SPEECH RECOGNITION FOR IMAGE ANIMATION AND CODING

Wu Chou

Rm 2D-526
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill,NJ 07974, USA
e-mail: wuchou@research.att.com

Homer H. Chen

Rm. 4C-522
AT&T Bell Laboratories
Crawfords Corner Road
Holmdel, NJ 07733, USA
e-mail: homer@big.att.com

ABSTRACT

In this paper, we discuss some issues related to acoustic assisted image coding and animation. An approach of talker independent acoustic assisted image coding and animation scheme is studied. A perceptually based sliding window encoder is proposed. It utilizes the high rate (or oversampled) viseme sequence from the audio domain for image domain viseme interpolation and smoothing. The image domain visemes in our approach are dynamically constructed from a set of basic visemes. The look-ahead and look-back moving interpolations in the proposed approach provide an effective way to compensate the mismatch between auditory and visual perceptions.

1. INTRODUCTION

Lip reading is an ancient art and is vital for hearing impaired people to understand verbal communication. Various approaches were proposed in the past to use visual information to augment speech recognition[1]–[4]. In this paper, we study the inverse problem in which the acoustic information is used to assist image coding and animation. This is motivated by application in two different areas: visual communication and animation. In low bit rate coding of facial images, the most active parts in the image are around the mouth area, and many bits are sent to update the changes in mouth shape, chin position, and other parts of the facial movements which are related to the action of verbal conversation. In animated movie and video games, alignment of the animator's voice with the character's mouth shape has been considered one of the most difficult and time consuming process. The difficulty in these applications lies in the fact that the image sequence must follow the acoustics of the talker and obey the basic physical laws of human articulator, which includes mouth shape, tongue position, facial muscle tension and so forth.

The idea of using speech for image coding and animation has been studied for some years [5][10] where the facial images are segmented into visemes. A viseme is a sequence of oral-facial movements (shapes) which often relates to uttering some linguistic based units such as phonemes. During the process of encoding, the viseme sequence is identified and transmitted as side information through the channel. At the receiver end, the image is reconstructed through an analysis-by-synthesis scheme from a stored or transmitted viseme inventory. This approach can lead to significant bit rate savings, since the bit rate for transmitting the side information regarding viseme identities is extremely low. In fact, this overhead can be eliminated if the viseme information can be robustly extracted from the decoded speech signal at the decoder. In some image animation applications, the entire talking scene can be animated from one single snapshot of the character. This is also attractive for some communication applications where the capacity of storage media is limited. The image sequence is, therefore, driven by the animator's voice. Thus, a higher level animation, such as providing a means of defining actors, is made possible.

Each viseme is associated with two different identities, one in the image domain and one in the acoustic domain. A traditional approach is to form two codebooks, and the viseme identity is determined by averaging the distorsions from the two codebooks. Although this approach is simple to implement, its capability is limited. In addition, averaging distortions in the two domains may not lead to a perceptually good representation in the coded image. As a matter of fact, one of the fundamental obstacles in this approach is that the acoustic and visual identity of the viseme may not match. The acoustic identity of the viseme is determined by the physical attributes of the human articulator, and the image identity of the viseme is determined by human visual perception, which often relates to a much slower frame rate. In fact, human speech often exhibits a much greater degree of variabilities, as illustrated in the change of speaking rate, pitch period and etc. Direct clustering the feature vectors generated by LPC-analysis into a VQ codebook often falls short to capture the dynamic variations in human speech, and the codebook is hard to be speaker independent.

In this paper, we propose an approach which can easily be made speaker independent and be real-time in a frame synchronous fashion. This is achieved by introducing more robust stochastic acoustic modeling for visemes in acoustic domain. The baseline viseme models are morphed through the use of a 3-D wire frame facial model and a knowledge based morphing scheme. It integrates acoustics and the physiological knowledge in coding the portion of the images

0-7803-2431-5/95 $4.00

related to human articulator. In addition, the proposed approach is vocabulary independent and does not require any human supervision, a feature which is attractive for many applications.

2. ACOUSTIC AND IMAGE IDENTITY OF VISEMES

Visemes are basic facial image units which relate to human articulatory actions in conversational speech. Like phonemes in speech, visemes are sometimes characterized as visual phonemes in lip reading literatures. Visemes in the image domain often stand for some distinct facial image sequence which can be identified visually. The number of basic visemes in English is around 10 [7]. However, visemes in acoustic domain are often short acoustic events typically corresponding to the acoustic events at the phoneme or sub-phoneme level. Therefore, a fine temporal resolution is needed for viseme identification in the audio domain.

There are several ways to increase the number of visemes from the basic viseme set. One way is to include short viseme sequence as new viseme units. One example of this approach is to include tri-visemes which are sequence of 3 visemes corresponding to the acoustic events of uttering CVC (consonant-vowel-consonant) sequences. The advantage of this approach is that the longer viseme units are often acoustically more stable and the image sequence for such viseme units can be concatenated at more stable anchor point, such as at the middle of the vowel where the position of the mouth is more reliable. However, viseme sequence units, such as tri-visemes, often correspond to a much longer image sequence. Real-time response of this approach is a problem. Moreover, in addition to the increased acoustic domain complexity, the image domain complexity is also greatly increased because more image domain viseme prototypes are needed to reduce the image distorsions due to their long time duration.

The other way to increase the number of visemes is to introduce context dependency where visemes with different surrounding viseme context are treated as new visemes. This is an approach similar to introducing context dependent tri-phone models in speech recognition. The image transition between different mouth shapes can be modeled more accurately. However, it requires more delicate training process in both acoustic and image domains. System complexity is also increased substantially. In addition, the transition region of the mouth shapes are very fuzzy. How to concatenate visemes at these unstable regions in the image domain deserves special attention.

In order to have a real-time response, a set of visemes corresponding to phoneme or sub-phoneme acoustic events is selected in our study. Introducing visemes corresponding to short acoustic events can reduce the complexity in viseme representation at the image domain, but requires more dedicated modeling schemes in the acoustic domain. Table 1 illustrates the mapping between phonemes and the visemes used in our study. Twelve basic visemes are selected and each phoneme is represented by one or two visemes. A phoneme corresponding to one viseme is denoted with a zero second coordinate. Instead of introducing cross viseme context dependency, as typical in speech recognition, visemes in our approach are associated with certain acoustically distinct events in which these visemes are presented. For example, phonemes "aa" and "ay" are mapped to the same viseme 2, but two instances are modeled separately to account for acoustic dissimilarity.

In acoustic domain, viseme instances are modeled statistically by hidden Markov models (HMMs). The acoustic identity of each viseme and its instances can be obtained directly by modeling the corresponding acoustic segments in the audio signal or by mapping each viseme to certain articulatory actions of uttering some linguistic based units. Most HMMs in our study consist of 3-states, and the observation probability distribution in each state is a mixture of Gaussians. The use of acoustic instance for visemes is a way to bypass the problem of differentiating those short sub-phoneme level visemes which do not possess stable identities in the acoustic domain. These short sub-phoneme level visemes are being associated with the acoustic differentiable instances, such as uttering some phonemes. Each state of HMM can be viewed as representing some sub-phoneme level acoustic units. The sub-phoneme level viseme instances can, therefore, be mapped into certain states of HMMs representing associated acoustic instances. For example, the acoustic instance of phoneme "oy", as in boy, is represented by two visemes (4, 3). Viseme 4 is mapped into the first state and viseme 3 is mapped to last two states of the HMM representing the acoustic instance of uttering "oy".

The audio signal in our approach is bandlimited to 4KHz, and a 10-th order LPC analysis is performed. The speech signal is converted into a feature vector sequence consisting cepstrum, delta cepstrum and delta-delta cepstrum features [9]. The viseme identification is performed through a frame synchronous Viterbi decoding algorithm which generate a viseme sequence with a temporal resolution corresponding to 100 frames per second. This audio rate is much faster than the required video frame rate ($15 \sim 30$ frames per second), but is needed for viseme identification in the audio domain.

In the image domain, visemes are represented as parameters which control a 3-D wire frame facial model. Fig. 2 shows an illustrative example of a 3-D wire frame facial model comprising a lattice of approximately 500 polygonal elements, of which approximately 80 are used to represent the mouth shape. The 3-D wire frame model is manipulated to express facial motions by controlling the lattice points of the wire frame. This operation is called structure deformation. It is unnecessary to control all of the lattice points on the 3-D wire frame independently, because motion of one lattice point influences neighbouring lattice points. Accordingly, the six lattice points corresponding to the six feature points shown in Fig. 3 are used as the control points for the mouth. With these control points, a sequence of mouth movements can be generated by deforming the 3-D wire frame according to the viseme parameters. A texture mapping operation is also performed in the image domain. It projects and maps a stored or received surface texture onto the 3-D wire frame image for each video frame to create an animated video sequence.

3. VISEME INTERPOLATION AND SMOOTHING

Visemes as basic image coding units are derived from two vastly different perceptual domains. Auditory perception and visual perception may not match each other. This problem becomes even more acute in spontaneous speech. The mouth of the talker often moves before the actual words or phrase being uttered, and continues to move even the mouth has stopped voicing. This phenomenon is often called *anticipation*. In addition, the frame rates for visemes in audio and video domains are also different and subject to different constraints.

To address these fundamental issues, a perceptual knowledge based sliding window encoder is developed according to the visual perception of visemes in the image domain and the physiological acoustic rules of visemes in the audio domain. The faster frame rate from audio domain can be used to interpolate and smooth the visemes in the image domain. It can look ahead several frames in the audio domain between two video frames. It comprises an integrated process for rate conversion, image smoothing, and error correction.

A diagram of the sliding window encoder is enclosed in Fig. 1. To account for the anticipation effects, a moving average interpolation scheme is employed. The mouth parameters of the current viseme is the normalized moving average between the last viseme, the current viseme and the one ahead in the future. Because of the look-ahead and look-back moving interpolations, mouth shape will start forming before the voice actually starts and will gradually close after the voice stops. Moreover, in this approach, new viseme prototypes are automatically generated for regions where the anticipation effects are dominant and the auditory perception and visual perception do not match.

The physiological acoustic knowledge of speech is useful in our approach. Errors in the decoded viseme sequence can be corrected or smoothed according to the physiological acoustic rules. Jerky jumps from mouth close to complete open within two video frames will be limited and smoothed. It should be noted that visemes in our approach still maintain a close relation to linguistic based units where movements of human articulator in these cases are well studied.

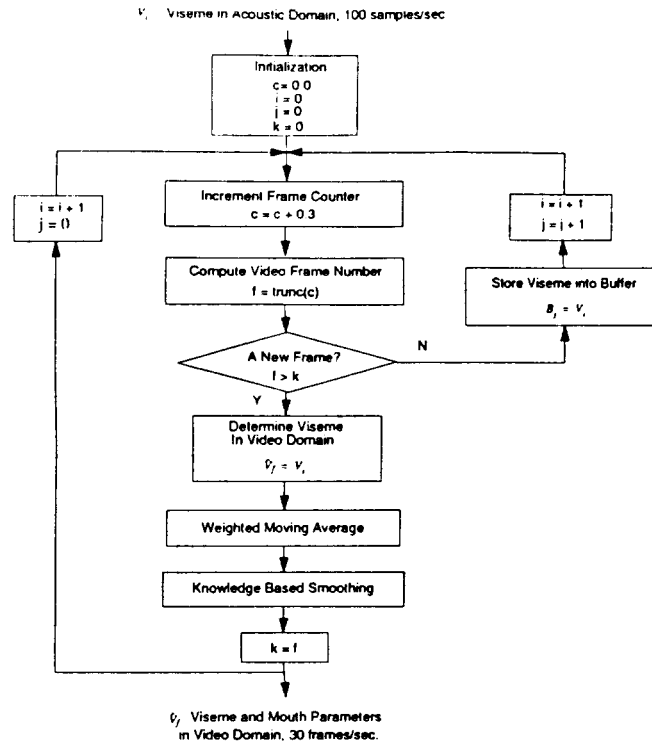

Figure 1: Diagram of the sliding window encoder

4. EXPERIMENTAL RESULTS

Two schemes under the proposed approach were studied. One scheme assumes that the word or ASCII transcription of the speech is known. This scheme requires a listener to transcribe the speech. The other scheme is based on automatic viseme sequence identification driven by the acoustics of the audio and does not require transcriptions. Both schemes are able to process free speech in live conversation. There is no limitation on the size of the vocabulary, the speaker and etc. The acoustic models for the visemes in our experiments are obtained from a speaker independent telephone data base. We have tested both schemes on a 30-second recorded audio-visual sequence *"grandma"*. Comparing the results of these two schemes, we found that the image sequence generated by the automatic viseme identification scheme is perceptually more favorable than the other one. One reason of this phenomenon may come from the fact that in live speech, many phonemes in the word sequence are omitted or merged with neighbouring dominant phonemes. Consequently, packing all phonemes in the word sequence into images can result in jerky and unnatural mouth movements, making the image sequence perceptually objectionable. The automatic viseme identification scheme follows the acoustics of the speaker and is not constrained by the rigid pronunciation rule from the dictionary. In addition, it does not require any human supervision, which is an attractive feature for many applications.

5. SUMMARY

In this paper, we discussed some issues related to acoustic assisted image coding and animation. An approach of talker independent acoustic assisted image coding and animation scheme is studied. A perceptually based sliding window encoder is proposed. It utilizes the high rate (or oversampled) viseme sequence from audio domain for image domain viseme interpolation and smoothing. The image domain visemes in our approach are dynamically constructed from a set of basic visemes. The look-ahead and look-back moving interpolation in the proposed approach provides an effective way to compensate the mismatch between auditory and visual perceptions.

Acknowledgement The authors would like to thank Fred Juang, Barry Haskell, and Paul Henry for their en-

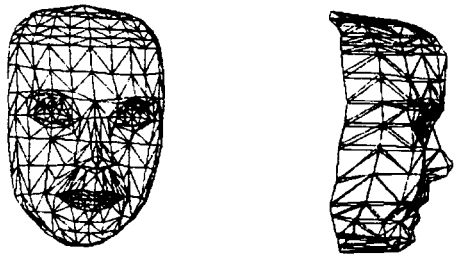

Figure 2: Diagram of 3-D wire frame model

h#	(1,0)	er	(6.0)	p	(8,0)
aa	(2,0)	ey	(2,0)	r	(6,0)
ae	(3,0)	f	(10.0)	s	(6.0)
ah	(6,0)	g	(6,0)	sh	(6,0)
ao	(5,0)	hh	(6,0)	t	(6,0)
aw	(2,7)	ih	(2,0)	th	(13,0)
ax	(3,0)	ix	(12,0)	uh	(5,0)
axr	(3,6)	iy	(3,0)	uw	(5,0)
ay	(2,0)	jh	(6.0)	v	(10.0)
b	(8,0)	k	(6,0)	w	(7,0)
ch	(6,0)	l	(9,0)	y	(6,0)
d	(6,0)	m	(8,0)	z	(6,0)
dh	(9,0)	n	(6,0)	zh	(6,0)
eh	(2,0)	ng	(12,0)	dx	(6,0)
el	(9,0)	ow	(4,0)	nx	(3,0)
en	(6,0)	oy	(4,3)		

Table 1: Visemes mapping table

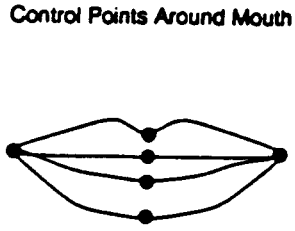

Figure 3: Control points around mouth

couragement and support of this work, Tsuhan Chen, Eric Petajan and Qiru Zhou for their helpful discussion, Kiyo Aizawa for his 3-D face modeling software, and S. Keshev for his viseme table.

6. REFERENCES

[1] Nishida, "Speech Recognition Enhancement by Lip Information," *ACM SIGHI Bulletin 17, no. 4 (1986), pp. 198–204.*

[2] S. Steven, "Computer Lip Reading to Augment Automatic Speech Recognition", Speech Technology (1989) pp. 175–181.

[3] B. Christoph, H. Hild, S. Manke and A. Waibel, " Improving Connected Letter Recognition by Lipreading", Proc. ICASSP93 vol. 1, pp. 557–560.

[4] G. Oscar, A. Goldschen and E. Petajan, "Feature Extraction for Optical Automatic Speech Recognition or Automatic Lipreading", GWUIIST-92, November 1992.

[5] S. Morishima, K. Aizawa and H. Harashima "An Intelligent Facial Image Coding Driven by Speech Phoneme", Proc. ICASSP88, pp1795–1798.

[6] R. Forchheimer and T. Kronander, "Image coding – From waveforms to animation ", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 37, no. 12, Dec. 1989, pp. 2008-2023.

[7] K. W. Berger, "Speechreading: Principles and Methods" National Education Press, 1972

[8] W. Chou, B.-H Juang and C.-H. Lee, "Segmental GPD Training of an Hidden Markov Model Based Speech Recognizer", Proc. ICASSP92 pp. 473–476, 1992

[9] C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini and A.E. Rosenberg, "Improved Acoustic Modeling for Speaker Independent Large Vocabulary Continuous Speech Recognition", Computer Speech and Language, pp. 103-127, 1992.

[10] A. Lipman, "Semantic Bandwidth Compression: Speechmaker" Picture Coding Symposium, pp. 29-30, 1981.