Automated Lip-Sync for 3D-Character Animation

T. Frank, M. Hoch, G. Trogemann
Academy of Media Arts
Peter-Welter-Platz 2, D-50676 Cologne, Germany,
e-mail:{micha,georg}@khm.de

Keywords: facial animation, speech recognition, neural networks

ABSTRACT

A central task for animating computer generated characters is the synchronization of lip movements and speech signal. For real time synchronization high technical effort, which involves a face tracking system or data gloves, is needed to drive the expressions of the character. If the speech signal is already given, off-line synchronization is possible but the animator is left with a time consuming manual process that often needs several passes of fine tuning. In this paper we present an automated technique for creating the lip movements of a computer-generated character from a given speech signal by using a neural net. The net is trained by a set of pre-produced animations.

INTRODUCTION

One of the major problems in character animation is the synchronization of the speech signal and lip movements. If the speech signal is already given, off-line synchronization is necessary by using a time-consuming process called "lip-sync" that is also used in classic cartoon animation. Here, the speech signal is marked with a time-code which is then used to manually determine when a certain expression is needed. The in-betweens are calculated by interpolation. Another possibility to create the lip movements uses an animation system to create the expressions of the character one after the other in several passes. Each pass, like creating the lip movements for the letter 'a', can be created in real time, i.e. while playing back the speech signal. Several passes are needed to create and fine tune the final animation. With all of these techniques either a high technical effort is needed or they stay a very time consuming task. A technique that facilitates this process would be of great help for the central task of creating convincing animations of computer generated characters, i.e. synchronizing speech signal and speech movements. One possibility is synthesizing the speech signal and creating the lip movements synchronized to the synthesized signal [12]. This approach does not work if a high quality speech signal, spoken by a professional speaker, is needed or if the speech signal is already given as it is in our case. Another approach would be to recognize phonemes, for example in [10], and create the lip movements that are appropriate for each phoneme. This approach does not take into account coarticulation effects, i.e. the context of the particular phonemes, and will not work for character animation where the context of a phoneme is important for the expression. Also this approach may not catch any mood or special characteristics that are important for a specific character.

In this paper we present an automated technique for creating the lip movements of a computer-generated character from a given speech signal by using a neural net. The net is trained by a set of pre-produced animations. We first discuss the relevant aspects of facial animation and speech. Thereafter, we explain the processing of the speech signal and the network architecture and present some results.

FACIAL ANIMATION AND SPEECH

Presenting a 3D - animated face and a corresponding speech source implies a combination of two information channels. This collaboration of visual and auditory cues signifies to observe certain consequences. Research in psycholinguistics provides a general study of the problems coming up with this situation, leading to the bimodality in speech perception.

Bimodality of visual and acoustic Speech

Regarding the lip-reading capability of the deaf, it is well known that this form of speech perception uses information recoverable from visual movements of the various articulators like lips, tongue and teeth. Moreover it has been discovered that not only hearing impaired people utilize this visual information. Several investigations [3, 5] demonstrate that intelligibility of speech is enhanced by watching the speaker's face, especially if acoustic information is distorted by background noise, even for normal hearers. This bimodality of visual and acoustic information shows the necessity of coherence. Experimental results obtained from perceptual studies reveal the need of spatial, temporal and source coherence.

Spatial coherence
Concerning speaker localization we can see that vision is dominant on audition. Such a "capture of source" is widely used by ventriloquists, where the audience is much more attracted by the dummy whose facial gestures are more coherent with the speech signal than those of its animator. This demonstrates the capacity of humans to identify coherence in facial gestures and their corresponding acoustic production, which is developed even by four-to-five month old [11]. This capacity is frequently used by listeners in order to improve the intelligibility of a single person in a conversation group, when the well known "cocktail party effect" occurs.

Temporal coherence
Acoustically and optically transmitted information contains inherent synchrony. Experimentally Dixon and Spitz [4] observed that subjects are unable to detect asynchrony between visual and auditory presentation of speech when the acoustic signal was presented less than 130 ms before or 260 ms after the continuous video display of the speaker's face. But it was also found that these bounds lower to 75 ms (before) and 190 ms (after) in the case of a punctual event, such as a hammer hitting an anvil.

Source coherence
McGurk [8] found that the simultaneous presentation of an acoustic /ba/ and a visual /ga/ makes the listener or viewer perceive a /da/. This effect shows that audio and video must present the same information content.

To sum up these psycholinguistics results we get:

- Vision greatly improves speech intelligibility
- Concerning speaker localization, vision is dominant on audition
- Synchrony holds even when the channels are slightly time delayed
- Vision can bias auditory comprehension, as in the "McGurk effect"

Animation parameters

In addition to the outlined psycholinguistic findings we will have to compete with one of the major problems in character animation: what parameters should be used for a particular facial model. Requirements for choosing the parameters are:

- all necessary movements are covered
- the parameters are intuitively to handle
- the animator can overcome the complexity

Considering the evaluation of character animation from cartoons, we can get a first impression which basic parameters are necessary to "lip-sync" a character by carrying over to cartoons. Basically, the cartoon designer defines 4-7 keyframes for lip-sync, including three for central "mouth-open" vocals' /a/, /e/, /o/, and one for the group of "mouth-close" consonant's /b,m,p/ (Figure 1). The animator generates the in-betweens by graphic interpolation. Depending on the aspired quality of a character, it is necessary to define additional keyframes for refining the in-betweens.

Figure 1: Key expressions for lip-sync of a cartoon character

Visemes

Another approach to check out the basic animation parameters is to estimate visual similarities between different phonemes. Visually equal mouth positions for different phonemes are collected into classes of **vis**ual phone**mes**. These so called visemes are then used as animation parameters to get keyframes for all possible mouth positions. Fischer [6] classified visemes for English, Owens and Blazek [9] found 21 visemes for "General American". Fukuda and Hiki [7] revealed 13 visemes for Japanese, Alich [1] found 12 visemes for German and Benoît [2] found 21 visemes for French.

Coarticulation Effects

But another reminding effect is that speaking is not a process of uttering a sequence of discrete units. A key difficulty associated with connected speech is the effect of coarticulation and plays a great role in the possibilities for subjects to process visual information. A speech signal for any particular word is influenced by the words that go before and after it. Even the signal for a single phoneme is influenced in the same manner by surrounding phonemes. Additionally in context of a sentence intonation and depending on speech speed vocal reduction alter the speech signal. On the other hand coarticulation takes also place in visual movements of the articulators and occurs at least as much in visual as in acoustic speech. For both sources, context distorts the visual and acoustic pattern that would ideally be expected and complicates the production of visual and acoustic cues.
In cartoons this effect is overcome by individual graphic interpolation of the transitions by the animator. When using visemes as animation parameters, workarounds have to be created, like special transition visemes in certain contexts or tuning the bias of a viseme by a particular context. But visemes remain speaker dependent and complex to handle if someone wants to consider coarticulation effects.

REALIZATION

Our approach is to use signal processing techniques to extract relevant features from the speech signal and evaluate the mapping to the animation parameters via neural net. Training of the neural net is done with manually created animations.

The Face Model

The face model that is being used is based on a software environment developed at the Academy of Media Arts in Cologne, Germany. The software consists of several distinct modules for every single step of 3D character animation. Available steps are sculpturing, modeling and animation. The characters consist mainly of one coherent polygon mesh with normal vectors and texture coordinates attached to it's vertices. Eyes and teeth and optional accessories are modelled as separate objects. Motions and facial expressions are achieved by deformations of the polygon surface, i.e. displacing the vertex's coordinates while the net-topology is preserved. Individual facial expressions and motions are complex movements of sets of vertices. They are defined by clusters of vertex transformations, where each vertex has it's own translation and rotation about a chosen axis. The set of basic expressions and motions is designed by the animator during the modelling phase of the character. By mixing and scaling subsets of these basic motions, complex expressions can be generated. The intensity of each expression can be controlled by setting a parameter value within the range of 0 and 1. For the lip-sync task, the expressions that are important have to be chosen from the list of mouth expressions of the character. As animation parameters we chose the vowels /a/, /i/, /o,u/, the mouth closing consonants /b,m,p/ and /open/. The parameter /open/ accounts for general mouth opening that can not be expressed by the other parameters. The most important parameters are in fact the mouth closing (/b,m,p/) and a clear /a/ and /o,u/. If these parameters are not matched very well the user gets the impression of asynchrony immediately.

Processing Speech Data

Due to the 25 frames/sec refresh rate of the animation system, a speech signal context of 112 ms is extracted every 40 ms. This context is segmented via a hamming window function in 9 frames computed every 10 ms, resulting in segments of 32 ms length. Thereafter, these speech segments are analyzed by using a fast fourier (FFT) algorithm. Next, the logarithmically scaled vocal spectrum is summed up in 27 distinctive areas to extract a feature vector of the vocal spectra (Figure 2). This feature vector is then input in the neural network.

Network Architecture

The resulting feature vector applies to the input and the set of animation parameters determines the output for a neural net. The neural net is a 3-layer feed-forward net with back-propagation learning rule. It is build of 27 x 9 input neurons to match the extracted feature vector, 18 hidden and 5 output elements to match the 5 animation parameters. The number of hidden units was found by experimentation and extensive testing. We made several test runs with a different number of hidden units, whereas 18 units gave the best results and more units did not improve the result.

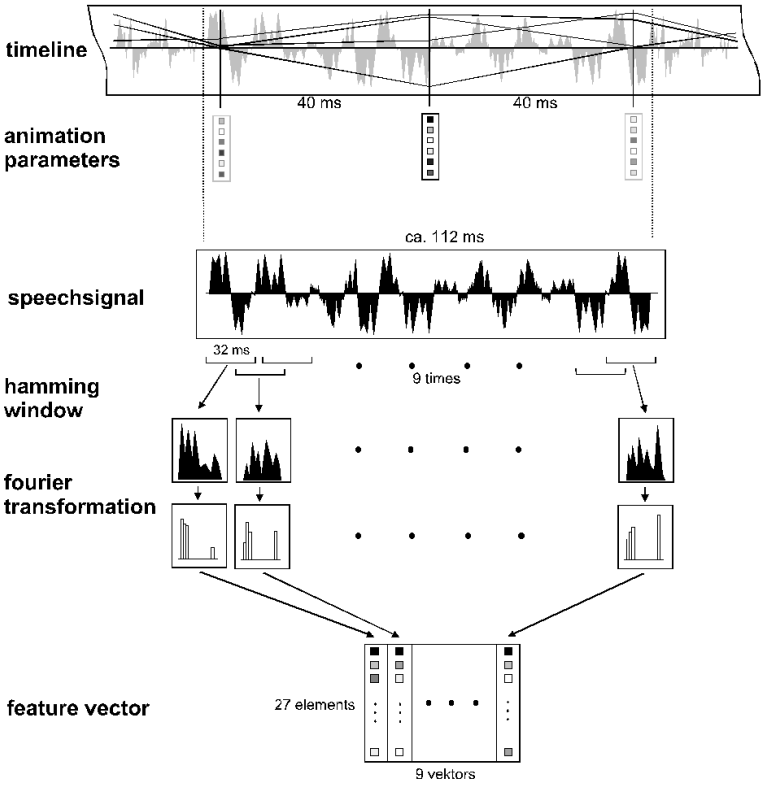

Figure 2: Schematic view of speech data processing

DISCUSSION / RESULTS

For testing the neural net it needs to be trained first. We used a 20 second sound sample which was animated manually using our character animation system. The system allows to output the chosen mouth expressions, i.e. for each animation parameter a value between 0 and 1 will be generated every 40 ms, which can then be used to train the net. Thereafter, we animated a 2 minutes sequence by presenting an unknown speech signal to the net and using the net output as animation parameters. In figure 3 part of the 2 minutes speech signal and the created animation parameters are shown. The labels indicate the corresponding animation parameter for each curve. The spoken text (in german) is written underneath the speech signal. Although the visual impression of the animation is rather convincing it is not perfect and manual correction is necessary. The correction took about 3 hours, which is a short period of time compared to 10 hours that would have been necessary when animating the hole sequence by hand.

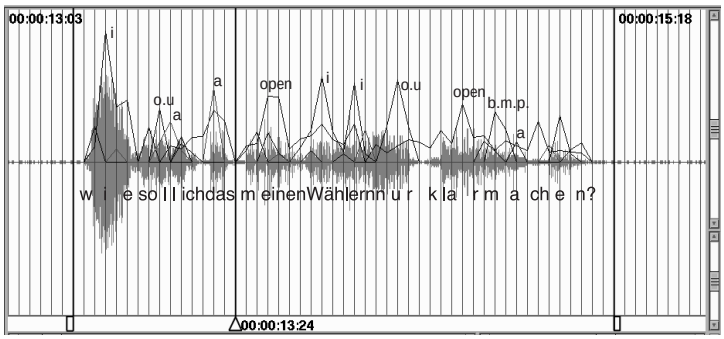

Figure 3: Animation parameters and speech signal

Problems

Some animation parameters like the vowel /i/ at the beginning of the sequence or the /u/ in the german word "nur" are correctly set by the net, other words, like "klar", are not represented correctly. There are several sources where these mismatches might stem from. Either the training sequence was not representative, or the speaker did speak in a different manner in the 2 minutes sequence than in the training sequence. An effect that often occurs with non-professional speakers (we achieved better results with professional speakers). The third source of error is the manual animation of the training sequence which has been animated by an professional animator but is not necessarily consistent concerning a spoken text to animation parameter match.

CONCLUSION

We presented a technique that facilitates the process of lip-synchronization for 3D-character animation. We use a neural net that was trained by a short sample sequence. The animation results given by the trained net with an unknown speech sample are not perfect but promising. They need manual correction but do reduce the production time to less than 30% compared to a complete manual approach. For further research we like to determine the sources of errors in more detail.

REFERENCES

[1] G. Alich, *Zur Erkennbarkeit von Sprachgestalten beim Ablesen vom Munde*, Dissertation Phil. Fak. Bonn, Germany (1961)

[2] C. Benoît, T. Lallouache, T. Mohamadi, C. Abry, *A Set of French Visemes for Visual Speech Synthesis*, In G. Bailly & C. Benoît (Eds.), Talking machines: Theories, Models and Designs Amsterdam: North Holland: 485-504 (1992)

[3] C. Benoît, *The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces*, Journal on Communications, 43; Scientific Society for Telecommunications, Hungary: 32-40 (Sep. 1992)

[4] N. F. Dixon, L. Spitz, *The Detection of Audiovisual Desynchronity*, Perception 9: 719-721 (1980)

[5] N. P. Erber, *Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli*, Journal of Speech & Hearing Research, 12: 423-425 (1969)

[6] G. G. Fisher, *Confusions amoung Visually Perceived Consonants*, Journal of Speech & Hearing Research 15: 474-482 (1968)

[7] Z. Fukada, S. Hiki, *Characteristics of the Mouth Shape in the Production of Japanese: Stroboscopic observation*, Journal of the Acoustical Societey of Japan 3: 75-90 (1982)

[8] H. McGurk, J. MacDonald, *Hearing Lips and Seeing Voices*, Nature, 264: 126-130 (1986)

[9] E. Owens, B. Blasek, *Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers*, Journal of Speech & Hearing Research 28: 381-393 (1985)

[10] A. Waibel et. al., *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Transaction son Acoustics, Speech, and Signal Processing, Vol. 37, No. 3, pp 328-339 (1989)

[11] D. H. Warren, R. B. Welch, T. J. McCarthy, *The role of visual-auditory compellingness in the ventriloquism effect: implications for transitivity among the spatial senses*, in Perception & Psychophysics, 30, 557-564 (1981)

[12] K. Waters, *DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces*, Imagina '95, pp 36-45 (1995)

Automated Lip-Sync for 3D-Character Animation

T. Frank, M. Hoch, G.Trogemann
Academy of Media Arts
Peter-Welter-Platz 2, D-50676 Cologne, Germany,
e-mail:{micha,georg}@khm.de

Keywords: facial animation, speech recognition, neural networks

ABSTRACT

A central task for animating computer generated characters is the synchronization of lip movements and speech signal. For real time synchronization high technical effort, which involves a face tracking system or data gloves, is needed to drive the expressions of the character. If the speech signal is already given, off-line synchronization is possible but the animator is left with a time consuming manual process that often needs several passes of fine tuning. In this paper we present an automated technique for creating the lip movements of a computer-generated character from a given speech signal by using a neural net. The net is trained by a set of pre-produced animations.

REFERENCES

[1] G. Alich, *Zur Erkennbarkeit von Sprachgestalten beim Ablesen vom Munde*, Dissertation Phil. Fak. Bonn, Germany (1961)

[2] C. Benoît, T. Lallouache, T. Mohamadi, C. Abry, *A Set of French Visemes for Visual Speech Synthesis*, In G. Bailly & C. Benoît (Eds.), Talking machines: Theories, Models and Designs Amsterdam: North Holland: 485-504 (1992)

[3] C. Benoît, *The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces*, Journal on Communications, 43; Scientific Society for Telecommunications, Hungary: 32-40 (Sep. 1992)

[4] N. F. Dixon, L. Spitz, *The Detection of Audiovisual Desynchronity*, Perception 9: 719-721 (1980)

[5] N. P. Erber, *Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli*, Journal of Speech & Hearing Research, 12: 423-425 (1969)

[6] G. G. Fisher, *Confusions amoung Visually Perceived Consonants*, Journal of Speech & Hearing Research 15: 474-482 (1968)

[7] Z. Fukada, S. Hiki, *Characteristics of the Mouth Shape in the Production of Japanese: Stroboscopic observation*, Journal of the Acoustical Societey of Japan 3: 75-90 (1982)

[8] H. McGurk, J. MacDonald, *Hearing Lips and Seeing Voices*, Nature, 264: 126-130 (1986)

[9] E. Owens, B. Blasek, *Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers*, Journal of Speech & Hearing Research 28: 381-393 (1985)

[10] A. Waibel et. al., *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Transaction son Acoustics, Speech, and Signal Processing, Vol. 37, No. 3, pp 328-339 (1989)

[11] D. H. Warren, R. B. Welch, T. J. McCarthy, *The role of visual-auditory compellingness in the ventriloquism effect: implications for transitivity among the spatial senses*, in Perception & Psychophysics, 30, 557-564 (1981)

[12] K. Waters, DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces, Imagina '95, pp 36-45 (1995)