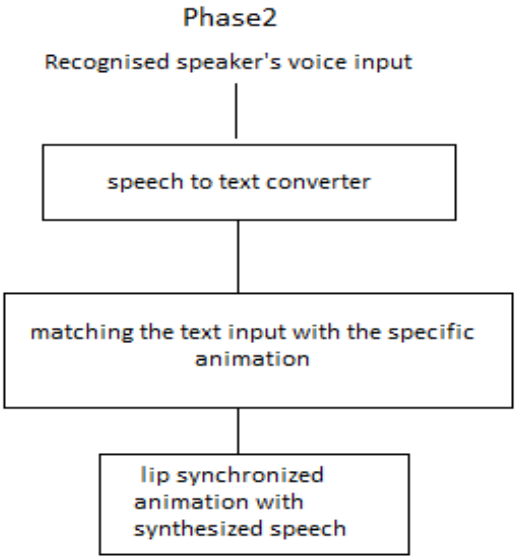


Figure 2(b)



A. *Phase 1:*

Phase 1 includes the two steps:

- Speech processing
- Speech recognition

1. *Speech processing*

For speech processing, we have followed Mel-Frequenct Cepstrum and Cepstral Coefficients (MFCC). Mel-frequency cepstral coefficients make up an MFC. These coefficients are derived from the type of cepstral representation of the speech/audio/wav files. The difference between the Mel-frequency cepstrum and cepstrum is that in MFC.

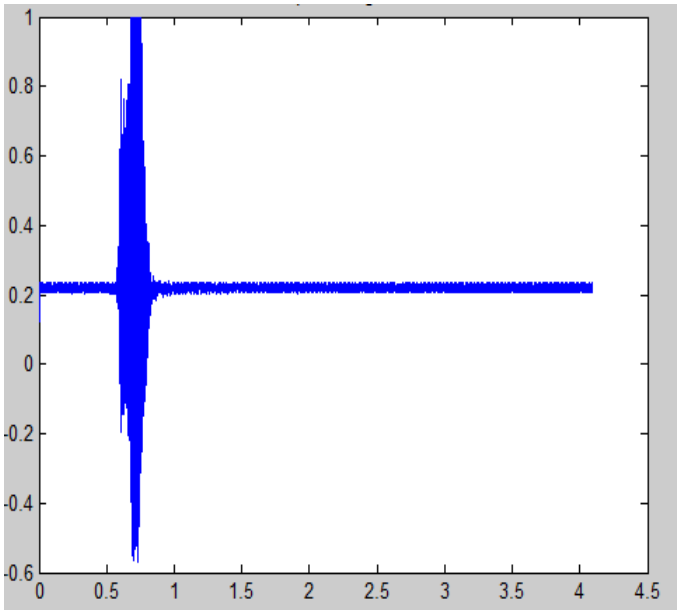
There are six computational steps for MFCC. The steps performed are described below. We have performed the steps using MATLAB 7.6.0 (R2008a).

- Pre-Emphasis

The signal of the wav file that has been stored in the database for recognition has to pass through a filter which emphasizes higher frequencies. Pre-emphasis is designed to increase the magnitude of some frequencies i.e. usually the higher frequencies with respect to the magnitude of lower frequencies so that to improve the overall signal-to-noise ratio. Equation of pre-emphasis is:

$$Y(n)=X(n) – 0.95*[n – 1] \quad (1)$$

The output after applying pre-emphasis is figured below:
Figure 2(a)

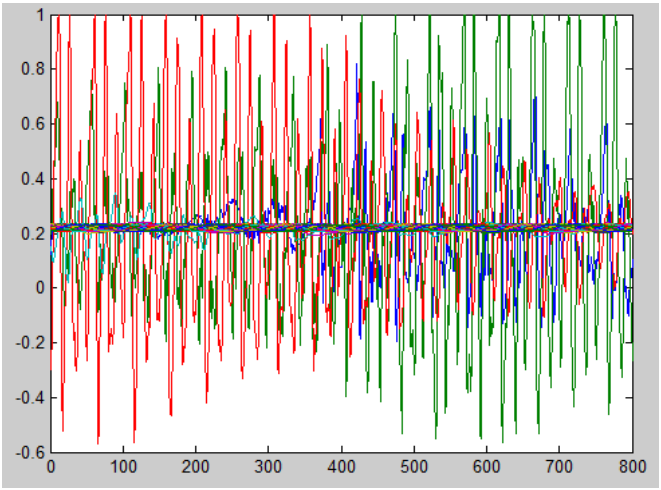


- Framing

The process of sampling the speech samples pre-emphasis into small frames. Frames contain N samples each. We have to take care of the window size during sampling the signal.

.

Figure 2(b)



- Hamming windowing

Hamming window integrates all the closest frequency lines. Equation of hamming window:

$$Y(n) = X(n) * W(n) \text{ where } 0<=n<=(N – 1) \quad (2)$$

Here Y (n) = output signal

X (n) = input signal

$$W(n) = 0.54 - 0.46 \cos(2*pi*n / N -1) \quad (3)$$

- Fourier transform of the signal

We have taken the Fast Fourier Transform of the wav signal which will help us to know what frequencies are present in our signal and in what proportions.

- Mel Filter Bank Processing

Maps the powers of the spectrum which is obtained above onto the mel scale. We have used triangular overlapping windows where magnitude of each filter's frequency response is triangular in shape. It is equal to unity at the centre frequency. After that it decreases linearly to zero at centre frequency of two adjacent filters. Each filter output is the sum of its filtered spectral components. Following equation is implemented after that to compute the Mel for given frequency:

$$F(Mel) = [2095 * \text{Log}_{10}(1 + f/700)] \quad (4)$$

- Discrete Cosine Transform

Then we have taken the discrete cosine transform of the above mel log powers.

2. *Speech Recognition*

The recognition i.e. comparing the preloaded pattern of voice input with microphones data stored in array using mean square error (MSE) method. The formula of Mean Square Error is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

Where $\hat{Y}$ is the vector of currents voice input and Y is the vector of the preloaded voice input that has to be matched with.

After calculating the MSE, if the MSE is lower than the threshold, it proves that the speaker's voice matches with the preloaded voice command.

B. *Phase 2*

Once the speaker is recognized, he/she is allowed to move for the phase2 of our system. The phase2 contains the following:

- Speech to text conversion and Matching the text with specific animation
- Lip synchronized animation with synthesized speech

1. *Speech to text conversion and matching the text with the specific animation*

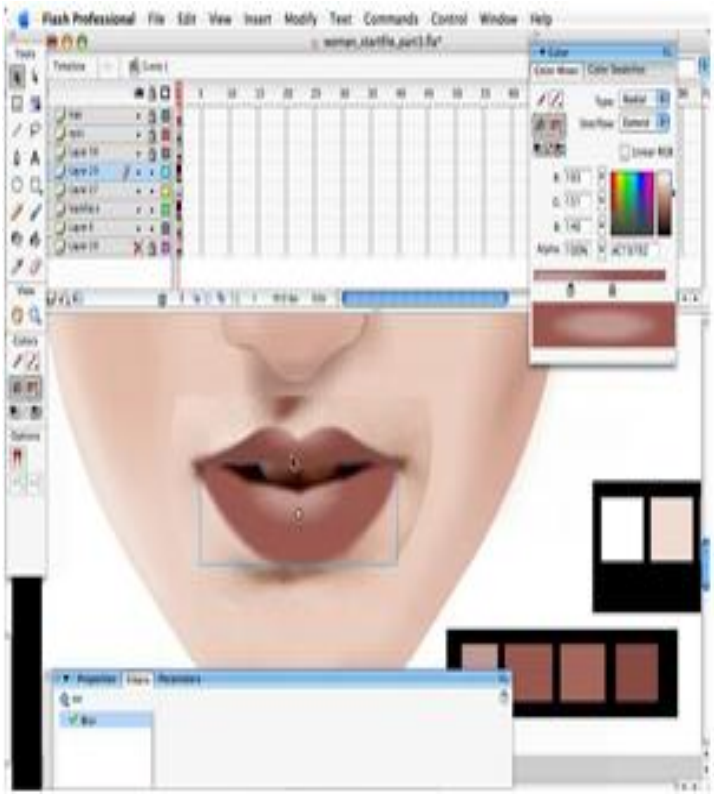
For speech to text conversion, we have used Microsoft Speech API (MSAPI). We have trained our system for some specific words. When these words are uttered by the speaker, the words are mapped with the preloaded files of the animations and required file is played. The vocabulary for our system included the following words:

I, we, you, have, has, she, he, see, saw, seen, drink, drank, drunk, dinner, lunch.

2. *Lip synchronized animation with synthesized speech*

This is the most important part of our model .A face animation is supported in MPEG4 standard. There are certain feature points (FP) as well as Facial Animation Parameters (FAP). Out of the 84 FP's, we have focused only on the FP's of the lips. We have constructed the facial model using MACROMEDIA FLASH PROFESSIONAL. For defining arbitrary face models, MPEG4 defines Facial Animation Parameter Units (FAPUs) which scales FAPs for any face model.

The figure of our facial model is figured below:



We have constructed animated lip expression for each word in our vocabulary. As MATLAB doesn't support .swf extension, so we have converted the .swf files to .avi files using an application Media Convertor [4]. The lip synchronized file of animation is integrated with the MSAPI codes for synthesized speech. Once an animated .avi file is called, the corresponding synthesized speech is uttered by a synthesized voice.

IV. *Results*

Speech recognition module is giving us the optimum result though in presence of noisy data, its performance slightly moves down.

Speech to text conversion using MSAPI depends totally on the training as well as on the input peripheral i.e. on microphone.

The lip synchronized speech animation module is giving the optimum result. The synchronization of facial animation and synthesized speech is very important. In spite of it, some time delay must be accepted as speech is spoken after its specified animation is played.

V. *Conclusion and Future Work*

In this paper, we have proposed a lip synchronized animation which performs lip synchronization for a particular user who has passed through voice recognition module at first and then provides its voice input for speech to text conversion. Once the voice input has mapped with the preloaded vocabulary of the words, a lip synchronized animation synthesizes the spoken word.

The concept of our work is focused on Human Computer Interaction (HCI). So, interaction of the user can be increased in more specific way by taking the facial value of the user as an input using webcam of the user's computer. This scope can be further extended to Human Robot Interaction (HRI) so that a robot can identify a user with the help of voice input as well as face value input.

References

[1]. Ren.C.Luo, Shu-Ruei Chang, Chien-Chieh Huang, Yee-Pien Yang , Human Robot
Interactions Using Speech Synthesisand Recognition with Lip Synchronization[2011]
[2] Face Tracking as an Augmented Input in Video Games:Enhancing Presence, Roleplaying
and Control by Shuo Wang,Xiaocao Xiong, Yan Xu, Chao Wang, Weiwei Zhang,
Xiaofeng Dai, Dongmei Zhang[2006]
[3] Scott A. King, and Richard E. Parent, Creating Speech-Synchronized Animation[2005]
[4] swftoaviconverter.org/

About Author (s):

Anurag Pujari
Student of Lovely Professional University
Department: CSE-IT

Abhishek Tyagi
Assistant Professor
Student of Lovely Professional University
Department: CSE-IT