Development of a lip-sync algorithm based on an audio-visual corpus

Jinyoung Kim, Seungho Choi** and Joohun Lee****

*Department of Electronics and computer Engineering, Chonnam National University, Korea
**Department of IT Engineering, Dongshin University, Korea
*** School of Science and Engineering, Waseda University, Japan
kimjin@dsp.chonnam.ac.kr

Abstract

In this paper, we propose a corpus-based lip-sync algorithm for natural face animation. Audio-visual (AV) corpus was constructed from the video-recorded announcer's facial shot, speaking the given texts selected from newspapers. To obtain lip parameters, we attached 19 markers on the speaker's face, and we extracted the marker positions by the color filtering followed by the center-of-gravity methods. Also, the spoken utterances were labeled with HTK and such prosodic information as duration, pitch and intensity was extracted as parameters. By combining the audio information with the lip parameters, we constructed audio-visual corpus.

Based on this AV corpus, we propose a concatenating method of AV units, which is similar to corpus-based Text-to-speech. For an AV unit search, we used a CVC-syllable unit as a basic synthetic unit. There are two procedures to get lip parameters for given texts and speech. First, top-N candidates for necessary CVC units are selected by two proposed distance measures. The one measure is a phonetic environment distance and the other is a prosodic distance. Second, the best path is estimated from the top-N AV unit sequence and Viterbi search algorithm is used for it.

From the computer simulation results, we found that the information not only about duration but also about pitch and intensity is useful to enhance the lip-sync performance. And the reconstructed lip parameters are almost equal to the original parameters.

1. Introduction

The ability of speech perception tends to decrease under noisy environments. Visible speech is particularly effective when auditory speech is degraded [1,2]. It is known that adding visual information to speech can increase human perception ability. Also, with the growth of Internet and animation techniques, it is possible to implement human-like avatar. Thus lip-sync technique becomes one of the promising techniques.

When developing lip-sync algorithm, there're some requirements as follows: 1) synchronization of audio and video [3], 2) modeling of natural coarticulation [4,5,6,7], and 3) parameter minimization for a real-time implement. Several approaches have been proposed to solve these problems, however, those methods did not consider coarticulation effect perfectly. In this paper, we propose a corpus-based lip-sync method similar to corpus-based TTS approach. As like the case of corpus-based TTS, we constructed AV-corpus and devised a method to concatenate some AV-unit. More details are explained in the following chapters.

2. AV corpus

In this paper, we constructed an AV corpus. AV recording was performed by SONY camcorder. The audio and video signals were digitized simultaneously by using a sound card and a video card, while the sampled AV signals were synchronized. In this chapter, we explain how the AV corpus was constructed from the digitized AV signals.

2.1 Markers

We extracted the lip-related parameters from the face image. For robust tracking facial movements, some markers were attached to the informative points of the announcer's face. The figure 1 shows the face with the markers attached. There are used 19 markers. 4 markers are placed around the lip contour and 7 markers are placed along the jaw's contour (see Figure 1). Auto-reflective round markers were used for distinguishing them from the face's skin. We used brightness and chromaticity information to track the marker's position. The positions were extracted through the well-known image processing techniques.

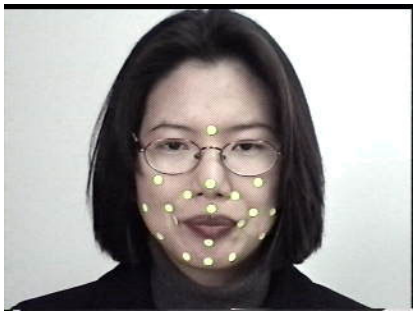


Figure 1. An announcer with some markers.

2.2 Data Processing

Audio signal was digitized with sampling rates of 8KHz. We captured 30 images per second. The sampled AV signals were processed with the procedure shown in figure 2 and as the following steps.

Step 1. Tracking each marker's position: Markers on the speaker's face had different intensities and colors enough to extract each position from original images. After converting color images to binary with some thresholds, the x- and y-axis profiles of these binary images were calculated. Based on this information, the markers' positions were estimated approximately. And we could get marker's positions with the center-of-gravity method.

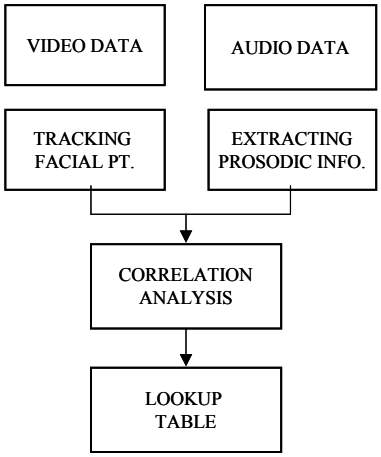


Figure 2. Data processing procedure.

Step 2. PCA (Principal component analysis) for data reduction: 19 markers have each *x*, *y* position and, so there are 38 visual parameters. These parameters are too many to handle. So PCA was applied to reduce 38 parameters into 8 ones which were containing 96% of the total information (see Figure 3).

Step 3. Extraction of prosodic information: From the speech files, we extracted pitch, duration and intensity information for each viseme. We used the AMDF method to get pitch information and modify the phoneme's length information from the labeling files. The labeling was performed by HTK (HMM Toolkit) with the predefined viseme units.

Step 4. Data analysis: Correlation analysis between audio and video information was performed. Analysis for visual parameters and prosodic ones for each viseme were performed, since there are some correlations between audio and video information. For example, when one speaks loudly, facial movement is getting bigger, and when speaks longer, the coarticulation has less effect than as usual.

Step 5. Establishment of a lookup table: Look-up table was constructed. This contains prosody, phoneme, lip and timing information for each instance of basic units. This table is explained later.

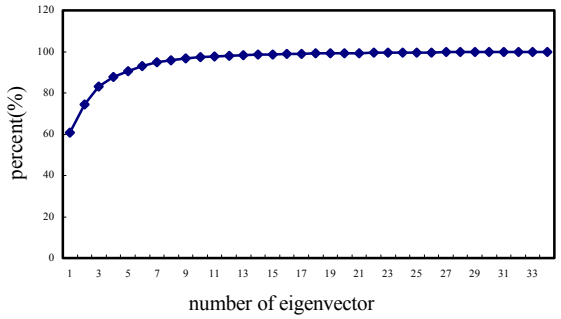


Figure 3. PCA result

2.3 Corpus construction

2.3.1 Concatenation unit

As we mentioned above, visual speech synthesis is performed with the concatenation method. Viseme is a simple unit, but poor to reflect a coarticulation unit. We need to select the concatenation unit that considers coarticulation effect. On the other hand, a CVC-syllable is an elementary structure of Korean pronunciation. And a CVC-syllable reflects the coarticulation effect as a tri-phone unit. A CVC-syllable unit is determined as a basic unit. Table 1 shows an example of the grouping and concatenation of a CVC-unit.

Table 1. Example of CVC-unit grouping and concatenation

Recording [(VC)+CVC+(CV)]
C V C V C V C C V C C V
sil 1: (g_eu) d e r eu n d e n g a
sil 2: g eu (d_e) r eu n d e n g a
sil 3: g eu d e (r_eu_n) d e n g a
sil 4: g eu d e r eu n (d_e_n) g a
Synthesis
g_eu + d_e + r_eu_n + d_e_n + g_a

2.3.2 Structure of Look-up table

Table 2 shows the structure of look-up table. The look-up tables were established for each CVC-syllable unit.

Table 2: Structure of look-up table

```
// Identity of CVC-syllable and prosody information
  Viseme1[C] pitch, intensity, duration
  Viseme2[V] pitch, intensity, duration
  Viseme3[C] pitch, intensity, duration

// Number of instance
   NofIns

// Instance 1
   // Preceding visemes
     Viseme1[V] pitch, intensity, duration
     Viseme2[C] pitch, intensity, duration
   //Following visemes
     Viseme1[C] pitch, intensity, duration
     Viseme2[V] pitch, intensity, duration
   // Number of frames
     Nof
   // Lip parameters(LP)
     Frame 1 : LP1
     Frame 2 : LP2
        ……..
      Frame Nof : LP_Nof

// Instance 2
   ……..
```

Some information of all instances for the given CVC unit is written in the each look. And prosody and information of CVC-unit and its phonetic environments (preceding and following visemes) are written. Also, lip parameters of frames from the instance of some viseme are written.

3. Concatenation Method

For a given-labeled speech, lip parameters are determined by the concatenation method based on the predefined AV-corpus. In this paper, the proposed method is very similar to that of the corpus-based TTS. In other words, the concatenation method is composed of two steps. The one is to select top-N candidates for the given CVC unit. The other is to determine the best path having the smallest concatenation cost from the pre-selected top-N candidate sequences. Figure 4 shows the proposed concatenation method and those steps in our method are explained as follows.

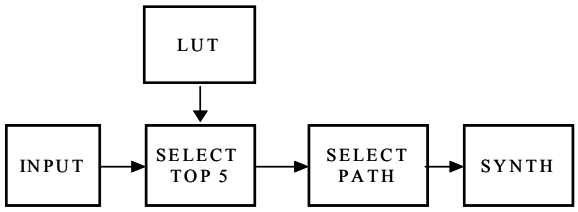


***Figure* 4**. Selecting corpus process

3.1 Candidate selection

Assume that the given speech utterance is "g-a-n-d-a"(in English, it means 'to go'). Then, the synthetic units are "g-a-n" and "d-a-x", where "x" means that there is no following consonant at that position). Now, the first work to do is to select the candidates for each CVC syllable unit. Searching the look-up table is performed for this. Assume that we labeled the utterance of "g-a-n-d-a" and analyzed the prosody parameters already. Then, we have the target values of prosody. And the phonetic environment is well defined. That is, for the target unit of "g-a-n", the preceding phonetic environment is "x-x" and the following phonetic environment is "d-a". We defined phonetic and prosodic distance measures.

Thus the dissimilarity of the source (from the look-up table) and the target patterns is defined as the weighted sum of the phonemic and the prosodic distances.

$$D = wD_{phon} + (1-w)D_{pros} \tag{1}$$

where D is a total distance summed D_{phon} and D_{pros} with the predefined weight parameter, and D_{phon} , D_{pros} are the phonemic and prosodic distance. In our approach, D_{pros} is defined based on Euclidean distance. Because prosody information includes duration, pitch and intensity information, D_{pros} is defined as follows.

$$D_{pros} = \lambda_1 D_p + \lambda_2 D_i + \lambda_3 D_d \tag{2}$$

, where λ_x ($x = i,p,d$) is weighting parameters for pitch, intensity and duration. And, for each of prosodic information, we use the Mahalanobis distance, which is defined as

$$D_x = \left| \frac{x_x - x_{ref}}{\sigma_x} \right|^2 \tag{3}$$

, where x means one of three prosodic parameters. Now we consider phonetic distance D_{phon} . D_{pros} is composed of the preceding and the following phonetic environments. So, D_{phon} is

$$D_p = D_{pP} + D_{pF} \tag{4}$$

, where D_{pP} is a distance for the preceding phonemes and D_{pF} is a distance for the following phonemes. By the way, each preceding or following environment has two visemes of C and V. So the distances of D_{pP} and D_{pF} are

$$D_{pP(F)} = D_C + D_V . \tag{5}$$

To calculate D_C and D_V , we should choose the parameters for calculating distances between the consonants or the vowels. In this paper, we used the features of coarticulation place and method. That is, the coarticulation method for vowel is front/back, high/low, round and etc. So, the features for visemes are not real-valued. Thus, we used the Hamming distance function for D_{pP} and D_{pF} . The distance for D_{phon} is the number of features having different class. The table 3 is an example of the features of vowels.

Table 3. Korean vowel (monophthong) and features

Korean vowel	Front/back	High/low	Round
(a)	Central	Low	
(eo)	Central	Middle	
(o)	Back	Middle	Round
(u)	Back	High	Round
(eu)	Central	High	
(i)	Front	High	
(ae)	Front	Low	
(e)	Front	Middle	
(oe)	Central	Middle	Round

3.2 Determination of optimum path

The figure 5 shows the series of top-N candidates. So, the total number of possible cases is M^N where M is the number of the target units.

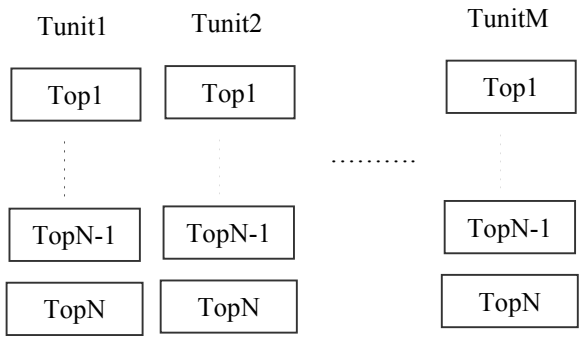


Figure 5. Sequence of top-N candidates

The problem is how we determine the optimum path from M^N cases. In this paper, we adopted Viterbi search algorithm to solve this problem. To implement Viterbi search, the concatenation cost should be defined. For that purpose, we use a Euclidean distance between the lip parameters at the boundary frames of the adjacent CVC-syllable instances. That is,

$$D_{ij} = \frac{1}{P}\sum_{0}^{P-1}(y_{jk} - y_{ik})^2 \ . \qquad (6)$$

D_{ij} : Distance between state *i* and state *j*

P : Number of lip parameters

y_{ik}, y_{jk} : k-th element of state *i* and state *j*

3.3 Recovery of lip movement parameters from PCA parameters

The concatenation of CVC-syllables results in the lip parameters versus the frame. This parameter should be converted to the original lip parameters and, for this, we use the PCA method for the data reduction. The conversion process is very simple and as follows.

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_{36} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_K \end{bmatrix}^T \begin{bmatrix} e_{1,1} \cdots e_{1,36} \\ e_{2,1} \cdots e_{2,36} \\ \cdots \\ e_{K,1} \cdots e_{K,36} \end{bmatrix} \qquad (7)$$

, where K is the reduced dimension of the lip parameters and $\lfloor e_{i,1} e_{i,2} \cdots e_{i,36} \rfloor$ is the *i*-th main axis of PCA analysis.

4. Experimental Results and Discussion

In our proposed method, there are some weighting parameters to be determined beforehand. Those are intra-prosodic weighting parameter λ_x and inter-weighting parameter w. There are 4 parameters to be adjusted. In this paper, we determine λ_x using the correlation analysis between three prosodic parameters and lip information. The used λ_x is

$$\lambda_p = \frac{\rho_{p,v}}{|\rho_{p,v}| + |\rho_{i,v}| + |\rho_{d,v}|} \qquad (8a)$$

$$\lambda_i = \frac{\rho_{i,v}}{|\rho_{p,v}| + |\rho_{i,v}| + |\rho_{d,v}|} \qquad (8b)$$

$$\lambda_d = \frac{\rho_{d,v}}{|\rho_{p,v}| + |\rho_{i,v}| + |\rho_{d,v}|} \qquad (8c)$$

Only the weighting w is left. Thus, we can determine the weighting w by the computer simulation. First, we can obtain the error values for various weighting values. The step size is 0.1. Then, the weighting w against the smallest error is the estimated optimum value.

4.1 Is only duration information useful?

Upon previous works, the main parameters for lip sync are duration and phonetic information. In this paper, we experimented when pitch and intensity information was not available. At first, only duration was considered. That is, $\lambda_i = \lambda_p = 0$ and $\lambda_d = 1$. Figure 6 show the average total square error depending on the weighting w.

According to the above result, the optimum weighting is 0.3 and the smallest error is 53. The following figure shows the results of the case that all the prosodic information is used. From the figure, it is observed that the smallest error is 48 and the optimum weighting is 0.3. Thus, figure 6 and 7 show that the performance is enhanced as much as 5/53=9.4%. The results tell us that pitch and intensity information is useful as well as duration in the lip sync problem.

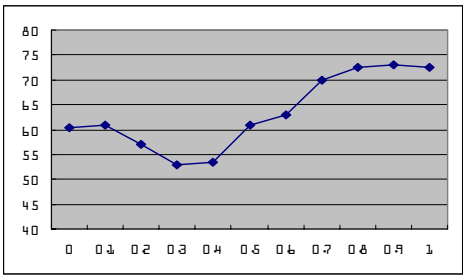


Figure 6. Weighting coefficient between prosodic and phonemic information for duration.

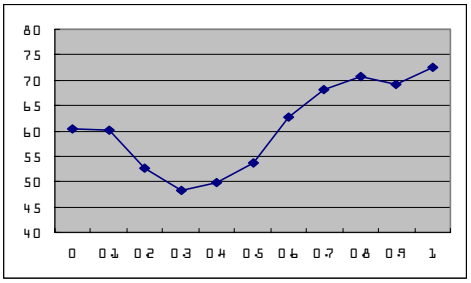


Figure 7. Weighting coefficient between prosodic and phonemic information for pitch, duration and intensity.

4.2 Example of recovering lip's parameters

Figure 8 show an example of recovering lip parameters. For convenience sake, the original lip heights and the estimated lip heights are compared. The two contours are very similar. Thus Figure (a) and (b) show that our proposed method is useful to solve lip sync problems.

Figure 9 shows the sample images of markers' position in case of utterance "amo"

5. Conclusion

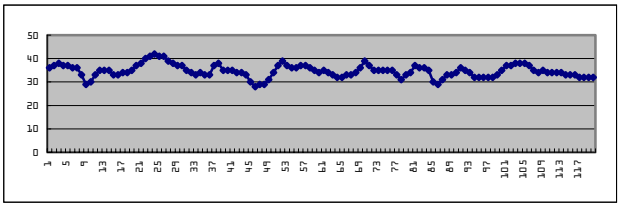
In this paper, a novel algorithm considering the lip-sync problem is described. An audio-video CVC-syllable DB is constructed based on CVC-syllable unit, prosodic information and lip parameters. As a concatenation of basic units, we proposed a method based on the similar concept of corpus-based TTS.

The simulation results show that our corpus-based method is useful to solve the lip-sync problem. Especially, It is shown that pitch and intensity information is not any more a garbage information in the lip-sync method. Adding this information to the conventional information of phoneme and its duration, we can improve the lip-sync performance of the corpus-based approach.

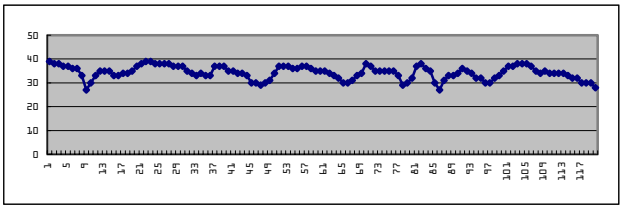
However, some works are left more to do in the future. In this paper, we used two distance measures and several weighting parameters. As a method of determining weighting values, we used the correlation values of prosodic parameters with the lip's width and height information. This method, however, may not be optimal. So, further study for how to determine the weighting values is necessary. Furthermore, our method needs to be compared with other previous works. Considering the complexity of our algorithm, our method needs much computational loads, which may not be an effective approach for lip-synchronization. Therefore, our algorithm needs such improvement more to be simple for the easy implementation.

Acknowledgement

Dr. Joohun Lee is sponsored as a postdoctoral fellow from *JSPS (Japan Society for the Promotion of Science)*.



(a) Original



(b) Synthesized plot of the lip height

Figure 8. The comparisons of the original and the synthesized lip information.

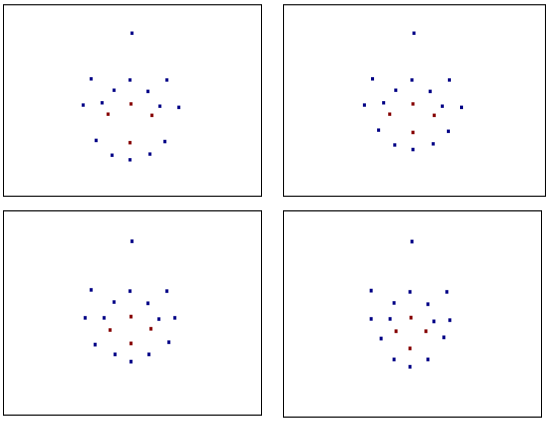


Figure 9. Recovered facial markers' point for a word ("a+m+o")

Reference

[1] W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise, " Journal of the Acoustical Society of America, vol.26, pp.212-215, 1954.

[2] Q. Summerfield, A. Mac Leod, M. McGrath and M. Brooke, "Lips, teeth, and the benefits of lipreading, " in Handbook of Research on Face Processing, A. W. Young and H. D. Ellis Editors, Elsevier Science Publishers, pp.223-233, 1989.

[3] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices, "Nature, Vol.264, pp.746-748, 1976.

[4] V. A Kozhevnikov and L. A. Chistovich, "Rech: artikulyatsiya I Vospriyatiye (Moscow-Leningrad, 1965). Trans. Articulation and perception". Washington, D.C : Joint Publication Research Service, Vol.30, pp.543,1965.

[5] F. Bell-Berti, and K. S Harris, "Anticipatory coarticulation: Some implications from a study of lip rounding," Journal of the Acoustical Society of America, Vol.65, pp.1268-1270,1982.

[6] J. S. Perkel, and C. Chiang, "Preliminary support for a hybrid model of anticipatory coarticulation," Proceedings of the 12th International Conference of Acoustics, A3-6, 1986.

[7] M. M Cohen and D. W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech", Models and Techniques in Computer Animation, N. M. Thalmann & D. Thalmann (Eds.), Tokyo : Springer-Verlag, pp.139-156.1993.

[8] C. Benoit, Tahar Lallouache, T. Mohamadi, and C. Abry "A set of French visemes for visual speech synthesis," Talking Machines : Theories, Models & Designs

[9] C. Benoit, C. Abry, M. A. Cathiard,, T. Guiard-Marigny "Read my Lips : Where? How? When? and so... What?," in Poster Book of the 8th international Congress on Event Perception and Aciton, July 9-14, 1995, Marseille, France.