International Journal of Humanoid Robotics

Biologically Valid Jaw Movements for Talking Humanoid Robots

SABRI GURBUZ
ATR Human Information Science Labs, Kyoto 619-0288, Japan
sabrig@atr.jp

KEISUKE KINOSHITA
ATR Human Information Science Labs, Kyoto 619-0288, Japan
kino@atr.jp

MARCIA RILEY
Georgia Institute of Technology, Atlanta, Georgia, USA
mriley@cc.gatech.edu

SUMIO YANO
ATR Human Information Science Labs, Kyoto 619-0288, Japan
yano@atr.jp

Here we describe our preliminary efforts toward building a phonetically and visually synchronized trainable talking robot mouth system. This work falls in the scope of learning from demonstration particularly for creating biologically valid jaw movements. Numerous studies show perception of naturalness and similarity to humans are important issues for the acceptance of social robots. For example, it is discomforting if a humanoid robot does not exhibit humanlike eye and jaw movements, or fails to pay attention to a task. Giving robots realistic mouth movements which match the auditory signal not only enhances the perception that the robot is talking, but can also increase the intelligibility of the speech uttered by the robot especially under acoustically noisy conditions. Learning jaw movement from demonstration utilizes auditory and visual sensory information to acquire perceptual motor skills from people. For each sound unit, the acquired visual motor trajectories are stored in a database called an experience library. Initial demonstration of our proposed system indicates that the correlation between the audio and the generated jaw movements is preserved for both the virtual and the hardware platforms.

Keywords: Biologically valid jaw movements, talking humanoid robots, text-to-jaw movement trajectory planning.

1. Introduction

Human-like behavior such as talking can increase a person's willingness to collaborate with a robot and helps create the social aspects of the relationship. [1,2,3,4] Our work focuses on biologically valid jaw movement synchronization during speech for a humanoid robot head. A person's jaw controls mouth opening and closing, which is an important cue in audio-visual speech communication. When implemented on a robot, the audio-synchronized jaw movements give people the compelling sense

that the robot is talking to them, even when the sound is coming from speakers rather than the robot's mouth. Potentially, lip motion could add another level of engagement between the user and the robot. Thus, we are also working on designing a robot lip structure which uses information available from our lip tracking system to create a mapping to the robot lips. For this work, however, we will focus on the naturalness of the jaw movement including its synchronization with audio.

In robotics, trajectory planning is based on building complex movements from mainly two kinds of approaches. In a *dynamical system approach* or *reflex chain hypothesis* the robot initially needs movement parameters and cost function(s). The cost functions are defined to generate the appropriate behavior. However, finding good cost functions is in many cases either difficult or complex.

A second kind of trajectory planning is based on a time-indexed movement primitives approach which is known in neuroscience as *motor tape theory.* [1] In motor tape theory, the robot stores explicit representations of all kinds of movement primitives called *motor tapes* in the experience library. When the robot needs to do a specific task indexed by time, it finds the associated motor tape and executes it. To produce a smoother trajectory for a complex task, a more desired approach would be blending and editing a set of tapes before the execution. Our research on jaw movement trajectory planning explores the motor tape theory by building jaw movements from movement primitives associated with motor tapes of sound units.

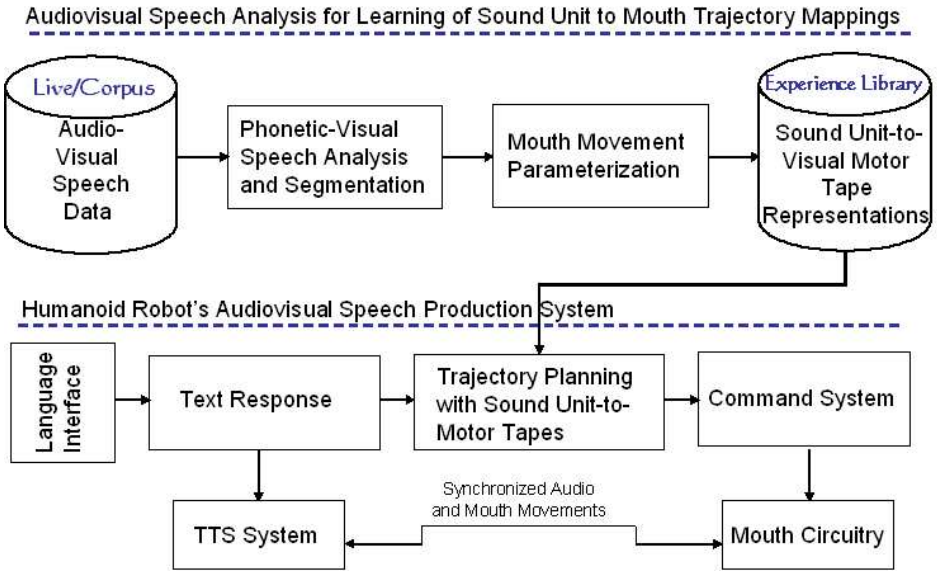

Fig. 1. Overview of the talking robot's proposed mouth system.

The human vocal track system involves movements of the vocal organs (lungs

and vocal cords) and articulators (tongue, lip, teeth and nasal cavity) to produce acoustic and visual signals that can be perceived by the auditory and visual sensory systems. Waseda University researchers developed speech production systems for talking robots (WT-1, WT-2, and WT-3) based on the human speech production system and include vocal organs and articulators. [3] This can be considered a dynamical system approach. The authors reported that they can generate Japanese vowels reasonably clearly, and can produce all Japanese consonants, stops, fricatives and nasal sounds, although not all utterances sound natural yet. [5,3] The researchers of Kismet are expanding their research efforts on naturalness and perception of humanness in robots. [6,4] Our work extends these efforts to combining a text-to-speech (TTS) system with our sound-unit-indexed jaw movement trajectory planning system.

One important goal of our system is to extract and store motor tapes by analyzing a human speaker's audio-visual speech data recorded from predetermined phonetically-balanced spoken text to create a mapping between the sound units and the time series of the mouth movement parameters representing the mouth movement trajectories. These motor tapes can then be executed with the same time index of the audio, yielding biologically valid jaw movements during audio synthesis.

We call our system the *text-to-visual speech (TTVS) synthesis* system. We plan to incorporate the Festival speech synthesis system developed by Black, Taylor and colleagues at the University of Edinburgh [7] into our work. Festival is a concatenative synthesis system: it creates indexed waveforms by concatenating parts (diphones) of natural speech recorded from humans. Using the same concatenative synthesis concept, our TTVS system concatenates corresponding mouth movement primitives. Thus, the system is capable of generating sequences of entirely novel visual speech parameters that represent the mouth movement trajectories of the spoken text. A humanoid robot equipped with TTS and TTVS systems can produce entirely novel utterances, and so is not limited to those recorded in the original audio-visual speech corpus. With these capabilities, the robot can robustly emulate a person's audiovisual speech. An overview of the system is shown in Figure 1.

2. Humanoid Head Platform

Designing and testing a physical system for research is often costly and time consuming, while developing ideas on a virtual system can be fast and thus useful for advancing the research. For these reasons we employ both a physical jaw robot system and a virtual 3D jaw robot system in our work. The virtual system allows us to test concepts without hardware limitations, and easily affords different degree of freedoms (DOFs) to help us decide what is important for improving the physical model. With both systems we also gain the ability to compare the perceptibility of their audiovisual speech articulation. Details of the systems are described in the following subsections.

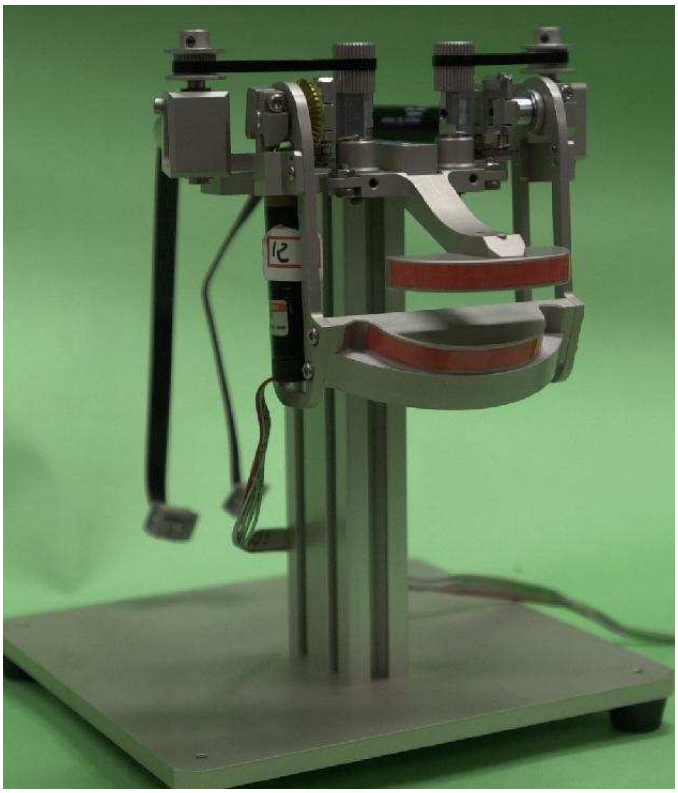

Fig. 2. The Infanoid's jaw mechanism.

2.1. *Physical System*

Our humanoid platform is a variation of *the Infanoid.* [8], and is currently limited to a head and a mechanical jaw. The mechanical jaw has three degrees of freedom: one controlling vertical motion (mouth opening), and two controlling lateral motions. Motors together with encoders are attached to each joint and are controlled from a PC using LabView software. Figure 2 shows the jaw mechanism only, and Figure 3 shows the Infanoid's head and jaw. Using this system, we aim to create biologically valid jaw movements synchronized with the TTS generated audio.

2.2. *Graphical System*

Our graphical system, shown in Figure 4, includes a 3D robot head model based on the physical robot described in Section 2.1 whose motion can be parameterized as a set of rigid body transformations, most notably rotations around joint axes for the degrees of freedom of the model. The DOFs include head nod and rotation; mouth rotation and mouth opening; eye pan and tilt; and inner and outer brow raise. The graphical model has additional DOFs not available on the physical device, such as

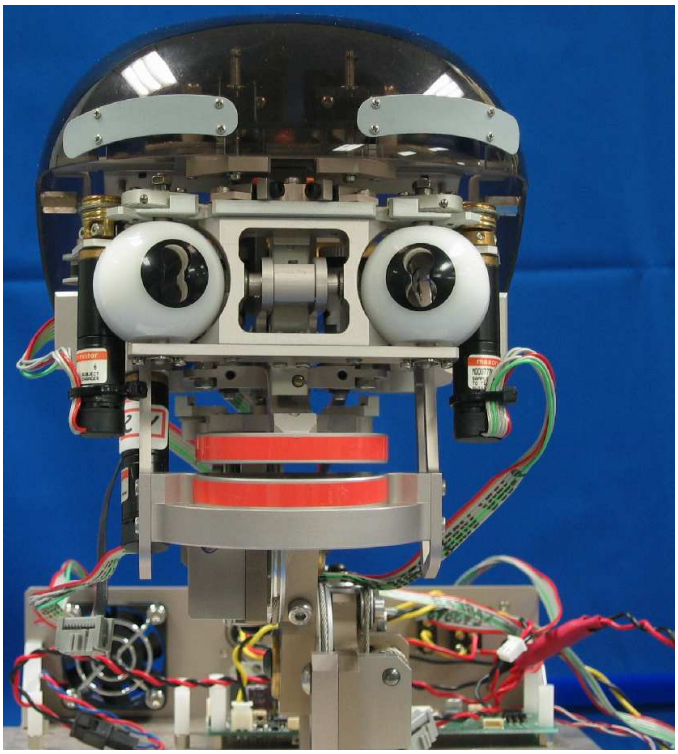

Fig. 3. Front view of the Infanoid's head.

head tilt, and allows uncoupled eyebrow motion.

We use twist coordinates to model the kinematics. [9] For a revolute joint, the twist has the form

$$\xi_{\mathbf{i}} = \begin{bmatrix} -\mathbf{n_i} \times \mathbf{q_i} \\ \mathbf{n_i} \end{bmatrix}, \tag{1}$$

where $\mathbf{n_i}$ is the unit vector in the direction of the joint axis and $\mathbf{q_i}$ is any point on the axis, both given in a global body coordinate system. See [9] for mathematical details. This representation allows us to easily model any arbitrary axis of rotation.

3. Audio-Visual Speech Analysis for Machine Learning

Audio-visual speech analysis requires an audio-visual speech corpus of a subject articulating a set of selected utterances. Each selected utterance is chosen so that it acoustically and visually instantiates one sound unit clearly. The one-to-one sound-unit-to-visual-mapping strategy is a reasonable approach for a concatanative TTS system's natural language processing (NLP) unit which produces a stream of sound units (phonemes or diphones) corresponding to the input text during the speech synthesis. Consequently, we need a mapping from sound unit to visual motor tape

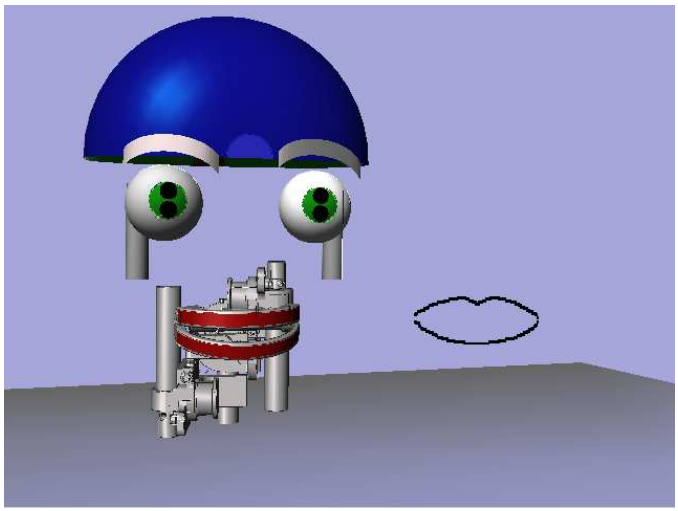

Fig. 4. The 3D graphic representation of the Infanoid: the virtual jaw is driven by mouth openning (height) information extracted from a speaker's outer lip contours (shown on the right).

representations so as to produce a stream of synchronized jaw movements. A visual motor tape representation of a sound unit is a set of parameters belonging to a set of lip shapes, while a *viseme* representation is a single static lip shape. We choose a one-to-one mapping between sound units and their visual motor tape representations to preserve smoothness and minimize the visual concatenation problem. A visual articulatory trajectory planning system takes a list of sound units as input, together with their time durations, and produces biological jaw movements for the given duration by the piecewise linear interpolation of their visual motor tape representations.

In [10] it is pointed out that in concatenative waveform synthesis, concatenating phonemes has the transition problem at the changing locations from one phone to the next because the beginning and ending of the phonemes are the least stable part of the waveform. On the other hand, the synthesizers that concatenate diphones (middle of one phone to middle of the next) join the more stable part of the signal. Of course, this increases the number of necessary sound units by the square of the number of phonemes. The number of required phones changes significantly with the language. For example the Japanese language has around 25 phones resulting in about 625 diphones, while English has around 40 phonemes resulting in about 1600 diphones. It is also reported that not all diphones are really the same, so most "diphone" synthesizers today include some commonly occurring triphone and nphone clusters. Therefore, we will create a one-to-one mapping between sound units used in the TTS system and their corresponding visual motor tape representations in order to preserve the correlation between audio and visual speech information.

3.1. *Visual Speech Analysis and Segmentation*

Facial movement analysis, especially mouth movement analysis, is an important topic in audio-visual speech perception, facial animation and talking robot research. There have been various approaches for sound-unit-to-visual-speech analysis and segmentation for facial animation applications. For the audiovisual speech synthesizer, Ezzat *et al.* [11,12] used single viseme image representations of the phonemes by manually analyzing and selecting them. They then automatically computed correspondence from every viseme to every other viseme by using optical flow methods, and constructed the visual utterance by concatenating viseme transitions. The authors reported that this approach does not handle the coarticulation effects in the visual domain and results in overly articulated lip movements. Hong *et al.* [13] used a 2D template model that covers the lower part of the face including the nostrils, lips and chin. To match the model to a natural face, they matched feature points and regions manually. Then, the feature points of a neutral face are tracked by utilizing an edge detection technique. Next they applied principal component analysis (PCA) to observed model sequences to build a mouth motion space for synthesis.

In our work, we use a contour-based parameterization method where mouth movements can be derived from a controlled video corpus data, and the mouth shape of the outer lip contour at each frame can be parametrically represented by a set of piecewise combined ellipses. The phonetic-visual speech analysis system extracts a shape-based sequence of visual representations of the sound units from the audio-visual corpus data. The characteristics of the sequence of mouth shapes are learned for each sound unit and stored in a database as a mapping from a sound unit to its visual motor tape representation. The mapping defines the physical and temporal space of the mouth movements. Varying time durations of the sound units can be accommodated by linear temporal interpolation (up sampling or down sampling) of the visual motor tape representations.

We need segmented audio in order to map the sound units to their visual representations in the corpus. Such alignment can be done either manually or by a hidden-Markov-model-based (HMM-based) machine learning system. [14] Given a text transcript and its associated audio sequence, an alignment system may use a HMM-based forced Viterbi search program where the goal is to find the optimal start and end of the phoneme boundaries for their visual motor tape representation mapping. Our visual representation mapping is based on the parametrized mouth contour which maps mouth movements to jaw movements as described in the following section.

3.2. *Mouth Contour Parameterization*

First, our mouth tracking algorithm locates the mouth region [15] and the directional outer edge of the speaker's mouth is detected. Then, the lip contour is parameterized by piecewise concatenation of the ellipses obtained from the edge data segments. A parametric contour is found that corresponds to the general quadratic equation

$ax^2 + bxy + cy^2 + dx + ey + f = 0$, where a, b, ..., f are constants, and a and c are non-zero. The upper lip is parameterized as a whole, and the lower lip contour is broken into three equal overlapping sub-contours due to lip deformation while speaking. Let us denote the 2D positions of data samples over a segment of traced lip contour as

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_N \\ y_1 & y_2 & y_3 & \dots & y_N \end{bmatrix}. \tag{2}$$

The basic form used in the elliptical parameter estimation in matrix notation is $M \times q = 0$ where $q = (a\ b\ c\ d\ e\ f)^T$. The dimensionality of M is the number of points, N, in the segment multiplied by 6 ($N \times 6$). Each row of M corresponds to one point in the segment. The parameters of each contour are then solved using the least-squares method to find a, b, c, d, e, and f.

Using the estimated parameters, parametric lip contour descriptions are generated for each segment. The generated parametric forms are combined by averaging the overlapping segments to form a final combined parametric lip contour shape. Five points are sufficient to represent each elliptical segment, leading to a significant data reduction and to the representation of lip movements in the future implementation of the lip structure.

3.3. *Current System and Visual Trajectory Mapping*

Our current physical jaw has three DOFs, one controlling mouth opening, and two controlling lateral motions. Generating the lip contour shape derived in Section 3.2 is not possible with the robot's current physical constraints, but our visual parameterization is flexible enough to implement the visual speech movements with the current design and allow implementation for a future design that includes a lip structure. So the mapping from sound unit to a visual parametric representation will have the capacity to generate articulatory speech movements of a robot for a phonetically transcribed text stream with synchrony to the TTS system.

In the current design, we need to transform the lip space, which is composed of four piecewise ellipses, to the jaw's vertical and lateral motions. The vertical motion can be modeled by the height, $h(t)$, of the mouth opening as shown in Figure 5, while the lateral motion, $\alpha(t)$, can be characterized by the skew angle of the lower lip to either side as shown in Figures 6 and 7.

4. The Robot's Audio-Visual Speech Articulation

4.1. *Language Interface and TTS System*

A humanoid robot's audio-visual speech articulation technology may work in tandem with audio only or with audio-visual speech recognition and natural language processing (NLP) systems. The NLP system may combine monolingual or multilingual speech recognizers, spoken language understanding, dialogue management,

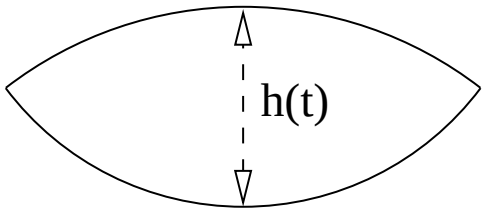

Fig. 5. Parameterization of the vertical jaw motion with $h(t)$, the height of the mouth opening between the outer lip contours.

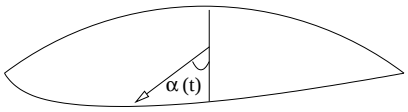

Fig. 6. Lateral jaw motion, $\alpha(t)$, with the lower lip skewed toward the left.

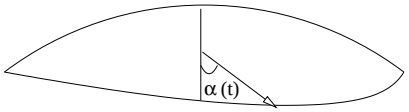

Fig. 7. Lateral jaw motion, $\alpha(t)$, with the lower lip skewed toward the right.

text-to-speech synthesis and text-to-visual speech articulation systems to create a talking humanoid robot.

Audio-based speech recognition and text-to-speech synthesis have became reasonably mature fields, and we plan to use known technology for our system. Our research efforts focus mainly on the text-to-visual speech articulation movements and their integration with the TTS system.

4.2. *Text to Visual Speech Articulation System*

In this work, we define text to visual speech articulations to be jaw movements phonetically synchronized with the synthesized speech. In this domain, we can think of audio-visual speech as follows: we see the low frequency speech information from the visual mouth movements, and we hear the high frequency speech information from the acoustic signal. Thus, biologically valid mouth movements matching the synthesized speech not only give the perception that the robot is talking, but also may increase the intelligibility of the articulated speech for poor audio signals or acoustically noisy conditions. A robot text to audiovisual speech articulation system (TTAVSAS) contains the following:

- An experience library of sound unit to visual motor tape representations
- A pipe passing the time-indexed list of sound units with their time durations to TTAVSAS
- A technique to drive articulatory movements in synchrony with the TTS system.

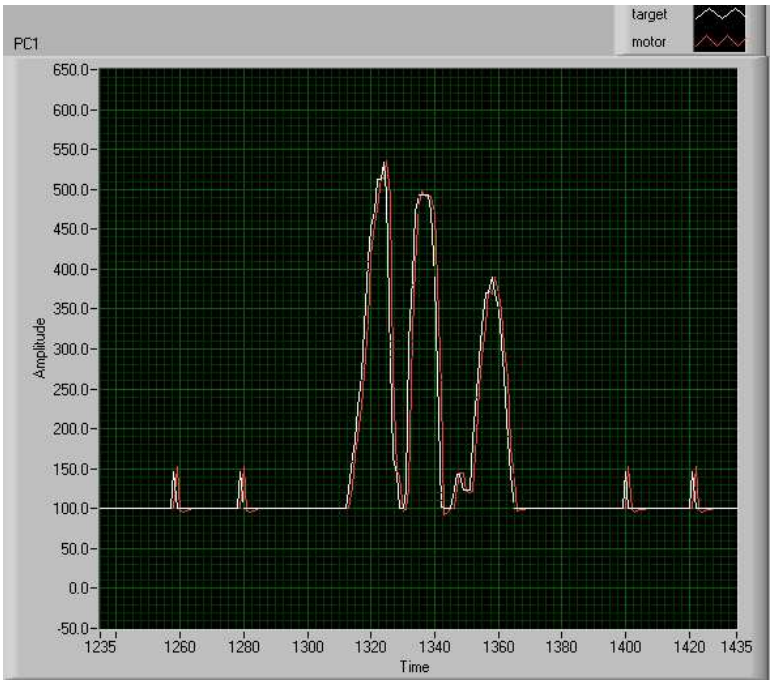

Fig. 8. Target jaw position and the motor response read from the jaw motor's encoder.

Visual articulatory parameters need to be driven from measurements of visual representations and re-aligned (fine-tuned) using apriori properties of the visible speech. Massaro *et al.* [16] reported that fine-tuning the facial movements with apriori properties of the visible speech produces more realistic and intelligible speech, approximating the visible speech produced by a natural talker. For example, during the re-alignment process, the mouth needs to be completely closed at the onset of */b/* and open at the onset of */d/*.

In our implementation, visual measurements are derived from outer lip contours using CCD camera recordings. [15] In some research, measurements are obtained by invasive techniques such as 3D motion capture or electro-magnetic articulation (EMA) systems. [17,18]

Our system will take a sound unit (phoneme or nphone) sequence and map it into the articulatory movements. Using the time-alignments of the sound units, corresponding visual representations are obtained and fine-tuned based on apriori properties of the visible speech. Then, these visual representations will be converted to visual motor tape representations of articulatory parameters (for the current jaw mechanism, height and lateral motion, $\{h(t),\ \alpha(t)\}$), that are then converted to electrical signals in the control system to drive the jaw motors.

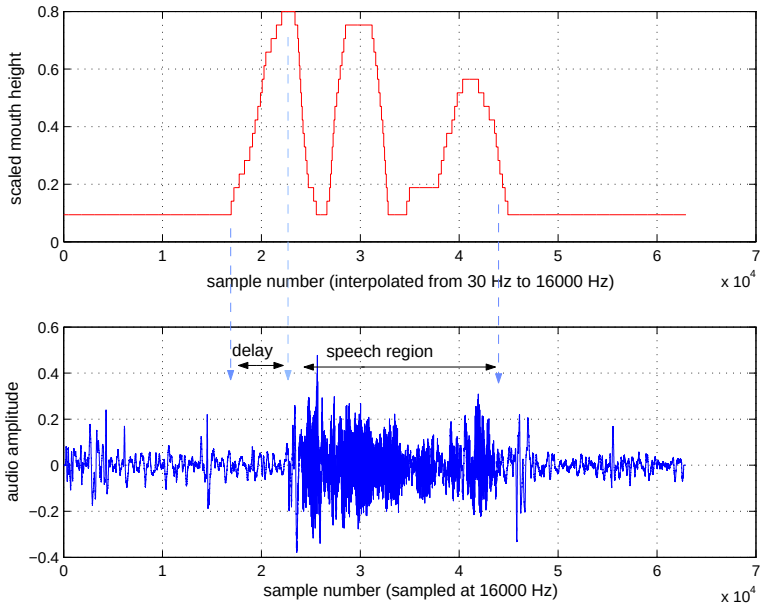

Fig. 9. Simultaneous plot of the mouth opening (top) and the audio waveform (bottom).

4.3. *Control of the Infanoid's Jaw*

The robot control system consists of a PC and a motor driver. LabView and FlexMotion (National Instruments) are used to control the jaw motor, which is equipped with a precise encoder that observes its rotation angle and uses it for PID feedback control. The PID gain coefficients are determined automatically and then manually optimized. The system can operate at up to a 120 Hz feedback cycle. The visual information was recorded at 30 frames per second (fps) using a video device, and was operated at the same speed to synchronize the jaw movements with the audio waveform. Figure 8 shows the response of the jaw motor to target jaw movements acquired from a human speaker. The delay is approximately 30 milliseconds, which is reasonable for audiovisual speech perception.

5. Experimental Task and Discussion

Asynchronous visual and audio signals can severely impair the quality of audiovisual speech perception. Figure 9 shows the simultaneous plot of mouth opening and audio waveform while a human speaker is uttering a sentence. Our mapping from sound unit to visual representation preserves the correlation between the audio waveform and visual speech (jaw) movements. The sound units can be phonemes, diphones, nphones or a combination of these, depending on the TTS system. The sound to visual motor tape mapping strategy is one to one, and the size of the visual corpus which needs to be recorded for diphone-based TTS is around 1600 for English.

The Infanoid's audiovisual talking mechanism may be viewed as a collection of two independent systems, TTS and TTVS, which are working in synchrony. During the waveform synthesis, the TTS system passes the sound unit's identification marks and their time durations to the TTVS system where jaw trajectories are planned.

To test our approach for visual speech articulation and its synchronization with the audio we have synthesized audio-visual sentences using both our virtual robot and the physical robot system. We informally observed from these experiments that the correlation between the audio and the visually articulated speech is preserved for both the virtual and the hardware platforms. Our demonstration may be viewed at *http://www.his.atr.jp/∼sabrig/Infanoid.htm.*

Our future work will address machine learning methods for smooth jaw behavior and enable the humanoid robot to learn visual articulatory motor tapes for any language with minimal human intervention.

6. Acknowledgements

This research was conducted as part of 'Research on Human Communication' with funding from the National Institute of Information and Communications Technology, and the CRL Keihanna Human Information-Communication Research Center. We also would like to thank Erhan Oztop for his suggestions on this research.

References

1. C. G. Atkeson, J. G. Hale, F. Pollick, M. Riley, S. Kotosaka, S. Schaal, T. Shibata, G. Tevatia, A. Ude, S. Vijayakumar, and M. Kawato, "Using humanoid robots to study human behavior," *IEEE Intelligent Systems*, vol. 15, no. 4, pp. 46–56, 2000.
2. F. Hara, K. Endou, and S. Shirata, "Lip-configuration control of a mouth robot for japanese vowels," in *RO-MAN '97. Proceedings., 6^{th} IEEE International Workshop on Robot and Human Communication, 29 Sept. -1 Oct.*, 1997.
3. T. Mochida, S. Hiroya, M Honda, K Nishikawa, and A. Takanishi, "Articulatory control on talking robot by mimicking formant trajectories of human speech," in *6^{th} International Seminar on Speech Production, Sydney, Australia*, 2003.
4. C. Breazel, A. Edsinger, P. Fitzpatrick, B. Scassellati, and P. Varchavskaia, "Social constraints on animate vision," *IEEE Intelligent Systems*, vol. 15, no. 4, 2000.
5. Takanishi Laboratory, "Talking robot," *Humanoid Robotics Institute, Waseda University.*
6. R. Brooks and *et al.*, "Humanoid robotics group," *http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html.*
7. A. Black and P. Taylor, *The Festival Speech Synthesis System*, University of Edinburgh, 1997.
8. CRL Japan, "Social interaction group," *http://www2.crl.go.jp/jt/a134/index.html.*
9. R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, New York, 1994.
10. A. W. Black and K. A. Lenzo, "Building synthetic voices," *http://festvox.org.*
11. T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of ACM SIGGRAPH 2002, San Antonio, Texas*, 2002.
12. T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," *International Journal of Computer Vision*, vol. 38, 2000.

13. P. Hong, T. S. Huang, and X. Lin, "Mouth motion learning and generating from observation," in *IEEE Workshop on Multimedia Signal Processing*, 1998.
14. S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *HTK Book*, Cambridge University Press, 1997.
15. S. Gurbuz, K Kinoshita, and S. Yano, "Mouth tracking from video sequences using trainable multivariate gaussian classifiers," in *PRMU 2003, Sendai, Japan*, 2003.
16. D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'99), pp. 133-138*, 1999.
17. E. Vatikiotis-Bateson and H. Yehia, "Physiological modeling of facial motion during speech," *Trans. Tech. Com. Physcho. Physio. Acoust.*, 1996.
18. M. M. Cohen, D. W. Massaro, and R. Clark, "Training a talking head," in *IEEE Fourth International Conference on Multimodal Interfaces (ICMI'02), Pittsburgh, Pennsylvannia*, 2002.