Statistical Correlation Analysis between Lip Contour Parameters and Formant Parameters for Mandarin Monophthongs

Junru Wu[1], *Xiaosheng Pan*[1], *Jiangping Kong*[1], *Alan Wee-Chung Liew*[2]

[1]Linguistic Prosody Laboratory, Dept. of Chinese Language & Literature,
Peking University, Beijing, China
[2]School of Information & Communication Technology, Griffith University,
Gold Coast Campus, QLD 4222, Australia

Email: yuerr@163.com, itol_xs@pku.edu.cn, kongjp@gmail.com, a.liew@griffith.edu.au

Abstract

In this study we examine quantitatively the correlation between the geometric lip contour parameters and the formant parameters for Mandarin monophthongs, and carry out a multiple linear regression study between the two parameters. We explicitly analyze the relationship between different geometric lip parameters and the formant parameters, which have some linguistic significance instead of the usual acoustic parameters such as MFCC. We analyzed the linguistic meaning of the regression formula, and found it in accord with the classical result on the relationship between vocal-track and speech acoustics. And those regions with relatively poor effect of estimation are related to specific phonetic conditions.

Index Terms: facial motion, monophthong, canonical correlation analysis, multiple linear regression

1. Introduction

The correlation between acoustics and visual features of speech is a fundamental issue in the field of audiovisual speech processing and an important aspect in phonetics. To examine this correlation, effective methods for the extraction and quantification of visual lip parameters, as well as statistical models are necessary.

In this study we examine quantitatively the correlation between the geometric lip contour parameters[1] and the formant parameters for Mandarin monophthongs, and carry out a multiple linear regression study between the two sets of parameters.

Several questions associated with speech acoustics and the geometric lip contour parameters are addressed in this paper: (1) what is the explicit relationship between different geometric lip parameters and formant parameters? (2) Is the non-linearity that exists between the two sets of features (lip parameters and formant parameters) distributed randomly or in accordance with some phonetics rules?

The widely cited study of Yehia, et al in 1998 [2] examined the degrees of correlation among vocal-tract and facial movement data and the speech acoustics. Using two corpuses of sentences in two languages, the 3D position data of markers placed on the face and in the vocal-tract was extracted. LSP coefficients and RMS amplitude of the signal were extracted from the acoustic signals in a separate experimental session. After temporal aligning the frames of different sets, for each two set of parameters linear estimation was carried out to build an estimator to recover a matrix, and then the mean correlation coefficient between measured and recovered vectors was measured through all possible combinations of training and test data. It was noted that estimation of the speech acoustics (f) from facial measures (x) is considerably better than from vocal-tract measures (y). Then, dimensionality analysis was carried out. Principle Component Analysis was used to reduce the dimensionality. However, there is not a priori reason to believe that the dimensionality reduction achieved in one space is optimum for describing the data measured in another space. Thus the authors perfomed singular value decomposition to map the data for the vocal-tract, facial and acoustic spaces onto a common coordinate system. It shows that between 4 to 8 components are sufficient to represent the mappings examined. In the discussion, the authors noted that in their study no temporal analyses were done; all correlations are based only on spatial properties of the data. They believe that the resulting global correlations suggest that correlated tongue-jaw behavior is basic to producing all speech rather than the result of some higher level phonetic control.

In later studies other data sources and parameters were tried. 2D face motion data has been examined by Almajai[3], Barker[4], Barbosa[5] and Jiang[6]. Some researchers tried to combine the study with visual feature extraction technique, for example 2-D DCT and cross-DCT (Almajai et al.) and 'Chroma-Key' processing (Barker et al.). As for Yehia's later study with Barbosa, a search algorithm is used for tracking the 2D facial motion of markers painted on the speaker's face.

As for acoustic data, Barker[4] tried LP, LSP and RMS parameters, and showed that the strongest correlations are achieved using the LSP parameterization. Almajai et al. used mel-scale filterbank vectors and the first four formant frequencies extracted using a combined linear predictive analysis-Kalman filter. Formant frequencies are closely related to speech production and correspond to resonant frequencies in the vocal tract. It has been shown that filterbank features exhibit higher correlation to visual features than formant frequencies. However, mel-scale filterbank parameters are hard to explain linguistically. Moreover, there is a lack of good method to extract linguistically meaningful formants in from these parameters.

As Yehia et al. [2] mentioned in their study, certain type of non-linearity exists between the acoustics and the visual features of speech. As Barker et al. [4] mentioned, Examination of the error distributions of the LP parameters reveals them to be multimodal i.e. clearly non-Gaussian. This shows that the linear estimates are essentially an inadequate model of the true mapping. No temporal analyses were done in [2]. Later studies tried to address these two points. Barbosa and Yehia[5] used time-invariant and time-varying linear models, as well as nonlinear (neural network) models

(Levenberg-Marquardt algorithm) in their study. As a result, the correlation coefficients between measured and estimated trajectories are as high as 0.95. This estimation of facial motion from speech acoustics indicates a way to integrate audio and visual signals for efficient audiovisual speech coding. On the other hand, relatively little studies from the perspective of linguistic phonetic principles have been done on this subject. In fact, there is complicated structure inside the so called 'non-linearity' related to linguistic phonetic notions.

Barker et al. [4] used a corpus of isolated nonsense words having a VCCV vowel-consonant structure (the systematic structure of the corpus allows the audio-visual correlation to be separately analyzed for both consonants and vowels). When examining the size and distribution of the errors, it is found that the lips can be more reliably reconstructed during vowels than during consonants. Jiang et al. [6] noted that, in general, predictions for CV syllables are better than those for sentences. Almajai et al. [3] measured the audio-visual correlation within each phoneme and then averaging the correlation across all phonemes and compared the result to the measurement across the whole corpus of sentences. As a result, there is an increase in correlation to R=0.9 when the audio-visual correlation is measured within each phoneme. All these indicated that the linearity of correlation is higher inside each phoneme than across different phonemes. In other words, different phonemes have different audio-visual correlations. A universal model would do well in some phonemes but do quite poorly in other phonemes.

2. Experiment

2.1. Database

The experiments for data acquisition are carried out for one female speaker of Chinese Mandarin. The data are acquired using a corpus of 61 Mandarin initials.

Mandarin has 22 initials. We choose 3 monophthongs /a, i, u/ for each initial and get 61 syllables (see Table 1). For those syllables that don't exist in Mandarin phonemic system the vowels /i/, /ʅ / or /ʅ / /y/ are used instead.

G1	b /p/	p /ph/	m /m/	f /f/	**G2**	d /t/	t /th/	n /n/	l /l/
a	ba1	pa1	ma2	fa1	a	da1	ta1	na2	la1
i	bi1	pi1	mi2	—	i	di1	ti1	ni2	li1
u	bu1	pu1	mu2	fu1	u	du1	tu1	nu2	lu1

G3	g /k/	k /kh/	h /x/	**G4**	j /tɕ/	q /tɕh/	x /ɕ/	**G7**	Ǿ
a	ga1	ka1	ha1	a	jia1	qia1	xia1	1	a
i	—	—	—	i	ji1	qi1	xi1	2	i
u	gu1	ku1	hu1	u	ju1	qu1	xu1	3	u

G5	zh /tʂ/	ch /tʂh/	sh /ʂ/	r /ʐ/	**G6**	z /ʦ/	c /ʦh/	s /s/
a	zha1	cha1	sha1	—	a	za1	ca1	sa1
u	zhu1	chu1	shu1	ru4	u	zu1	cu1	su1
/ʅ /	zhi1	chi1	shi1	ri4	/ɿ /	zi1	ci1	si1

Table 1: The 61 syllables used in the study (Group1 (G1) Bilabial, Group2 (G2) Alveolar, Group3 (G3) Velar, Group4 (G4) Coronal, Group5 (G5) Retroflex, Group6 (G6) Alveolar2)

These AVI clips are part of an existing audio-visual corpus recorded by our laboratory in Peking University. The two domains of data are taken simultaneously using Adobe Premiere Prof 1.5 in AVI format (Video: 720×576 pixel, 24bits, 25 fps; Audio: 639kbps, PCM 16bits, 32 kHZ, 1024kbps) and then separated using Virtual Dub 1.7.0.1c1.x by Avery Lee. Then the Audio was re-sampled to 11025 HZ, and the area of Lip (96×80 pixel) was extracted from the original video.

2.2. Parameterization

At this point, the data available are the audio and video signal. This section describes suitable parametric representations that will help in the study of the relationship between the two domains.

2.2.1. Lip contour extraction

The lip contour parameters are characterized by using the deformable template by Liew et al. [1] from video images of the speaker's face. The parameters of the model are adjusted manually to correct any fitting error since the aim of this study is to investigate the relationship between visual lip features and monophthongs, instead of lip segmentation methodology.

In [1], a robust deformable model-based technique for lip contour extraction from a color RGB lip image is proposed. The method uses a region-based stochastic cost function to find an optimum partition of a given lip image into lip and nonlip regions. Spatial fuzzy clustering using both luminance and chrominance features from the CIELAB and CIELUV color spaces is used to produce a probability map of the lip image. The optimum model parameters are then found by performing a conjugate gradient search on the cost function. Extensive experimental results show the feasibility of their approach.

Given a lip model as shown in Figure 1 , the equations describing the lip shape are given by [1]:

$$y_1 = \frac{-h_1}{(w - x_{off})^2}\left(|x - sy_1| - x_{off}\right)^2 + h_1 \tag{1}$$

$$y_2 = h_2\left(\left(\frac{x - sy_2}{w}\right)^2\right)^{1+\delta^2} - h_2 \tag{2}$$

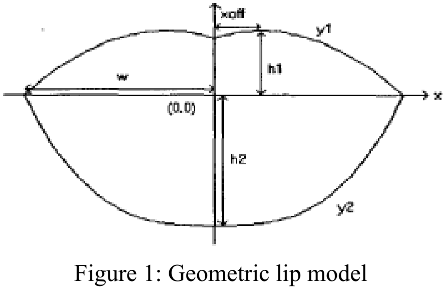


Figure 1: Geometric lip model

The parameters we used are a little different from the one presented in [1]. In this study we added the inner lip parameters into the lip model of [1]. Our lip model are as shown in Fig.2 and is parameterized as follows: the horizontal position of the lip in the rim (y1), the vertical position of the lip in the rim (y2), the width of the outer contour of the lip (y3), the distance from the lower outer contour to the level line (y4), the distance from the upper outer contour to the level line (y5), the degree of concavity of the philtrum (y6), the curvature of the lower lip counter (y7), the obliquity of the lip (y8), the width of the inner lip (y9), the distance from the lower inner contour to the level line (y10), the distance from the higher inner contour to the level line (y11), the Skewness of the lip (y12), enantiomorphism (y13). Finally the set of lip parameters are given by

Y= {y1, y2, y3, y4, y5, y6, y7, y8, y9, y10, y11, y12, y13} (3)

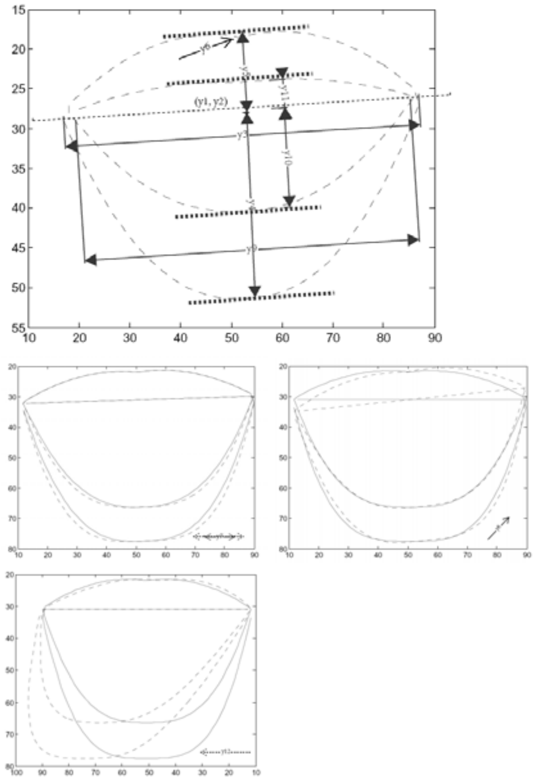


Figure 2: Geometric lip model (13 parameters). First plot: y1, y2, y3, y4, y5, y6, y9, y10, y11; Second plot: y7; Third plot: y8; Last plot: y12

2.2.2. *Formants*

The first three formant parameters (fmt1, fmt2, fmt3) are obtained from audio data taken simultaneously with the video data using the method of LPC (period in-phase), using *Wavefinal*, which is written by our laboratory. The boundary of the consonant and the vowel in a syllable is manually marked. Only the vowel frames are used:

F= {fmt1, fmt2, fmt3} (4)

2.2.3. *Alignment*

Since the video sampling rate of 25 fps is lower than the audio sampling rate (11025HZ), we performed appropriate time alignment of the two sets of parameters. The formants taken using LPC is period in-phase, and the interval between pulses is always changing, which is known as jitter. As a result, the frames of formant data don't have identical time length, and an indefinite number of formant frames are related to a frame of lip contour parameters.

We calculate the average of F through all the frames of formants and align it with the relative frame of lip contour parameters. The relative bandwidth (bw) is also extracted. For those formant frames at the beginning or end of the vowel with null, the next or the former data was filled into the null. Table 2 shows a frame of the aligned parameters.. 800 frames of data are obtained from the vowel parts of these syllables.

pinyin	mark_inx	y1	y2	y3	y4	y5	y6	
a1	1	47.9	38.6	34.33	39.43	11	6.84	…

y7	y8	y9	y10	y11	y12	y13	
1.1	0.03	33	25.95	0.62	0	0	…

fmt0	fmt1	bw1	fmt2	bw2	fmt3	bw3	lip mark_idx	position
239	1204	69	1702	94	3227	122	24	0.96

Table 2. Tthe first frame of data from /a1/

2.3. Canonical correlation analysis

Canonical correlation was first established by Hotelling in 1936 to analysis the correlation between two sets of random variables. The idea of reducing dimensionality is borrowed from PCA. The correlation between two sets of parameters is reduced to the correlation between two canonical variables.

We perform statistical matrix analysis on the two sets of parameters. The correlation analysis result for the set of lip parameters (Set1) shows that many of the lip parameter pairs have Pearson correlation coefficient larger than 0.5 (y2-y3,y2-y5,y2-y9,y3-y4,y3-y5,y3-y6,y3-y9,y3-y11,y3-y12,y4-y6,y4-y9,y5-y6,y5-y9,y5-y11,y6-y9,y6-y11,y6-y12,y9-y11,y9-y12). For the set of formant parameters (Set2), no Pearson correlation coefficient is larger than 0.5. This shows that there is much redundancy between the lip parameters whereas the formant parameters are largely independent. In the correlation between the set of lip parameters and the set of formant parameters, several pairs have Pearson correlation coefficient larger than 0.5 (y3-fmt1 (R=0.5712), y3-fmt2 (R=0.6109), y4-fmt1 (R=0.7571), y5-fmt2 (R=-0.5826), y9-fmt1 (R=0.5818), y9-fmt2 (R=0.6461), y10-fmt1 (R=0.7414)), indicating that there are strong relationships between the two sets of parameters. Canonical Correlation analysis is performed on the two sets of parameters and the results are verified using hypothesis testing (Wilk's and Chi-Sq. tests). The first canonical correlation coefficient (L1-F1) is 0.911 which is larger than any correlation coefficient between any two individual variables from Set1 and Set2. Standardized canonical coefficients for Set1 indicates that L1 = 0.018y1-0.039y2+ 0.328y3+ 0.119y4+ 0.049y5- 0.046y6- 0.013y7-0.086y8- 0.972y9- 0.531y10- 0.291y11+ 0.047y12, whereas standardized canonical coefficients for Set2 indicates that F1=-0.778fmt1-0.648fmt2+0.052fmt3.

Canonical loadings for Set1 shows that y3, y4 y6, y7, y8, y9, y10 have negative correlation to L1. Canonical loadings for Set2 shows that fmt1, fmt2, fmt3 have negative correlation to F1. Cross loadings for Set1 shows that y3, y4, y6, y9, y10, y12 can be better estimated by F1，y5 can also be estimated by F1 to certain extent. Cross loadings for Set2 shows that fmt1, fmt2 can be better estimated by L1. The difference of sign between canonical coefficient and canonical loadings of y3, y4, y11 indicates that they may be compensation parameters for y9, y10, y5.

Redundancy analysis shows that the proportion of variance of Set1 explained by its own canonical variant is 41.7%. The proportion of variance of Set2 explained by its own canonical variant is 34.3%. The proportion of variance of Set1 explained by opposite canonical variant is 34.6%. The proportion of variance of Set2 explained by opposite canonical variant is 28.5%. Hence, the formant parameters are better in explaining the lip parameters than vice versa.

Canonical correlation analysis is later carried out separately between the inner lip parameters (y9-y10-y11) and formant parameters or outer lip parameters (y3-y4-y5) and formant parameters. The parameters in the set are chosen in accordance with the above-mentioned CCA. It was found that the set of inner lip parameters is also better in explaining F than the set of outer lip parameters.

2.4. Multiple Linear Regression

Multiple linear regression (MLR) is used to fit the linear combination of the components of the multiple-dimension vector L, i.e. (y3, y4, y5) or (y9, y10, y11) as independent variables, to the single-dimension vector fmt1, or fmt2, or fmt3, which is the dependent variable. It is also used to fit the linear combination of the components of the multiple-dimension vector F (fmt1, fmt2, fmt3) as the independent variables to the single-dimension vector, i.e. y3, or y4, or y5, or y9, or y10, or y11, as the dependent variable. In this process, the method of stepwise regression is used to exclude the independent variables which do not fit the criterion and include the independent variables which contribute most to the dependency. So the independent variable fmt3 is excluded from the regression formula of F and y3, y9 is excluded from the regression formula of L (y9, y10, y11) and fmt1, y11 is excluded from the regression formula of L (y9-y10-y11) and fmt2. (see Table 3)

	formula	R	AESq	coll
1	fmt1=25.024y4+19.685y5+8.758y3-494.021	0.779	0.606	+
2	fmt2=89.974y3-34.583y4-47.513y5+296.385	0.683	0.645	+
3	fmt3=-46.532y5-15.672y4+12.456y3+3919.007	0.483	0.230	+
4	y3=0.005fmt2+0.013fmt1+14.838	0.838	0.701	
5	y4=0.022fmt1+0.004fmt2-0.001fmt3+10.729	0.821	0.673	
6	y5=-0.002fmt2-0.002fmt3-0.003fmt1+25.022	0.647	0.417	
7	fmt1=24.461y10+26.840y11+392.405	0.792	0.626	+
8	fmt2=65.313y9-41.547y10+578.949	0.698	0.486	+
9	fmt3=-45.985y11-19.298y10+14.908y9+3225.469	0.513	0.260	+
10	y9=0.011fmt2+0.028fmt1+0.002fmt3-18.616	0.872	0.760	
11	y10=0.026fmt1+0.005fmt2-0.001fmt3-12.254	0.858	0.736	
12	y11=-0.001fmt2-0.003fmt3-0.001fmt1+11.745	0.546	0.295	

Table 3. Regression formula, where ARSq denotes adjusted R-Square, coll denotes collinearity.

Similar to the result of canonical correlation analysis, the formant parameters works better than the lip parameters, and the inner lip parameters works better than the outer lip parameters as the independent variables in that the Adjusted R-Square is larger.

Figure 3 shows all the frames of virtual data (i.e., predicted from regression formula) and estimated data (LPC estimated) when the virtual data is ranked in ascending order.

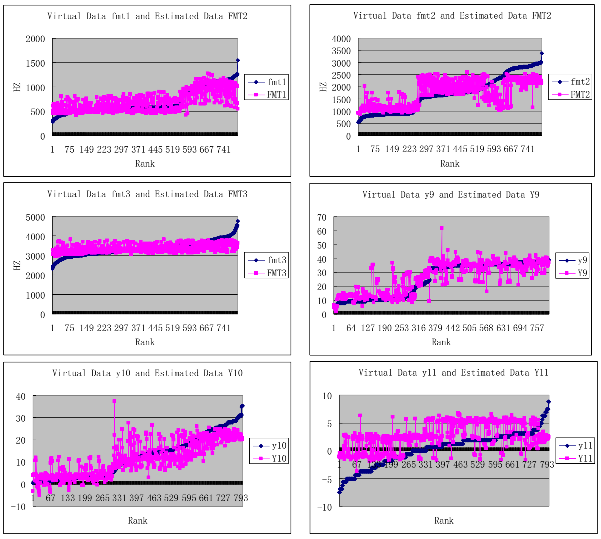


Figure 3. All the frames of virtual data and estimated data. The virtual data are ranked in ascending order.

3. Discussion

3.1. The phonetic meaning of the coefficients

The sign of the coefficients in the regression formula show that the larger the opening of the lower lip is, the larger the first formant is; the higher the opening of the upper lip is (from the level line), the larger the first formant is; the wider the lip is, the larger the second formant is; the smaller the opening of the lower lip is, the larger the second formant is; the wider the lip is, the larger the third formant is; the smaller the opening of the upper lip is, the larger the third formant is; the smaller the opening of the lower lip is, the larger the third formant is. The finding is in harmony with the study of speech articulation and formant frequencies in linguistic research, which state that the first formant is positively related to the opening aperture of the mouth, the second formant is negatively related to the posteriori of the tongue and the roundness of the mouth, and the second and the third formant come closer when the lip is rounded. Figure 4 shows the fmt1, fmt2, and fmt3 plot for different vowels, whereas Figure 5 shows the plot for different vowels as a function of lip parameters y9, y10, and y11. It can be seen that the vowels can be separated to some extend based on the three lip parameters, although the separation is not as good as that using fmt1, fmt2, and fmt3.

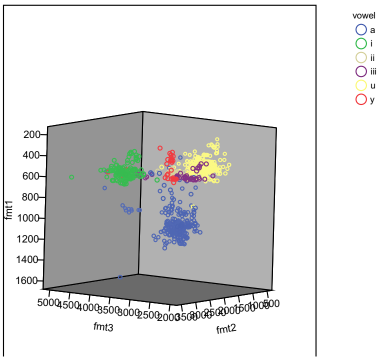


Figure 4. fmt2 as x-axis, fmt1 as y-axis, fmt3 as z-axis

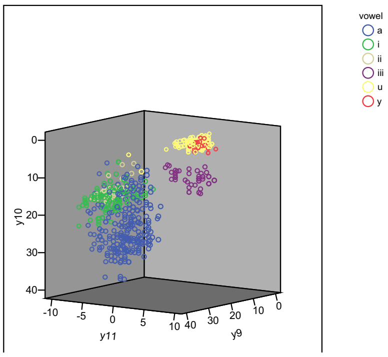
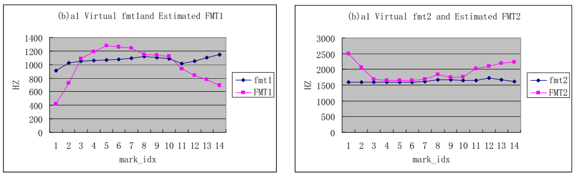


Figure 5. y9 as x-axis, y10 as y-axis, y11 as z-axis

3.2. The Distribution of Non-linearity

Although the MLR analysis reveals some useful insights, the linearity assumption has its limitation. First, we observed that the distribution of fmt1, fmt2 is far from being Gaussian. Second, there is noticeable co-linearity between y3 and y4, y3 and y5, y9 and y10, y9 and y11 and the set of outer lip parameters and inner lip parameters can be recovered from each other. Third, the distribution of the residuals is not Gaussian, indicating that there is residual correlation not extracted by MLR. This residual correlation may be due to the non-linear relationship between the two sets of parameters.

It is important to note that this non-linearity neither distribute randomly within a phoneme nor across different phonemes. Figure 6 shows some examples of the virtual and estimated formant data in specific vowels.



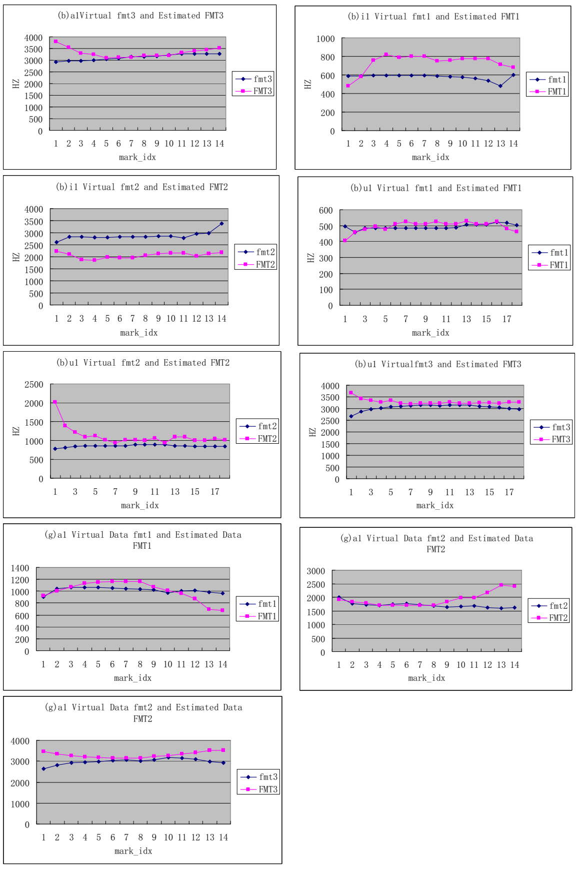


Figure 6 the virtual and estimated data in (b)a, (b)i, (b) u, and (g)a

First, the position of a specific frame in the vowel influences the relation between its lip and formant parameters. We could see from the figures that the linear estimators do quite well in the middle of the vowel but poorly at the beginning or at the end of the vowels. The estimated loci usually take the shape of an arc, which is related to the loci of the lip parameters which indicate the movement feature of the lips. While pronouncing, the muscles first move fast in acceleration to a specific point, then keep the state until the fast release comes. At the same time, the formants do not experience such dramatic change but keep a relatively stable state.

Second, the estimators perform differently depend on the interaction between the vowel and the initial circumstance. As for /a/, the changes of lips are very sensitive to the initial circumstance. As shown in the figure, the lip parameters change dramatically after (b)/p/, but not so much after (g)/k/, which lead to distinctly different estimated formant loci between these two /a/s, while the virtual formant loci are less different. This is obviously not a linear correlation. In /u/, the state of lips does not change much while pronouncing this vowel, regardless of the initial circumstance, so linear estimators fit well.

From the viewpoint of phonetics and the study of vocal tract, different articulation can be used to produce the same set of formants. For example, to produce /u/, which is marked with lower second and third formant, the speaker may have more posterior tongue position and less rounded lip or more rounded lip but less posterior tongue position. On the other hand, since the same lip shape may combine with different tongue position, different formant structures can be obtained.

For example, /u/ and /y/ have almost the same lip contour parameters but different formant set. This may explain why it is not as effective to estimate formants from lip contour parameters as vice versa. Hence, combining it with the study of vocal track may be a useful attempt. [7-9]

4. Conclusions

Canonical Correlation Analysis and Multiple Linear Regression are carried out across the lip contour parameters and formant parameters of Mandarin monophthongs and the effect is evaluated. It was observed that formant parameters works better than the lip parameters, the inner lip parameters works better than the outer lip parameters in estimation, and the first and second formants are better estimated than the third one. These are in accordance with former studies. It was also found that the phonetic meanings of the coefficients of MLR are in harmony with the study of speech articulation and formant frequencies in linguistic research. The distribution of non-linearity of the correlation is not random, but influenced by the position of the frame in the vowel and the interaction between vowel and initial circumstance. Future work would be to incorporate this findings into the automatic speech and lipreading application of Mandarin language.

5. References

[1] A.W.C. Liew, S.H. Leung, and W.H. Lau, "Lip Contour Extraction from Color Images Using a Deformable Model", Pattern Recognition, Vol. 35(2), pp. 2949-2962, 2002.

[2] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," Speech Communication, vol. 26, pp. 23-43, 1998.

[3] I. Almajai and E. Milner, "Maximising Audio-Visual Speech Correlation", proceedings of the Auditory-Visual Speech Processing 2007 (AVSP2007), Kasteel Groenendaal, Hilvarenbeek, The Netherlands, 2007.

[4] J. P. Barker and F. Berthommier, "Evidence of Correlation Between Acoustic and Visual Feature of Speech," Proceedings of the ICPhS '99, San Francisco 1999.

[5] A. V. Barbosa and H. C. Yehia, "Measuring the relation between speech acoustics and 2D facial motion", Proceedings of the IEEE Int. Conf. Acoustics, Speech, Signal Processing, Salt Lake City, Utah, USA, 2001.

[6] J. Jiang, A. Alwan, L. E. Bernstein, P. Keating, and E. T. Auer, "On the correlation between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1174-1188, 2002.

[7] 汪高武, 鲍怀翘, and 孔江平, "从声道截面积推导普通话元音共振峰," *中国语音学, 商务印书馆, 北京*, vol. 第一辑, 2008.

[8] G. Wang, T. Kitamura, X. Lu, J. Dang, and J.P. Kong, "MRI-based Study on Morphological and Acoustic Properties of Mandarin Sustained Vowels ", Journal of Signal processing, 2008, in press.

[9] G. Wang, "A Stuy of Mandarin Chinese Using X-ray and MRI," *Journal of Chinese Phonetics (中国语音学报)* , 2008.