ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Evaluation of a formant-based speech-driven lip motion generation

Carlos T. Ishi [1], Chaoran Liu [1], Hiroshi Ishiguro [2], Norihiro Hagita [1]

[1] ATR Intelligent Robotics and Communication Labs.
[2] ATR Hiroshi Ishiguro Labs.

carlos@atr.jp, chaoran.liu@is.sys.es.osaka-u.ac.jp, ishiguro@sys.es.osaka-u.ac.jp, hagita@atr.jp

Abstract

The background of the present work is the development of a tele-presence robot system where the lip motion of a remote humanoid robot is automatically controlled from the operator's voice. In the present paper, we introduce an improved version of our proposed speech-driven lip motion generation method, where lip height and width degrees are estimated based on vowel formant information. The method requires the calibration of only one parameter for speaker normalization, so that no training of dedicated models is necessary. Evaluation was conducted in a female android robot and in animated lips. Subjective evaluation indicated that naturalness of lip motion generated in the robot is improved by the inclusion of a partial lip width control (with stretching of the lip corners). Highest naturalness scores were achieved for the animated lips, showing the effectiveness of the proposed method.

Index Terms: lip motion, formant, tele-operation, humanoid robot.

10.21437/Interspeech.2012-39

1. Introduction

We have been developing tele-operation systems for transmitting human tele-presence through humanoid robots, such as androids. Previous tele-operation systems have used motion capture or vision-based lip tracking techniques. However, it has been experienced that the performance of vision-based approaches is dependent on the speaker, besides other factors such as good lighting conditions and image resolution. On the other hand, a motion capture system is more robust to these factors, but it is too expensive. Thus, we decided to develop a tele-operation system where the lip motion of a remote humanoid robot is automatically controlled from the operator's voice.

Several approaches have been proposed for converting speech or text to lip motion. When text or phonetic transcriptions are available, methods like concatenation, trajectory generation, or dominance functions (which are linear combinations of trajectories selected according to a phonetic transcription) can be applied [1,2]. Lip motion generation methods based on audio-only can roughly be categorized in phone-based methods or direct audio-visual conversion. Phone-based methods model the audio-visual data using different phone models, mostly artificial neural networks (ANN) and hidden Markov models (HMM) [3-5]. However, in a tele-operated system, an online real-time conversion of audio to lip motion is required, so that text-based methods or phone-based models are not appropriate, since the input is speech and different phones have different durations, leading to variable and large latencies.

Direct audio-visual conversion, without using phone models, have been shown to be more effective than HMM-based phone models [6,7]. For example, in [7], these two approaches are directly compared, and it was shown that ANN-based method was judged significantly better than the HMM method. An explanation for that is because ANN-based approaches typically work on a frame-wise basis, and can offer closer, more direct synchronization with the acoustic signal than HMM, in which the mapping is mediated through longer phone-size units.

Another class of direct audio-visual conversion methods uses Gaussian Mixture Models (GMM) [8]. In [8], a maximum likelihood estimation of the visual parameter trajectories are adopted using an audio-visual joint GMM, and a minimum converted trajectory error approach is proposed for refining the converted visual parameters.

Another approach of direct audio-visual conversion is the use of information related to formants (acoustic resonances of the vocal tract) [9,10]. Although most approaches use Mel-frequency cepstral coefficients (MFCC) as acoustic parameters, the interpretation of their values with regard to phonetic contents are not straightforward as the formant frequencies. Further, all MFCC-based methods require construction of dedicated models prior to their use. Thus, we decided to adopt a formant-based lip motion generation approach.

Vowels can be represented in a two-dimensional space formed by the first and second formants (F1, F2). It is well-known that there is a relationship between formant space and vocal tract area functions (including lips) [11]. For example, F1 is related to the jaw lowering, while F2 is related to front-back position of the tongue. However, lip opening and closing may occur without jaw lowering, so that the relationship between F1 and lip opening is not straightforward. In our previous work [12], we proposed a method for lip opening degree estimation from the first and second formants. It was shown that the lip height control by the audio-based approach performed better than vision-based and motion capture-based approaches.

In the present work, we improved the parameter extraction and evaluated the performance of a partial lip width control in a female android robot. The term "partial" is because the android cannot round the lips but has an actuator for stretching the lip corners. A comparative evaluation was also conducted on animated lips generated by the formant-based method.

2. The proposed method

Fig. 1 shows a block diagram for the proposed lip motion generation method.

Firstly, formant extraction is conducted on the input speech signal. Then, a transformation of the formant space given by the first and second formants is realized according to a speaker-dependent parameter to obtain a normalized vowel space. This normalization is necessary since the vowel space differs depending on the speaker's feature, such as gender, age and height. The lip shape is then estimated from the normalized vowel space

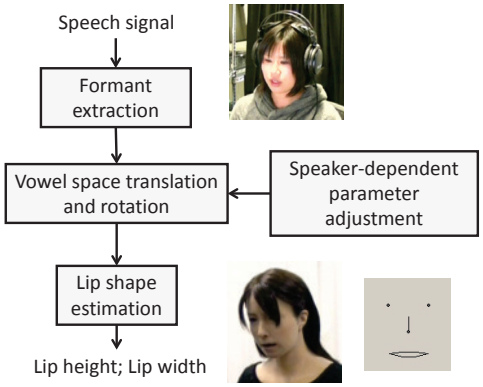

Figure 1. *Block diagram of the proposed lip motion generation.*

The following sections describe each block in detail.

2.1. Formant extraction

The formant extraction implemented in the present work is a conventional method of picking the peaks of the LPC (Linear Predictive Coding) smoothed spectrum [13]. Nonetheless, any other formant extraction method with better performance could be used instead.

The input signal is captured at 16kHz/16bit resolution, pre-emphasized by $1\text{-}0.97z^{-1}$ for reducing the effects of the glottal waveform and the lip radiation, and framed by a 32ms Hamming window, in 10ms intervals. Then, 19-th order LPC coefficients a_k (k = 0~18, with a0 = 1) are extracted for each frame. The LPC smoothed spectrum is then obtained by taking a 512-point FFT of the LPC coefficients a_k with zero-padding.

Finally, the first and second peaks are picked from the LPC smoothed spectrum, searching from low to high frequencies, for the first and second formants (F1 and F2).

In our previous work, the LPC analysis order was fixed to 19. However, it is known that the order should be selected according to the actual number of formants in the analyzed bandwidth. So, the LPC order should be lower for female voices and higher for male voices. Thus, in the present work, the LPC order was made adjustable according to the speaker normalization factor, as explained in the following sub-section.

2.2. Calibration and speaker normalization

As a first step for speaker normalization, the origin of the coordinates in the formant space (centerF1; centerF2) are adjusted according to the speaker, since the vowel space changes for different speakers (e.g. due to differences in gender, age and height). Specifically, the new origin is moved to the center of the speaker's vowel space (corresponding to the schwa vowel in English) in the logF1 vs. logF2 space (i.e., the space given by the logarithms of the first and second formants). (Strictly, the logarithm of physical quantities does not exist, so that the values used in the present work correspond to using a reference frequency of 1 Hz.)

The center point of the speaker's vowel space is made adjustable by a GUI (graphic user interface), so that after uttering isolated vowels, the user (operator) is able to identify visually an approximate position of the center of his/her vowel system. In practice, only centerF1 is made adjustable, in the range of 400 to 800 Hz, in steps of 10 Hz. The centerF2 can be automatically estimated from centerF1, according to the following expression:

$$\text{centerF2} = 2.9 * \text{centerF1} \quad (1)$$

Theoretically, an open-closed straight tube (which is a quarter-wave resonator) would have F2 equals to three times F1 [14]. However, the neutral vocal tract configuration is not like a straight tube, so that expression (1) was found to fit better in our preliminary analysis.

Regarding the GUI-based manual adjustment, a reasonable criterion is to adjust the vowel center to lie between the plots of /e/ and /o/ utterances in the formant space.

The next step of speaker normalization is the scaling of the coordinates. This step would be equivalent to a vocal tract length normalization, where the logF1 and logF2 coordinates are stretched or enlarged. Preliminary analysis indicated that scaling factors around 2 for male and 1.8 for female speakers are good approximations. In the present work, the scaling factor is automatically estimated from centerF1, so that values around 450 ~ 500 Hz (average for male speakers) produce scaling factors around 2, and values around 540 ~ 600 produce scaling factors around 1.8.

The LPC order described in the previous sub-section was also adjusted according to the centerF1 value, so that values around 15 ~ 19 are obtained for female voices, and values around 19 ~ 23 are obtained for male voices. This improves the correspondence between the formants and the peaks in the LPC spectrum.

2.3. Vowel space rotation and lip shape estimation

After moving the origin of the formant space to the center vowel position, the axes are rotated counter-clockwise by about 25 degrees. The new coordinate axes after rotation will be represented by logF1' and logF2'. This rotation process was motivated by the observations that the logF1' axis (after rotation) has good correspondence with the (vertical) aperture of the lips. Fig. 2 shows examples of distributions of the formant maps for isolated vowels uttered by two speakers (one male and one female), superimposed by the average vowel spaces for Japanese male and female speakers. The new coordinates after translation to the center of the vowel space and rotation are also shown. Note that logF1' values are ordered as /a/ > /e/ > 0 > /i/ ≅ /o/ > /u/, which correspond to the relative lip height variations between the different vowels. The center (schwa) vowel has logF1' = 0. Note also that /i/ and /o/ have different lip width (lip spreading in /i/ and lip rounding in /o/) but have approximately the same lip height.

Normalized lip height values are estimated from the formants according to the following expression:

$$\text{lip\_height} = 0.5 + \text{height\_scale} * \text{logF1'}, \quad (2)$$

where lip_height = 0 corresponds to closed mouth, lip_height = 1 corresponds to a maximally opened mouth, the factor 0.5 corresponds to the aperture for the center (schwa) vowel, and height_scale is the scaling factor described in Section 2.2.

For the lip width, we use the F2 value before rotation. Although F2 is known to be more related to the front/back tongue position, there is also relationship between lip spreading/rounding in most languages. F2 (or logF2) values are ordered like /i/ > /e/ > /a/ > /u/ > /o/, which correspond to the degree of lip spreading in /i/ to lip rounding in /o/. Lip width values can be estimated from the formants according to the following expression.

$$\text{delta\_lip\_width} = \text{width\_scale} * (\text{logF2} - \log \text{centerF2}) \quad (3)$$

For delta_lip_width, positive values are obtained when F2 is higher than centerF2, like in /i/ and /e/ where the lips are spread, while negative values are obtained when F2 is lower than centerF2, like in /o/ and /u/ where the lips are rounded. The scaling factor width_scale determines the degree of lip spreading/rounding relative to lip height. Values around 0.5 were found to produce human-like lip shapes. Nonetheless, it is important to mention that the relationship between F2 and lip width will become ambiguous in languages that distinguish

rounded front vowels from rounded back vowels (as in French).

Representative lip shapes generated for each vowel are also shown in Fig. 2. In the figure, we can observe that similar lip shapes are generated for the vowels of different speakers, thanks to the normalization processing.

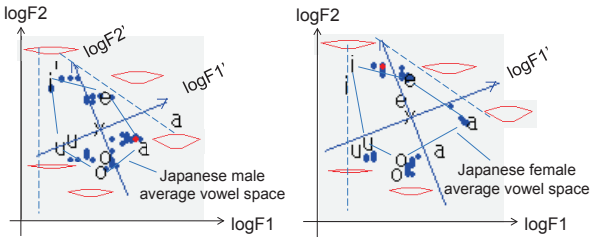

Figure 2. *Examples of the distributions of single vowels uttered by a male and a female speaker.*

The mapping between formants and lip shapes can be constructed in vowel or semivowel intervals. In consonants, where there is a constriction in the vocal tract, the formants are more difficult to be estimated, and its relationship with lip shape is less straightforward.

Then, constraints on formant range and power values are imposed for discriminating consonants, and a fix lip height of 0.35, corresponding to an average aperture in consonants, is forced. In the present work, the constraints for accepting the detected formants for lip motion generation were improved by establishing an upper limit for the vowel formant space and detection of fricative and affricative consonants like /s/, /sh/, /ts/, /ch/. The constraints are as follows.

$$F1 > centerF1*0.5, \quad (4)$$

$$logF2 < f(logF1), \quad (5)$$

$$sonorant\_power > power\_threshold, \quad (6)$$

$$fricative\_power < sonorant\_power, \quad (7)$$

where f (.) is a straight line function defining the upper limit of vowel formants in the logF1 vs logF2 space, sonorant_power is the power value computed in the frequency band of 100 ~ 3000 Hz, where the power of vowels are concentrated, while fricative_power is the power in the frequency band of 3000 to 8000 Hz, where the power of fricative and affricative consonants are concentrated. The coefficients of the straight line f (.) are obtained from two points: (log centerF1; log centerF2 + 0.55/height_scale) and (log centerF1 + 0.7/height_scale; log centerF2). These points were defined from observations of the distributions of the vowel formants in the logF1 vs. logF2 space. These lines are shown by dashed lines in Fig. 2.

If the low-power interval time exceeds a threshold (200 ms), it is judged to be a non-speech interval, and the mouth is gradually closed by a multiplying factor of 0.95, so that the mouth is totally closed after 200 ~ 400 ms.

Finally, a moving average smoothing filter with 9 taps (4 past and 4 future points, with intervals of 10 ms between two points) is passed through the generated lip height and width sequences.

3. Evaluation of the proposed method

3.1. Experimental setup for evaluation of lip height control

Simultaneous recordings of audio, vision-based face parameter data and motion captured data were conducted for seven speakers (four male and three female) speaking in several languages. The speakers are researchers, intern students or research staffs working in our laboratory, which could speak more than one language. Table 1 shows the languages spoken by the female speakers (F1 ~ F3). The male speakers' data were excluded from the evaluation of lip width control, since a female android robot was used in the evaluation.

Table 1. List of the speakers' origin and the spoken languages.

Speaker ID	Origin	Spoken languages
F1	Japanese	Japanese, English
F2	Iranian	Persian, Japanese, English
F3	Turkish	Turkish, English

The motions were generated in a female android robot, Geminoid-F, and in animated lips, as shown in the bottom pictures of Fig. 1. For the animated lips, the lip corners were drawn with a height of 1/3 of the estimated lip height, and with a width changing according to the estimated delta_lip_width value.

For Geminoid-F, the normalized lip motion parameters generated by the method described in Section 2.3 is linearly mapped to the actuator commands of the robot.

The lip motion is basically controlled by the jaw actuator, which linearly controls the degree of mouth opening (or equivalently the lip height). The allowed actuator command values are from 0 to 255, where 0 corresponds to closed mouth, while 255 corresponds to the maximum opening of the mouth. The maximum actuator command value was limited to 200, for avoiding an excessive opening of the mouth during speech.

Lip width control cannot be thoroughly evaluated in Geminoid-F, since there are no dedicated actuators for rounding the lips. However, Geminoid-F has one actuator for stretching the lip corners to outside and upward direction, for realizing a smiling behavior. Although this lip corner stretching is not strictly in the same direction for the lip corner spreading in /i/ and /e/ (due to the presence of a simultaneous upward direction), we considered that lip width control can be partially evaluated in Geminoid-F.

We then mapped the lip width control parameters to this lip corner actuator, and conducted an experiment to verify whether or not the use of this actuator improves the perceived naturalness.

The lip corner actuator commands were generated by a linear mapping between the lip width parameters to the lip corner actuator according to the following expression:

$$LipCornerActuator = delta\_lip\_width*64 + 64 \quad (8)$$

For this actuator, the maximum value was limited to 127, to avoid the appearance of a strong smiling expression.

Two motion types were generated in Geminoid-F, for the female voices (in Table 1) by using only the jaw actuator ("jaw only") or using both jaw and lip corner actuators ("jaw+lip corner"). Video clips were recorded for the motions generated in Geminoid-F for each speaker. Segments of 10 ~ 20 seconds were selected from the utterances spoken by each speaker and each language (as shown in Table 1), resulting in seven stimuli for each motion type.

In the evaluation experiment, the stimuli (video clips) were played in sequence, in random order for the motion types. Subjects were asked to grade the naturalness scores in a 7-point scale, where 1 is the most unnatural, 7 is the most natural, and 4 is "difficult to decide". A "natural" motion means a "human-like" motion. Ten subjects (native speakers of Japanese) evaluated the naturalness of the generated motions.

3.2. Evaluation results

Fig. 3 shows the subjective naturalness scores for lip motion generated by jaw actuator only ("jaw only") and by both jaw and lip corner actuators ("jaw+lip corner") in Geminoid-F, and

in the animated lips ("animation"). The naturalness of individual motion types were obtained by using the scores and normalizing to the scale from 0 (most unnatural) to 100 (most natural). Repeated measure ANOVA was conducted for the significance tests of the scores.

The subjective scores in Fig. 3 indicate that the use of the lip corner actuator increases overall naturalness of the lip motion. However, some of the subjects preferred "jaw only" than "jaw+lip corner". The main reason of unnaturalness was that the lip corner motion was a little bit "jerky" in some of the stimuli. The explanation for this jerky motion was in formant extraction errors mainly in /o/ and /u/ where F2 was misdetected as the true F3, and in consonant portions where high F2 values are detected. These eventual misdetections in F2 were causing a sudden change in lip corner actuation, resulting in an unnatural motion. However, the results show that even though there are still problems in formant extraction, the use of lip corner actuator can lead to a more natural lip motion. Finally, the subjective naturalness in the animated lips was higher than in the robot, suggesting that if lip rounding can be controlled in the robot, higher naturalness can still be achieved.

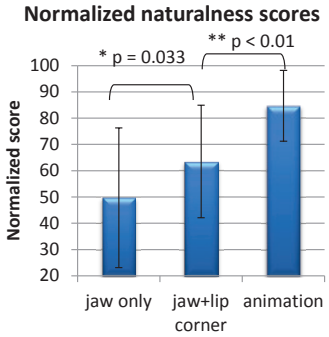

Figure 3. *Subjective naturalness scores for lip motion generated by jaw actuator only ("jaw only") and by both jaw and lip corner actuators ("jaw+lip corner"), for the audio-based method in GeminoidF(F), and in the animated lips ("animation").*

4. Discussions

Regarding the constraints of the present method, a reasonably good formant extraction is desirable. We have observed that a simple method of peak picking in the LPC smoothed spectrum often fails when the first two formants are close (such as in /o/ and /u/). A relatively natural motion could be generated by the lip height control of the present method, even under the current limitations in the formant extraction. However, we also observed that errors in formant extraction cause more severe problems in lip width control. Further improvement on formant extraction is the next step of our work.

Another remaining issue is lip shape estimation in the consonants. With the current approach, the lip closure can be correctly detected in the bilabials /m/ and /b/ when the transitional parts to/from the neighbor vowels shows F1 and F2 lowering curves, but these curves are seldom observed in /p/ (where the lips should close). There is also a tradeoff of doing or not doing smoothing of the generated actuator commands, since it avoids the occurrence of jerky motion, but also avoids the lips to completely close in bilabials.

However, even though the proposed method cannot generate perfect lip motion, it is more effective than vision-based approaches for purposes of generating natural lip motion. It is worth to mention that additional evaluation confirmed the same trends of our previous work, where higher subjective scores were achieved by the proposed formant-based approach compared with vision- and motion capture-based approaches.

Finally, no clear differences were found among different languages, so that the audio-based lip motion was consistently judged as more natural than other motion types, regardless of the language. However, the absolute naturalness judgment could be different for subjects with different origins. This would be a topic for future investigation.

5. Conclusions

With the aim of tele-operating the lip motion of a remote humanoid robot in synchrony with the operator's voice, we developed and evaluated a formant-based lip motion (height and width) generation method.

Evaluation of the lip motion generated by the proposed formant-based method indicated that even a partial lip width control through the lip corner stretching actuators improve the perceived naturalness in the android robot. Highest subjective naturalness above 80 (for a 0 to 100 scale) was obtained for the animated lips, where lip rounding can also be generated, indicating that natural lip motion can be generated by the proposed method.

Analysis of the reasons of low naturalness scores revealed remaining issues on formant extraction and detection of bilabial consonants. These are targets for our future work.

6. Acknowledgements

This work was supported by JST/CREST.

7. References

[1] Cohen, M., and Massaro, D. 1993. Modeling coarticulation in synthetic visual speech. In Models and techniques in computer animation, Jan. 1993.

[2] Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T. 1998. Text-to-visual speech synthesis based on parameter generation from HMM. In Proceedings of ICASSP98, 3745-3748.

[3] Hong, P., Wen, Z., and Huang, T. 2002. Real-time speech-driven face animation with expressions using neural networks. IEEE Trans. on Neural Networks 13, 4, (Jul. 2002), 916-927.

[4] Beskow, J., and Nordenberg, M. 2005. Data-driven synthesis of expressive visual speech using an MPEG-4 talking head In Proceedings of Interspeech2005, 793-796.

[5] Hofer, G., Yamagishi, J., and Shimodaira, H. 2008. Speech-driven lip motion generation with a trajectory HMM. Proc. Interspeech 2008, 2314-2317.

[6] Takacs, G., 2009. Direct, modular and hybrid audio to visual speech conversion methods – a comparative study. Proc. Interspeech09, 2267-2270.

[7] Hofer, G., and Richmond, K. 2010. Comparison of HMM and TMDN methods for lip synchronization. Proc of Interspeech2010, 454-457.

[8] Zhuang, X., et al. 2010. A minimum converted trajectory error (MCTE) approach to high quality speech-to-lips conversion. In Proceedings of Interspeech 2010, 1726-1739.

[9] Wu, J., et al. 2008. Statistical correlation analysis between lip contour parameters and formant parameters for Mandarin monophthongs. In Proc. AVSP2008, 121-126.

[10] Erber, N. 1979. Real-time synthesis of optical lip shapes from vowel sounds. J. Acoust. Soc. Am. 66(5), 1542-1546.

[11] Ladefoged, P., Harshman, R., Goldstein, L., Rice, L. 1978. Generating vocal tract shapes from formant frequencies. J. Acoust. Soc. Am. 64(4), 1027-1035.

[12] Ishi, C., et al, 2011. Speech-driven lip motion generation for tele-operated humanoid robots. In Proc. AVSP2011, 131-135

[13] Markel, J.D., and Gray, A.H. 1976. Linear prediction of speech. Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[14] Titze, I.R. 1994. Principles of voice production, Prentice Hall, New Jersey, 136-168.