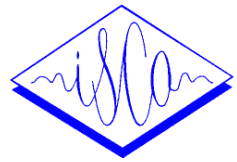

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

SUBJECTIVE EVALUATION FOR HMM-BASED SPEECH-TO-LIP MOVEMENT SYNTHESIS

Eli Yamamoto, Satoshi Nakamura, Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science & Technology
8916-5 Takayama, Ikoma, Nara 630-01, JAPAN
Tel:+81-743-72-5287, Fax:+81-743-72-5289
e-mail:{eli-y,nakamura,shikano}@is.aist-nara.ac.jp

ABSTRACT

An audio-visual intelligibility score is generally used as an evaluation measure in visual speech synthesis. Especially an intelligibility score of talking heads represents accuracy of facial models[1][2]. The facial models has two stages such as construction of real faces and realization of dynamical human-like motions.

We focus on lip movement synthesis from input acoustic speech to realize dynamical motions. The goal of our research is to synthesize lip movements natural enough to do lip-reading.

In previous research, we have proposed a lip movement synthesis method using HMMs which can incorporate a forward coarticulation effect and confirmed its effectiveness through objective evaluation tests.

In this paper, subjective evaluation tests are performed. Intelligibility test and acceptability test are conducted for subjective evaluation.

1. INTRODUCTION

In research of human perception, the integration of auditory and visual modalities has been investigated by tests of audio, visual, and audio-visual speech intelligibility. These intelligibility scores are evaluated by not only a natural human face but also a synthetic talking face. Especially the intelligibility score of talking heads represents the accuracy of the facial models . The accurate synthesis of talking heads needs to be elaborately re-synchronized with input acoustic speech signals, lip synchronization. The lip synchronization of talking heads are normally realized for the text-to-(audio-visual) speech systems, where the visual parameters for the lip movement synthesis are prepared in advance. However the visual parameters can be synthesized from the input acoustic speech signals using the correlation between auditory speech and visual facial parameters. This speech-to-lip movement synthesis can save the transmission bit rate of facial images in multimedia communication.

Thus we focus on lip movement synthesis to realize dynamical motions from input acoustic speech. The goal of our research is to synthesize lip movements natural enough to do lip-reading. If such elaborate lip motions can be synthesized, hearing impaired people may be able to recover auditory information by reading the visualized lip motion.

Mapping algorithms from acoustic speech signals to lip movements have been reported based on: Vector Quantization(VQ) [3][4] and Gaussian Mixtures [5] These methods require extensive training sets to account for context information. The required audio-visual data increases in proportion to the time length over the preceding or succeeding frames.

A different approach uses speech recognition technique, such as phonetic segmentation [6] and Hidden Markov Model (HMM) [7][8][9][10]. These methods convert the acoustic signal into lip parameters based on information such as a phonetic segment, a word, a phoneme, an acoustic event and so on. The HMM-based method has the advantage that explicit phonetic information is available to handle coarticulation effects caused by surrounding phoneme contexts. In speech recognition, the forward or backward coarticulation effects are generally dealt by using with biphone or triphone HMMs. However these biphone and triphone models require a large speech data and extensive training.

In previous work, we have proposed an HMM-

based lip movement synthesis method that is driven by a speech recognition process and that incorporates lip movement coarticulation effects. Moreover it has been verified that the proposed method is more effective than the conventional VQ method by the objective tests.

However, it is important to confirm the performance by not only objective evaluation but also subjective evaluation. Therefore subjective evaluation tests are performed in this paper. In addition to the intelligibility test, we adopt the acceptability test in line with subjective evaluation of audio speech synthesis.

2. SPEECH-TO-LIP MOVEMENT SYNTHESIS METHODS

We introduce the speech-to-lip movement synthesis methods before describing the subjective evaluation tests. We describe conventional VQ-based and the HMM-based methods, then describe the proposed context dependent HMM-based method.

2.1 VQ Method

The VQ-based method maps a VQ codeword vector of an input acoustic speech signal to visual parameters frame-by-frame.

The VQ-based method generates both audio and visual speech VQ codebooks by training data. Each audio VQ code is assigned to the output visual parameter VQ code by the training.

In the synthesis process, an acoustic codeword vector for input auditory speech is selected to minimize the distance between the acoustic speech parameter and an acoustic codeword vector of the acoustic speech VQ codebook. Then the output visual parameters are retrieved as a visual speech VQ code associated with the input acoustic codeword vector.

The visualized lips images are reconstructed by the visual parameters.

2.2 HMM Method

The HMM-based method maps from an input acoustic speech signal to lip parameters through HMM states determined by the Viterbi alignment. The Viterbi alignment assigns an input frame to the optimal HMM state by maximizing the likelihood of input acoustic speech. In this paper, phoneme HMMs are used so that the combination of phoneme HMMs can produce any utterance sequence.

The HMM-based method is composed of two processes: a decoding process that converts an input acoustic speech signal to a most likely HMM state sequence by the Viterbi alignment and a table look-up process that converts an HMM state to corresponding lip parameters. The lip parameters for each HMM state in the look-up table are also trained using the Viterbi alignment.

2.3 SV-HMM Method

we have proposed a new HMM-based method taking into account the succeeding viseme context, which is called the Succeeding-Viseme-HMM-based method (SV-HMM-based method).

This proposed HMM-based method produces visual parameters through incorporation of coarticulation information. In speech recognition, coarticulation effects are generally dealt using HMMs. Biphone or triphone HMMs are modeled as depending on preceding and succeeding phoneme contexts. However, the number of the biphone or triphone HMMs is the squared or cubed number of the monophone HMMs so that the training for all models requires large amounts of acoustic speech data. Moreover, collection and preparation of such large corpora of synchronized audio-visual data is quite expensive.

On the other hand, the proposed method continues to use the monophone HMMs of context independent models but synthesizes visual parameters with context dependency. The proposed method generates context dependent visual parameters by looking ahead to context independent HMM state sequence.

Although coarticulation effects are bidirectional, the focus in this paper is on forward coarticulation because estimation errors in the HMM-based method are affected more by forward than backward coarticulation.

Moreover, the use of visemes reduces the number of context dependent lip parameters. Visemes are defined by the distinguishable postures of the visible mouth articulation, associated with speech production. The postures of the mouth in the same places of articulation can not be distin-

guished. Visemes of succeeding phonemes have a strong coarticulation effect on current visual parameters.

The training algorithm of the SV-HMM-based method differs from the HMM-based method with context independent lip parameters only in its use of the viseme classes of succeeding phonemes. The training and synthesis steps are as follows.

Training Algorithm

1. Prepare and parameterize data of synchronized acoustic speech signals and 3D lip position data.
2. Train the acoustic phoneme HMMs using the training acoustic speech parameters.
3. Align the acoustic speech parameters into HMM state sequences using the forced Viterbi alignment.
4. Classify all frames into viseme classes by looking ahead to succeeding phoneme contexts.
5. Average synchronous lip parameters for the frames associated with the same HMM state and the same viseme class of the succeeding phoneme context.

Synthesis Algorithm

1. Align an input acoustic speech parameters into an HMM state sequence using the Viterbi alignment.
2. Determine the viseme classes of the succeeding phoneme contexts at each frame.
3. Retrieve the output lip parameters associated with the HMM state and the viseme class.
4. Concatenate the retrieved lip parameters as a lip movement sequence to synthesize the visualized lips.

3. SUBJECTIVE EVALUATION EXPERIMENT

In the previous section, the Euclidean error distance gave an objective distance between the synthesized lip parameters and the original lip parameters. However, it is not certain whether these values reflect the error perceived by a viewer. Therefore we performed subjective tests to evaluate the quality of the synthesized lip motion. In the field of speech processing, standard subjective testing consists of two procedures: the intelligibility test and the acceptability test. The intelligibility of speech test quantifies the amount of transmitted information perceived by the subjects. The acceptability of speech test deals with perceiver judgements of naturalness. Both of these preliminary tests were used to assess the quality of the lip movement synthesis.

3.1. Intelligibility Test

In this paper, the intelligibility score is defined as the percentage of all syllables that were identified. The test was performed by four normal Japanese subjects unfamiliar with lip-reading.

3.1.1 Method

Subjects

Ten young adults participated in the experiment. The subjects were all native speakers of Japanese, and they have normal hearing and normal or corrected vision. They were volunteer students of NAIST.

Stimuli

The stimuli were CVCV nonsense Japanese words. They are selected out of the 102 disyllabic nonsenese words list, which is reported by the project which aims to standarize the evaluation of the intelligibility of Japanese synthetic speech[12]. The second syllable in the list is selected from all the Japanese syllables whose vowels /a/,/i/,and /u/. The vowel of the first syllable in the list was the same as the one of the second syllable except for eleven C/a/C/i/ syllables. The lip movement synthesis methods are evaluated using the second consonant of the disyllable, which have been verified to reflect well the intelligibility of the normal sentence[13].

The 102 nonsense words were recorded by a video camera for this subjective evaluation. The other 326 word audio-visual data for construction of HMMs and VQ codebooks was recorded by the 3D position sensor system. The data consists of 3D positions of the markers.

The test audio data was digitized at 12kHz with 16 bit resolution. The length of each word is almost 2 second.

The trial words are conducted considering the distinction of the visual lip shape out of the non-

sense words list. Audio-visual perception studies for Japanese show that Japanese speakers rely less on visual speech information than do English speakers[14][15]. However another report describes that the intelligibility of only the bilabial consonants show 90% even in the Japanese. Moreover Sekiyama et al.[16] have investigated the consonant ranking to do easy lip-reading. The result shows the consonant /w/,/p/,/h/,/py/, /m/ are easy to be recognized.

We selected the 9 stimuli out of the 102 words list for the subjects. The consonants are selected from three category of /w/, the bilabial consonants, and the other consonants except bilabial consonants unaccustomed lip-reading.

The synthesized visual parameters are visualized by the ICP 3D lip model[11]. The orientation of the lips was fixed frontally. The lips were synthesized at 125Hz. The actual output 25Hz frames were constructed by lip parameters averaged in every five frames.

Experimental Design

The presentation is classified as the five synthesis methods ; lip image sequence from natural human visual parameter, synthetic VQ, synthetic HMM, synthetic SV-HMM visual parameters. In the intelligibility test, the stimulus by the HMM and SV-HMM methods are prepared without the forced Viterbi alignment.

The noise comtaminated speech is presented for the purpose of degrading the estimation from audio speech. White Gaussian noise was added to degrade the acoustic speech signal. The Signal-to-Noise Ration are selected the identical conditions to the ICP intelligibility tests[11], $-18, -12, -6, 0, 6$dB. Then another SNR $-\infty$ is added for evaluating visual only speech synthesis.

For presentation, the trial words are repeated by the synthesis methods and Signal-Noise-Ratio. Therefore the words arranged in random order.

Procedure

Subjects viewed the stimuli on a color monitor. Stimulus presentation and response collection were controlled by a computer. Subjects were given an audio-visual stimulus by pushing Enter or Space keys by themselves. They were asked to look at visual speech and listen without concentrating to the audio utterances.

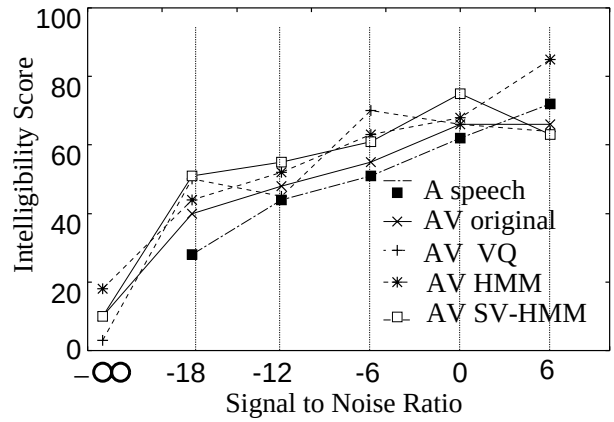

Figure 1: Result of the Intelligibility Tests

3.1.2 Results

Fig. 1 shows the /C-V/ intelligibility scores for the original and the SV-HMM-based method at different levels of acoustic degradation. The auditory and visual stimuli synthesized from the original lip position and using the SV-HMM-based method gave higher intelligibility scores than the purely auditory stimulus. This suggests that both natural and model-based lip synthesis enhanced intelligibility.

3.2. Acceptability Test

3.2.1 Method

In the acceptability test, we investigate the differences among the synthesis methods by focusing on the naturalness of the synthesized lip movement. The indicators of the naturalness may be the smoothness of synthesized lip parameters or the synchronization between the lip shapes and the audio speech. The acceptability score is evaluated by the criterion whether the synthesized lip movement is as natural as a human lip motion.

Subjects

The subjects for the acceptability tests were used the identical group of intelligibility tests.

Stimuli

The test utterances were composed of 3 Japanese words out of the 3D position data. The words were randomly selected out of 100 test words. The

3 words are presented two times randomly

The synthetic visualized lips were created using the same ICP 3D lip model software.

Experimental Design

The presentation is classified as the five synthesis methods ; lip image sequence from natural human visual parameter, synthetic VQ, synthetic HMM with forced Viterbi alignment, synthetic HMM, synthetic SV-HMM with forced Viterbi alignment, synthetic SV-HMM visual parameters.

In the same manner as the intelligibility test, the presentation order of the words was chosen at random. To imagine the human lip motion, the clean acoustic signal is provided to the subjects in addition to the visualized lips.

The Mean Opinion Score (MOS) (IEEE Recommendation, 1969) was used as the measure to evaluate the naturalness of the synthesized lip movement. The MOS is the most widely used subjective quality measure for evaluating telephone systems and speech coding algorithms. The subjects assigned scores on a five point scale defined by Japanese category labels analogous to excellent, good, fair, poor, and bad.

Procedure

Subjects viewed the stimuli on a color monitor. Stimulus presentation and response collection were controlled by a computer. Subjects were given an audio-visual stimulus by pushing Enter or Space keys by themselves. In evaluating the lip movement synthesis, subjects were instructed that the displayed lip movement should be marked excellent when it had natural enough to be human lip movement.

3.2.2 Results

Fig.2 shows the mean opinion scores. Scores for the VQ-based method, the HMM-based method, and the proposed SV-HMM-based method were 2.97, 3.42, and 3.17, respectively. However, the difference in mean MOS over the synthesis methods shows no significance with the F-test. The result might be caused by the very small number of test utterances, and by an insufficient number of subjects. Moreover the visualized lip model is somewhat limited, as it uses a cartoon animation whose parameters are fine-tuned to a specific speaker. In the future, the acceptability test will be further investigated while improving the data and the visualization tool.

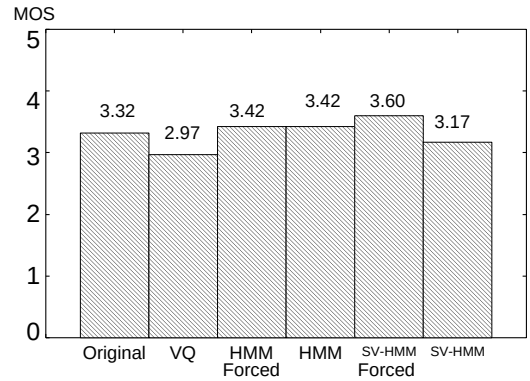

Fig. 2: Result of the Acceptability Tests

4. CONCLUSION

In this paper, subjective evaluation tests are performed for speech-to-lip movement synthesis. Intelligibility test and acceptability test are conducted for subjective evaluation.

The subjective evaluation did not show the effect of the proposed method clearly; this will be investigated under better experimental conditions in the future.

In synthesis, the HMM-based method has the intrinsic difficulty that the precision of lip movement synthesis depends upon the accuracy of a Viterbi alignment. The Viterbi alignment deterministically assigns a single HMM state for each input frame. Incorrectly decoded frames of the HMM state sequence may give rise to wrong lip shapes. This problem could be solved by extending the Viterbi algorithm to the Forward-Backward algorithm, which can take all the HMM state sequence probabilities into account.

5. REFERENCES

1. Massaro,D.W.,"Perceiving Talking Faces", MIT Press (1997).
2. Le Goff,B.,Guiard-Marigny,T.,Benoit,C., "Analysis-Synthesis and Intelligibility of a Talking Face", *in "Progress in Speech Synthesis", J. van Santen et al.,Eds, Springer-Verlag* (1996).
3. Morishima, S. and Harashima, H.: A Media Conversion from Speech to Facial image for Intelligent Man-Machine Interface, *IEEE Journal on sel. areas in Communications*, Vol. 9, No. 4, pp. 594–600 (1991).
4. Lavagetto, F.: Converting Speech into Lip Movements: A Multimedia Telephone or Hard of Hearing People, *IEEE Trans. on Rehabilitation Engineering*, Vol. 3, No. 1, pp. 90–102 (1995).

5. Rao, R.R. and Chen,T., "Cross-Modal Prediction in Audio-Visual Communication", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.4, pp.2056-2059(1996)
6. Goldenthal,W., Waters,K., Van Thong,J.M. and Glickman,O. "Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe!", Eurospeech'97 Proceedings,Vol.4, pp.1995-1998 1997)
7. Simons, A. and Cox, S.: Generation of Mouthshape for a Synthetic Talking ead, *Proc. of the Institute of Acoustics*, Vol. 12, No. 10 (1990).
8. Chou, W. and Chen, H.: Speech Recognition for Image Animation and Coding, *ICASSP 95*, pp. 2253–2256 (1995).
9. Chen, T. and Rao, R.: Audio-Visual Interaction in Multimedia Communication, *ICASSP 97*, pp. 179–182 (1997).
10. Yamamoto, E., Nakamura, S. and Shikano, K.: Speech-to-Lip Movement Synthesis by HMM, *ESCA Workshop of Audio Visual Speech Processing*, pp.137–140(1997).
11. Guiard-Marigny, T., Adjoudani, T. and Benoit, C.: 3D Models of the Lips and Jaw for Visual Speech Synthesis, *in "Progress in Speech Synthesis", J. van Santen et al.,Eds, Springer-Verlag* (1996).
12. Speech Input/Output Systems Expert Committee, *"Commentary on the Guideline of Speech Synthesizer Evaluation"*, Committee on Standardization of Human-Media Information Processing, the Japan Electronic Industry Development Association (1997)
13. Kasuya,H.and Kasuya,S.(1992), "Relationships between syllable, word and sentence intelligibility of synthetic speech", *" Proc. Int'l Conf. Spoken Language Processing"*, Vol.2,pp.1215-1218.
14. Sekiyama,K.(1994),"Differences in audio-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility",*J. of the Acoustic Society of Japan*,Vol.15,pp.143-158.
15. Sekiyama,K.and Tohkura,Y.(1991),"McGurk effect in non-English listeners: Few visual affects for Japanese subjects hearing Japanese syllables of high auditory intelligibility",*J. of Acoustic Society of America*,Vol.90,pp.1797-1805
16. Sekiyama,K.and Joe,K.and Umeda,M.(1987), "Perceptual components of Japanese syllables in lipreading:A multidimensional study (English Abstract)",IEICE Technical Report IE,pp.29-36(1991)