Visual Analysis of Lip Coarticulation in VCV Utterances

Aseel Turkmani, Adrian Hilton, Philip J.B. Jackson and James Edge

Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK
{a.turkmani,a.hilton,p.jackson,j.edge}@surrey.ac.uk

Abstract

This paper presents an investigation of the visual variation on the bilabial plosive consonant /p/ in three coarticulation contexts. The aim is to provide detailed ensemble analysis to assist coarticulation modelling in visual speech synthesis. The underlying dynamics of labeled visual speech units, represented as lip shape, from symmetric VCV utterances, is investigated. Variation in lip dynamics is quantitively and qualitatively analyzed. This analysis shows that there are statistically significant differences in both the lip shape and trajectory during coarticulation.
Index Terms: Coarticulation, Visual speech analysis

1. Introduction

One of the most common approaches for speech synthesis is based on the concatenation of speech units, commonly phonemes, to produce a novel speech output. In the same way visual synthesis of speech can be the rendering of visemes, the visual shape and appearance of the face associated with the pronunciation of a phoneme. Numerous application areas benefit from visual speech synthesis including facial animation, human-computer interfaces, research into audiovisual speech perception, speech therapy and telecommunication.

Coarticulation is the variability of an articulators pose, dependent on context, caused by the assimilation of a speech unit to a preceding unit. Variability due to articulatory planning affects the subsequent speech unit. Thus, coarticulation is bidirectional. Many theories of coarticulation have sought to encode the relationship between articulatory planned sequence of speech gestures and their physical realisation [12] [9] [15] [7] [10] [17]. The constraints of physiology, effort minimisation, linguistic contrast and inter-articulator coordination all affect the average articulatory behavior and its variability. Most of these studies have been based on acoustical data and analysis of formant transitions. In recent years, as greater amounts of articulatory data have become widely available, researchers have concentrated on statistical approaches that allow the properties of articulatory configurations to be learnt from annotated measurements [16] [14] [6].

Visual speech synthesis provides examples of coarticulation modelling in a practical setting. In [3] [5] static phonetic units are blended to generate synthetic articulatory trajectories. The parameters of the blending functions used in these models determine how the lips move between viseme targets. Such models are similar to the theoretical models of [16] [13] in how they attempt to synthesize speech movements from discrete phonetic targets. Another popular method, used in both audio and visual speech synthesis, is the concatenative approach. In these models, short segments of real speech (e.g. syllables or triphones) are blended to generate synthetic trajectories. In terms of visual synthesis this approach has been demonstrated using both video [2] and motion-captured point trajectories [11] as the underlying speech data. Finally, in [8] an optimisation approach is used to fit a trajectory through targets represented as the mean and covariance of each viseme at its center.

Dynamics is especially crucial to the multi-modal perception of spoken utterances because our visual perception is highly sensitive to movement on human faces. We can detect minor occurrences of unnatural motion, e.g., from a discontinuity, from poor interpolation or from physiologically implausible gestures. The lips, jaw, teeth and occasionally the tongue are the only parts of the human vocal apparatus that are obviously visible in the face, so not all articulators are relevant in the study of visual speech.

In this work we look at the underlying dynamics of labeled visual speech units, represented as lip shape, from VCV utterances. The aim of this paper is to provide detailed quantitative statistical analysis, from ensemble data, of variation in lip dynamics due to coarticulation. Here, we use tracking of both inner and outer lip contours to describe the lip configuration and its visual appearance in video.

The remainder of the paper is structured as follows: Section 2 describes data acquisition. Section 3 shows the results of the analysis and Section 4 concludes our findings so far and discusses future work.

2. Data Acquisition

Data is acquired from color video, captured at a rate of 25Hz, of an English speaker, under uniform lighting conditions. The speaker is always facing towards the camera and there is minimal head movement. A rectangular area, (100x170 pixels), initialised manually, around the mouth is extracted from the video sequences. In our experiments we parameterise the lips (outer and inner lip contour) as a set of *N* landmarks, Figure 1. Each landmark on the lip boundaries is represented as a Cartesian coordinate. All landmark coordinates around the lip contours are concatenated to form a 1D vector, x.

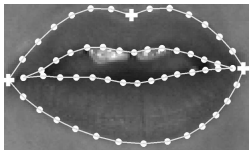

Figure 1: N landmarks of a parameterised lip. Crosses are reference landmarks.

The landmarks are labeled manually for half the number of frames, training data. The lip contours for the remaining frames are extracted using a standard Active Appearance Model (AAM) tool [4]. Three reference marker points around the lips form an axis of orientation (in the 2D plane) for the alignment of all the data. We translate, rotate and scale each lip shape so

as to minimise the sum of squared distances to the first frame.

For the work presented in this paper our analysis is based on the case of a voiceless bilabial /p/ in three VCV contexts. Three vowels /a/ (low back), /i/ (high front) and /u/ (high back and rounded) have been chosen. The three VCV nonsense words are placed in the carrier phrase "Is ... it?" The consonants in the carrier phrase form visually consistent stops (mouth open, wide and teeth showing and together). All three phrases are repeated 15 times (a total of 45 speech tokens). Lip data for frames associated with the VCV utterances are considered; frames associated with the carrier phrase are ignored. For this experiment there are a total of 481 frames, T, of lip shape data, x, accumulated into a matrix, X.

$$X = [x_t]', t = 1 \rightarrow T \quad (1)$$

Time synchronised audio is captured at 16 kHz. The audio stream is manually labelled into the known phonemes of the carrier phrases, using the audio waveform, wide-band and narrow-band spectrograms as a reference, Figure 2.

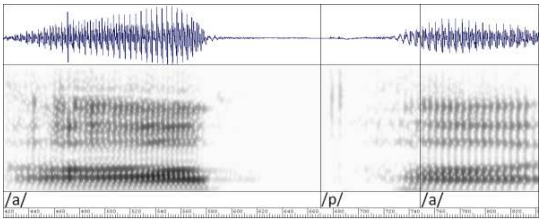

Figure 2: An example waveform and wide-band spectrogram for the phrase "Is apa it?"

The audio labels are used to time align the lip data with a common reference (the first carrier phrase utterance). The lip data is time aligned between the beginning and end of each phoneme in the utterance. The largest adjustment of any phone required a 18% scaling onto the reference utterance. The time alignment allows for comparison of lip data across all the utterances at any time instant.

3. Analysis

In the articulatory and acoustic domains, there have been attempts to provide statistical models of the dynamics of speech movements that account for coarticulatory effects. Coarticulations can exhibit dependency on neighbouring phonetic context, this can be modelled in three ways:

1. in the articulator's 3D coordinates,
2. in the relative configuration of articulators,
3. in terms of the articulators' motion over time.

The study presented in this paper investigates all three aspects for the bilabial plosive consonant /p/ which, together with /b/, is arguably the most crucial speech gesture in English from the perspective of visual impact. The pronunciation of /p/ involves context dependent combinations of lip movement, an interaction of lip and jaw positions and a rapid transition at release of the plosive, which is key to multi-modal perception of spoken utterances. Here, ensemble analysis is performed on the lip shape data acquired from a speaker pronouncing /p/ in three contexts.

The lip shape data obtained, X, is subjected to Principal Component Analysis (PCA). By calculating the principal components and removing those corresponding to low variance, the dimensionality of the feature vectors can be reduced. The lip shape consists of N lip points and the number of total frames is F resulting in an $N \times F$ matrix S. PCA is performed by calculating the eigenvectors of the $N \times N$ covariance matrix of S. The output of this is a small set of principal components (PC) and the variance of each component. In this experiment 12 PC's accounted for 98% of the total variation.

The work presented here is focused on the analysis of only the first two principal components, that account for 85% of the variance. The first principal component represents the general variation of the opening and closing and the widening and rounding of the mouth, 76% of the variance. The second PC represents a local variation of the inner-lip that displays protrusions associated with the rounding of the mouth, representing 9% variation. All other components represent subtle local variations, which are less significant for analysis, but may be considered for modelling in future work.

3.1. Analysis of the First Two Principal Components

The time-aligned shape data is projected onto the first two principal components, Figure 3. A polynomial is fitted onto the data to allow for a continuous interpolation so that the mean and variance of the data could be found at any time instant, as shown in Figure 5. It has been found that a cubic polynomial provides a good fit to the observed data.

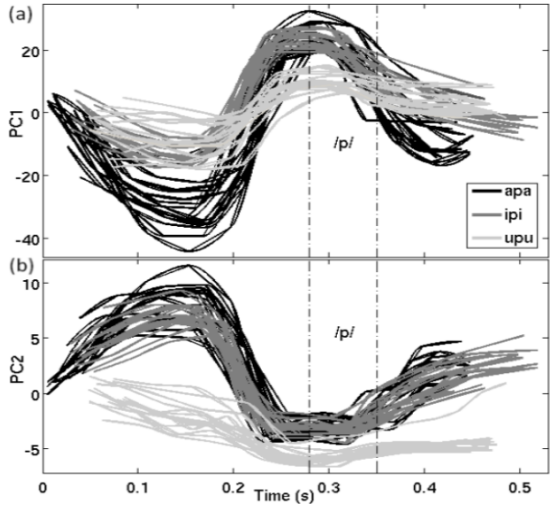

Figure 3: This figure shows all the data samples over time: (a) is for the 1st PC and (b) for the 2nd PC.

Figure 4 (a) shows how the data for the three VCV phrases, /apa/, /ipi/ and /upu/, vary over time for the first component. It can be seen that there is a separation between /apa/ and /upu/, particularly at the peaks and troughs that lie within the intervals of the phonemes. There is also some separation between /ipi/ and /upu/ in the time interval leading to lip closure. A separation also occurs between /apa/ and /ipi/ in the region of time where the first vowel occurs. The onset of the phoneme /p/ seems to be delayed for /apa/ than for the other two utterances. This indicates that the duration of the first /a/ is greater than the first vowel for /ipi/ and /upu/.

Figure 4 (b) shows how the data for the three VCV phrases, varies over time for the second component. It can be seen that there is a separation between /apa/ and /upu/ and between /ipi/ and /upu/ across the entire time interval that the utterances oc-

cur. For this component the spread of data for /apa/ and /ipi/ is almost identical.

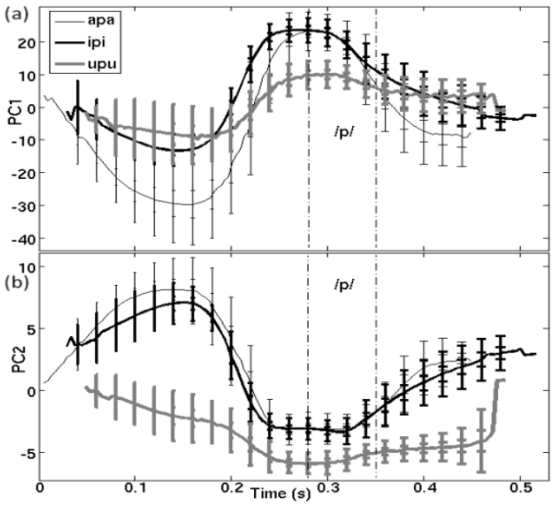

Figure 4: The mean (continuous lines) and variance for ± 3 standard deviations(vertical bars) of the data, over time: (a) is for the 1st PC and (b) for the 2nd PC.

3.2. Linear Discriminant Analysis

To quantitively measure which principal component gives the greatest separation between the three contexts of the phoneme /p/, Fisher's linear discriminant analysis (LDA) is applied [1]. LDA, a widely used technique for pattern classification, finds the linear boundary that yields optimal discrimination between two classes. Data is projected onto a line and classification is performed in the 1D space. The projection maximises the distance between the means of the two classes, while minimising the variance within class. LDA uses the projection that maximises the following ratio:

$$J(\Theta) = \arg\max_{\Theta} \frac{\Theta^T S_B \Theta}{\Theta^T S_W \Theta} \tag{2}$$

where S_B is the *between classes scatter matrix*, S_W is the *within classes scatter matrix* and Θ is the covariance matrix of all the data. The *between classes scatter matrix* is defined as:

$$S_B = \sum_{c=1}^{N_c} p_c(\mu_c - \overline{x})(\mu_c - \overline{x})' \tag{3}$$

where, μ_c is the mean vector for class c, p_c is the fraction of data belonging to class c and $\overline{x}$ is the mean of all data. We define the *within classes scatter matrix* as:

$$S_W = \sum_{c=1}^{N_c} p_c \Theta_c \tag{4}$$

where Θ_c is the covariance matrix of class c.

$$\Theta_c = \frac{1}{N_c} \sum_{c=1}^{N_c} (\overline{x}_c - \mu_c)(\overline{x}_c - \mu_c)' \tag{5}$$

where $\overline{x}_c$ is the mean of the data associated with class c.

This class discrimination is applied to the data projected on all principal components. Table 1 shows the total measure of discrimination for the three classes of /p/. It can be seen that the greatest discrimination occurs when /upu/ is compared to /ipi/. It was observed that for comparison, the greatest discrimination is found when data is projected onto the second principal component.

Table 1: Measures of separation, between /apa/, /ipi/ and /upu/.

	/apa/	/ipi/	/upu/
/apa/	0	3.5	10.7
/ipi/	3.5	0	19.6
/upu/	10.7	19.6	0

The variation in the separation between these classes in the first two principal components is presented in Figure 5. Figure 5(a) shows how the measure of discrimination varies over time between the three contexts when data is projected onto the first principal component. It can be seen that the greatest separations occur in the time interval that corresponds to the movement of the lips to closure associated with /p/. Figure 5(b) shows a similar discrimination when data is projected into the second principal component. However, in this case there is greater separation between the classes in the regions corresponding to the vowels in the VCV utterances, particularly for the time interval associated with the first vowel.

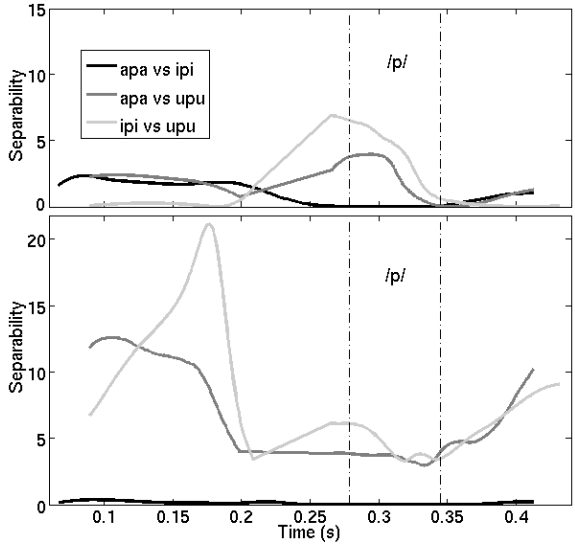

Figure 5: Variation, in terms of Fisher's class discrimination measure, over time, when the data is projected onto (a) the 1st PC and (b) the 2nd PC.

3.3. Qualitative Analysis

Using PCA, the lip shapes can be reconstructed as a weighted sum of the principal components. Figure 6 shows how the lip geometry varies over time. This provides a qualitative visual analysis of the difference between the lip shapes associated with the coarticulations /apa/, /ipi/ and /upu/. The first column shows how the lips generally vary over time for the first principal component. The second column shows variation in lip shape for the second component. The difference in lip shape between the coarticulations is visible, particularly at time stamps within the regions of the vowels.

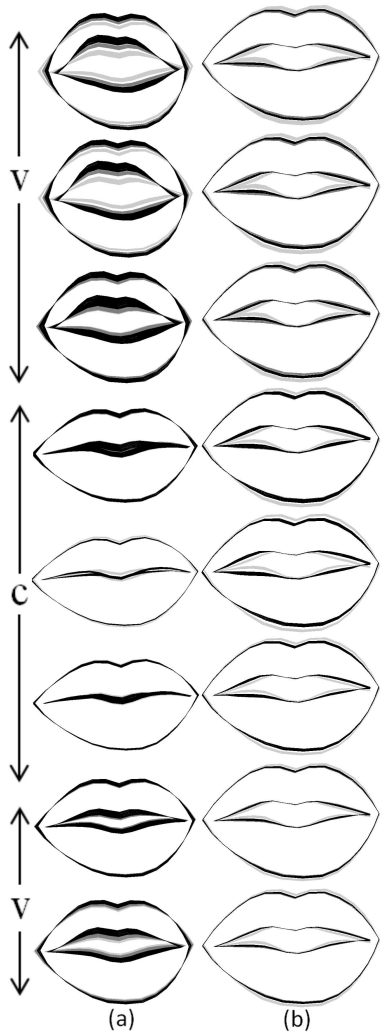

Figure 6: A sequence of the reconstructed lip shapes, frame rate, along the VCV time interval, for (a) 1st PC, (b) 2nd PC. Black represent variation (± 3 s.d.) for /apa/, dark grey lines for /ipi/ and light gray for /upu/.

4. Conclusions and Future Work

The work presented in this paper investigates the visual variation on the bilabial plosive consonant /p/ in three coarticulation contexts. The effect of coarticulation is analysed based on ensemble analysis of repeated utterances of symmetric VCV coarticulation to derive the temporal mean and variance characteristics. Results show that temporal influence of coarticulation is significant both in lip shape variation and timings of lip movement during coarticulation. Linear discriminant analysis of different VCV utterances shows that dynamic separations exist, in terms of lip shape. It can be concluded that the effect of temporal variation due to coarticulation is statistically significant and should be taken into account in modelling visual speech synthesis.

In the future we aim to extend our work to look at inter person variation. We aim to build dynamic models of visual speech for distinct utterances, for each principal component. We plan to expand our analysis to the voiced bilabial stop /b/ and nasal /m/, commonly grouped with /p/ in the same visemic category, to see if dynamic patterns found in /p/ overlap for /b/. We also plan to extend our analysis to some alveolars, labiodentals interdentals and alveopalatals.

5. References

[1] S. Balakrishnama, Ganapathiraju, “Linear Discriminant Analysis - A Brief Data”, Institute of Signal and Information Processing, 1998.

[2] C. Bregler, M. Covell, and M. Slaney, “Video Rewrite: Visual Speech Synthesis from Video”, Proceedings of the Workshop on Audio-Visual Speech Processing, 1997, pp. 353-360.

[3] M.M. Cohen, and D.W. Massaro, “Modeling and coarticulation in synthetic visual speech”, In Thalmann N. M. and Thalmann D., editor, *Models and Techniques in Computer Animation*, Tokyo: Springer-Verlag, 1993.

[4] T. F. Cootes, G. J. Edwards and C. J. Taylor, “Active appearance models”, Proceedings of the European Conference on Computer Vision, 1998, pp.484-498.

[5] E. Cosatto , H. Graf, “Sample-Based Synthesis of Photo-Realistic Talking Heads”, Proceedings of Computer Animation, , 1998, p.103-110.

[6] J. Dang, J. Wei, T. Suzuki and P. Perrier, “Investigation and modelling of coarticulation during speech”, Eurospeech, Lisbon, 2005.

[7] R. Daniloff and R. Hammarberg, “On defining coarticulation”, Journal of Phonetics, Volume 1, 1973, pp. 239-248.

[8] T. Ezzat, G. Geiger, and T. Poggio, “Trainable Videorealistic Speech Animation”, Proceedings of ACM SIGGRAPH, 2002, pp. 388-398.

[9] W.L. Henke, “Dynamic articulatory model of speech production using computer simulation”, MIT, Cambridge, MA, 1965.

[10] P.A. Keating, “The window model of coarticulation: articulatory evidence”, UCLA Working papers in Phonetics, Volume 69, 1988, pp. 3-29.

[11] S. Kshirsagar and N. Magnenat-Thalmann, “Visyllable Based Speech Animation”, Computer Graphics Forum 22(3), 2003, pp. 631639.

[12] B. Lindblom, “Spectrographic study of vowel reduction”, Journal of Acoustical Society of America, Volume 35, 1963, pp. 1773-81.

[13] A. Löfqvist, “Speech as audible gestures”, In W.J. Hardcastle and A. Marchal (Eds) Speech Production and Speech Modelling, Dordrecht: Kluwer Academic Publishers, 1990, pp. 289-322.

[14] A. Löfqvist and V.L. Gracco, “Interarticulator programming in VCV sequences: Lip and tongue movements”, Journal of the Acoustical Society of America, Volume 105(3), 1999, pp. 1864-1876.

[15] S.E.G. Öhman, “Coarticulation in VCV utterances: Spectrographic measurements”, Journal of Acoustical Society of America, Volume 39(1), 1966, pp. 151-68.

[16] S.E.G. Öhman, “Numerical model of coarticulation”, Journal of Acoustical Society of America, Volume 41(2), 1967, pp. 310-20.

[17] E.L. Saltzman and K. Munhall, “A dynamic approach to gestural patterning in speech production”, Ecology Psychology, Volume 1(4), 1989, pp. 333-82.