COMPARISON OF PHONEME AND VISEME BASED ACOUSTIC UNITS FOR SPEECH DRIVEN REALISTIC LIP ANIMATION

Elif Bozkurt[1], *Çiğdem Eroğlu Erdem*[1], *Engin Erzin*[2], *Tanju Erdem*[1], *Mehmet Özkan*[1]

[1] Momentum A. Ş.
TÜBİTAK – MAM – TEKSEB, A-205, Gebze, Kocaeli, Turkey
E-mail: {ebozkurt, cigdem.erdem, terdem, mozkan}@momentum-dmt.com
[2] Electrical and Electronics Engineering Dept., Koç University, İstanbul, Turkey
E-mail: eerzin@ku.edu.tr

ABSTRACT

Natural looking lip animation, synchronized with incoming speech, is essential for realistic character animation. In this work, we evaluate the performance of phone and viseme based acoustic units, with and without context information, for generating realistic lip synchronization using HMM based recognition systems. We conclude via objective evaluations that utilization of viseme based units with context information outperforms the other methods.

***Index Terms*—** Lip animation, Hidden Markov Models, character animation, viseme recognition, lip synchronization

1. INTRODUCTION

Natural looking lip animation is a very important and challenging problem for realistic character animation in computer graphics. Humans are very sensitive to the slightest glitch in the animation of the human face. Therefore, it is necessary to achieve realistic lip animation, which is synchronous with a given speech utterance.

There are methods in the literature for achieving lip synchronization based on audio-visual systems that correlate video frames with acoustic features of speech [1]. A major drawback of such systems is the scarce source of audio-visual data for training. Other methods use text-to-speech synthesis, which utilize a phonetic context to generate both speech and the corresponding lip animation [2], [3], [4], [5]. However, current speech synthesis systems sound slightly robotic, and adding natural intonation requires more research. If the lip synchronization is generated using speech uttered by a real person, the animation will be perceived to be more natural. In such systems, a phonetic sequence can be estimated directly from the input speech signal using speech recognition techniques [6], [7], [8].

This paper focuses on the limited problem of automatically generating phonetic sequences from pre-recorded speech for lip animation. The generated phonetic sequence is then mapped to a viseme sequence before animating the lips of a 3D head model, which is built from photographs of a person [9]. Note that, a viseme is the corresponding lip posture for a phoneme, i.e. visual phoneme.

In this work, we experimentally compare four different acoustic units within HMM structures for generating the viseme sequence to be used for synchronized lip animation. These acoustic units are namely phone, tri-phone, viseme and tri-viseme based units. First, we define a viseme set and the associated lip shapes. Then, we present a viseme sequence generation method, which uses an HMM trained using viseme based acoustic units. The viseme based HMM is easier to train and also gives smaller number of classes with respect to the phone-based HMM method.

The rest of this paper is organized as follows. In Section 2, the details of the HMM structures for each acoustic unit are given. In Section 3, training of the HMMs is explained and objective test results for lip animation are provided. Finally, conclusions are discussed in Section 4.

2. SPEECH DRIVEN VISEME RECOGNITION

Our lip animation method is based on 16 distinct viseme classes as defined in Table 1. After the generation of the 3D head model [9], a graphic artist defines the mouth shapes for the 16 visemes using a graphical user interface. Sample visemes corresponding to various phoneme classes are shown in Figure 1. Since speech has both an auditory and a visual component [10], it is very important that the definitions of visemes are done accurately. These mouth shapes (visemes) are properly interpolated (smoothed) during the actual animation.

In order to synthesize viseme sequences, it is possible to construct HMM models using different parameters and language models. In this work, we compare four different HMM structures, which are described below. The phone based and tri-phone based HMM structures try to recognize the phoneme sequence from a given speech utterance. Then

1-4244-0722-2/07/$20.00 ©2007 IEEE

the recognized phoneme is mapped to one of the 16 viseme classes as shown in Table 1. On the other hand, the viseme and tri-viseme HMM structures try to recognize the viseme classes directly from speech.

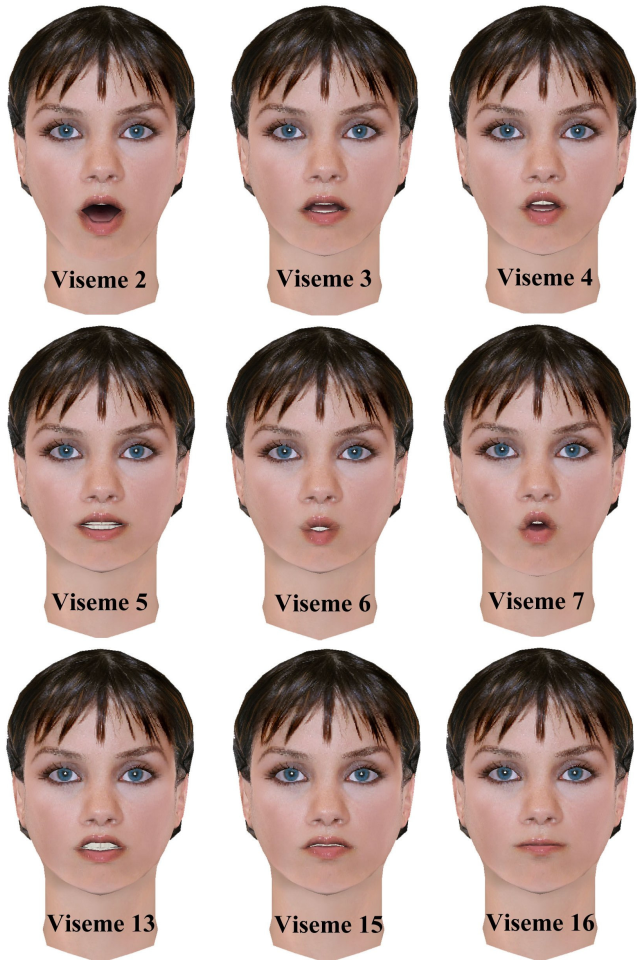


Figure 1. Example visemes for phoneme classes given in Table 1.

2.1. Phone Based Acoustic Units

A phone is the basic acoustic speech unit. We use the TIMIT speech database [12], which has 46 defined phonemes for American English. Therefore, we need to train 46 phone based HMM models. This set of HMM models and a phone-level grammar model are used to extract speech-synchronous phone sequences.

Since the output phone sequence is mapped into a viseme sequence, phone boundaries coincide with viseme boundaries. Recognizing the phonemes individually, without taking its neighboring phonemes into consideration is error prone. Therefore, this method is not satisfactory for animations where the context of a phoneme is important. To overcome this problem, we need to generate context-dependent HMM models.

TABLE 1
PHONEME TO VISEME MAPPING

Viseme Classes	Timit Phoneset	Examples
1	pau	-
2	ay, ah	bite, but
3	ey, eh, ae	bait, bet, bat
4	er	bird
5	ix, iy, ih, ax, axr,y	debit, beet, bit, about, butter, yacht
6	uw, uh, w	boot, book, way
7	ao, aa, oy, ow	bought, bott, boy, boat
8	aw	bout
9	g, hh, k, ng	gay, hay, key, sing
10	r	ray
11	l, d, n, en, el, t	lay, day, noon, button, bottle, tea
12	s, z	sea, zone
13	ch, sh, jh, zh	choke, she, joke, azure
14	th, dh	thin, then
15	f, v	fin, van
16	m, em, b, p	mom, bottom, bee, pea

Hence, we used tri-phone based HMM structures to generate context-dependent models. A tri-phone is a phoneme with left and right context (left phoneme – phoneme + right phoneme). Given a set of phone HMMs, it is possible to create context-dependent tri-phone HMMs. The total number of tri-phones in the TIMIT dictionary is 11076. Since it is very difficult to obtain sufficient data for training thousands of HMMs, we use a decision-tree based state tying strategy for sharing data and achieving robust parameter estimates.

State tying, meaning HMM states with similar phone contexts share the same set of parameters, decreases the number of parameters to be trained. We use the decision tree model that attempts to find those contexts, which make the largest difference to the acoustics for distinguishing the clusters. In the decision tree, we define 202 questions conveying left and right contexts, also phonetic classifications such as stops, fricatives, vowels etc.

2.2. Viseme Based Acoustic Units

For lip animation, it is not crucial to recognize the exact phonemes composing the speech. It is sufficient to recognize the visual components, i.e. visemes. We aim to obtain acceptable viseme sequences comparable to phone-based method by narrowing the set of acoustic units to 16 viseme classes.

However, we do not have viseme-labeled ground truth data for training the HMM models. Therefore, in order to convert phoneme-labeled ground truth data to viseme labeled ground truth data, we map the phonetic transcriptions to their corresponding viseme classes as defined in Table 1.

Viseme based HMM models have been previously studied [11], [15]. Different from the work of Dongmei et. al. [11], for training the viseme HMM models, we increase the number of Gaussian mixtures to 30 with the purpose of improving the viseme recognition rate. The number 30 is estimated by observing the plot given in Figure 2, which shows the number of Gaussian mixtures versus the viseme recognition rate.

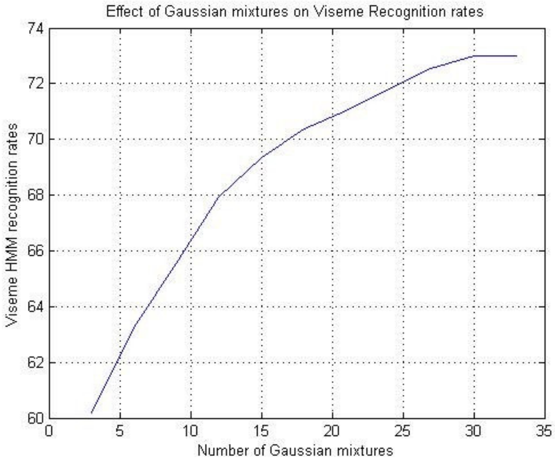


Figure 2. The plot of number of Gaussian mixtures versus viseme HMM recognition rates.

We also evaluated the tri-viseme based acoustic units. Analogous to a tri-phone, a tri-viseme is also a context-dependent structure: a viseme with left and right context makes a tri-viseme. The total number of tri-visemes for English (i. e. existing in the TIMIT database) is 1941.

As in the tri-phone case, we employ the state tying approach based on a decision-tree with 72 questions, to overcome the problem of scarce training data. Also, as in the viseme HMM case, we increase the number mixtures to 6 for the tri-viseme HMM models to achieve better recognition rates. Since, a tri-viseme is a context-dependent structure, we expect more realistic lip animation results than the viseme based method.

3. EXPERIMENTAL RESULTS

We use the TIMIT speech corpus [12] to build speaker-independent HMM models and to generate appropriate language models corresponding to each model unit.

Supervised training of the hidden Markov models are performed over the labeled TIMIT speech corpus. In the TIMIT database, there are 630 female and male speakers; each of which has 10 utterances, where two of them are the same for all speakers. We exclude these identical utterances both from training and test sets so that no utterance appears in both groups. Since we want to train speaker independent HMMs, we include 462 speakers for training, resulting in 3696 phonetically balanced utterances. TIMIT speech recordings have a sampling rate of 16 kHz and we use 100 frames per second to analyze the audio. Each utterance is analyzed over 25 ms windows with 10 ms frame shifts.

3.1 Training the Hidden Markov Models

We utilize HTK 3.1 to build and manipulate the hidden Markov models [13]. We model each phoneme or viseme using a five-state HMM with three emitting states. We use Mel Frequency Cepstral Coefficients (MFCCs) to parameterize the speech signals. Each acoustic observation consists 12 MFCCs, the energy term, and the corresponding delta and acceleration coefficients resulting a total feature length of 39. During parameterization, we also employ the cepstral mean normalization (CMN) technique that compensates for long-term spectral effects such as those caused by different microphones and audio channels.

The training of HMMs is an incremental process. We first train the 46 phone HMMs using the available phonetic transcriptions for the utterances. Then, we generate context-dependent models for tri-phones from the phone-based models.

The TIMIT phonetic transcriptions are mapped to viseme transcriptions using Table 1. Afterwards, we generate HMM models for the context-dependent tri-viseme HMM models using the viseme-based HMM models.

Our system is speaker-independent and no phonetic transcription of the input speech signal is needed for viseme generation. In order to test the proposed viseme generation approaches described in the previous section, we use the test groups of the TIMIT speech corpus consisting of 168 speakers, who are different from the training group speakers. This gives us a total of 1334 test utterances.

3.2 Objective Evaluations

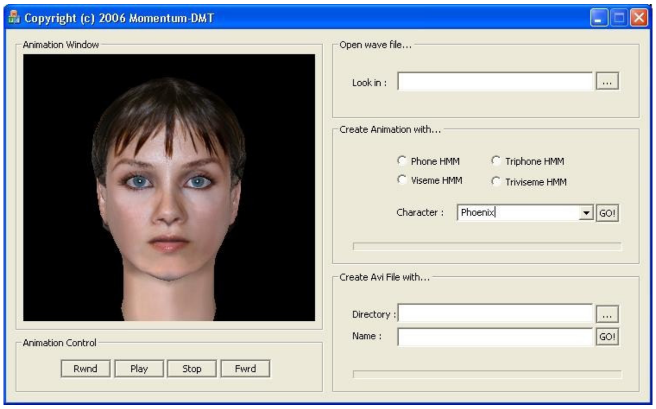
Recognition results are objectively evaluated in terms of viseme recognition rates, using HTK's performance analysis tool. Recognized phone, tri-phone, viseme and tri-viseme sequences are mapped to the corresponding viseme sequences. The recognition rate of the system is defined as the ratio of correctly recognized viseme labels (matching viseme labels in both ground truth and recognized sequences) to total number of viseme labels. The correctly recognized viseme label rates for the four strategies are summarized in Table 2 given below.

We can observe that the recognition rate of the tri-viseme based HMM method is the highest, which is followed closely by the tri-phone based HMM method.

The results of the lip synchronization can be viewed using a user interface shown in Figure 3, which is also available at our web site [16]. We are currently planning a set of subjective evaluation tests for the described HMM approaches.

TABLE 2
VISEME RECOGNITION RATES FOR PHONE, TRI-PHONE, VISEME AND TRI-VISEME HMM MODELS

Method	Recognition Rate
Phone HMM	68.36 %
Tri-phone HMM	78.75 %
Viseme HMM	73.01 %
Tri-viseme HMM	79.49 %



Character : Phoenix

Figure 3. The 3D graphical user interface used for viewing the lip synchronization results.

5. CONCLUSIONS

We compared four different HMM structures for realistic lip animation given a speech file as the only input. The performances of the phone, tri-phone, viseme and tri-viseme acoustic units are considered for HMM based viseme recognition. Based on the objective viseme recognition rates, we conclude that the tri-viseme based HMM structure outperforms the other structures.

ACKNOWLEDGMENT

This work was supported by TÜBİTAK – TEYDEB under grant 3050135 and by EC within FP6 under grant 511568 with the acronym 3DTV NoE.

REFERENCES

[1] C. C. Chibelushi, F. Deravi, and J.S.D. Mason, "A review of speech-based bimodal recognition", *IEEE Trans. on Multimedia*, vol.4, nr.1, pp. 23-37, March 2002.

[2] Magnenat-Thalmann, N., and D.Thalmann, *Synthetic Actors in Computer Generated Three-Dimensional Films*, Springer-Verlag, Tokyo, 1990.

[3] T. Ezzat, and T. Poggio, "MikeTalk: A Talking Facial Display Based on Morphing Visemes", *Proc. Computer Animation Conference*, Pennsylvania, 1998.

[4] D. Hill, A. Pearce, and B. Wyvill, "Animating speech: an automated approach using speech synthesis by rules", *The Visual Computer*, vol. 3, pp. 277-289, 1988.

[5] J. Lewis, and F. Parke, "Automated lip-synch and speech synthesis for character animation", Proc. CHI87, ACM, New York, Toronto, pp. 143-147, 1980.

[6] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on Hidden Markov models," *Speech Communication*, vol. 26, no. 1-2, pp. 105-115, 1998.

[7] L. M. Arslan, and D. Talkin, "Codebook Based Face Point Trajectory Synthesis Algorithm using Speech Input", *Speech Communication*, vol. 27, pp. 81-93, 1999.

[8] T. Ohman, "An audio-visual speech database and automatic measurements of visual speech," Quarterly Progress and Status Report, Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden, 1998.

[9] A.T. Erdem, "A New method for Generating 3D Face Models for Personalized User Interaction", *13th European Signal Processing Conference*, Antalya, September 4-8, 2005.

[10] H. McGurk, and J. MacDonald, "Hearing Lips and Seeing Voices", *Nature*, vol 264, pp. 746-748, December 1976.

[11] J. Dongmei, X.. Lei, Z. Rongchun, W. Verhelst, I. Ravyse, and H. Sahli, "Acoustic Viseme Modelling for Speech Driven Animation: A Case Study", *IEEE Benelux Workshop on MPCA*, November 2002.

[12] S. Seneff, and V. Zu. "Transcription and Alignment of the Timit Database", NIST, CD-ROM TIMIT, 1988.

[13] S. J. Young, D. Kershaw, J. Odell, and P.Woodland, "The HTK Book (for HTK Version 3.1)", http://htk.eng.cam.ac.uk/, 2001.

[14] "Methodology for the Subjective Assessment of the Quality of Television Pictures", Technical Report, Recommendation ITU-R BT.500-11, 2002.

[15] A. Verma, N. Rajput, L. V. Subramaniam, "Using Viseme Based Acoustic Models for Speech Driven Lip Synthesis", *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Proc (ICASSP)*, 2003.

[16] http://www.momentum-dmt.com/paper/tv_fdhc0.avi