Production of Mandarin lexical tones: Auditory and visual components

Virginie Attina[1], *Guillaume Gibert*[1], *Eric Vatikiotis-Bateson*[2], *Denis Burnham*[1]

[1] MARCS Auditory Laboratories, University of Western Sydney, Australia
[2] Department of Linguistics, University of British Columbia, Canada

v.attina@uws.edu.au, g.gibert@uws.edu.au, evb@interchange.ubc.ca, d.burnham@uws.edu.au

Abstract

This paper presents a study of audio-visual production of the four Mandarin lexical tones on words in citation form and in sentences. OPTOTRAK motion capture data of the head and face of a Mandarin speaker were modelled using both PCA and guided-PCA. For each tone, correlations between F0 values and the different face and head components were calculated. Results show that there are visual parameters related to the different F0 patterns of each tone. Moreover differences were found in both duration and correlational patterns between words produced in citation and in sentential forms. The results show that there are identifiable visual correlates of lexical tone but the difference between citation and sentential forms has implications for materials used in production and perception studies of Mandarin lexical tones, and possibly those in other languages.

Index Terms: audiovisual speech production, tone languages, Mandarin, OPTOTRAK, motion capture.

1. Introduction

It is well-known that speech perception is not only auditory but also visual [1]. A necessary corollary of this is that visual information in the production of consonants and vowels helps to disambiguate speech segments under certain conditions (for a review, see [2]). For example there is an acoustic confusion between /ba/ and /va/ syllables in noise which is solved when visual information is also provided [3].

Some languages like Mandarin Chinese use pitch variations in addition to consonants and vowels to distinguish among words in speech. These pitch modulations are called lexical tones and occur mainly on the vowel. Lexical tones are found in 70% of the world's languages [4], but are not so well studied compared to segments (consonants and vowels). Mandarin Chinese has four lexical tones which differ mainly in height and contour patterns of the fundamental frequency (F0) (see Figure 1): tone 55 is known as high level (Chao's description of tones [5]), tone 35 is mid rising, tone 214 is low dipping and tone 51 is high falling. In Mandarin we can find some quadruplet sets of words differing only in lexical tone; for example the word 'fu' (pinyin transcription) means 'husband' for 'fu55', 'to support' for 'fu35', 'to comfort' for 'fu214' and 'rich' for 'fu51'. While F0 variations are the main acoustic features for tones, other acoustic properties of tones such as duration and amplitude have been shown to differ (see for e.g. variations of durations in Figure 1) and to be perceptually salient as well (e.g. [6]).

Even though most studies of lexical tone have investigated the acoustic and auditory features of tone languages, there is evidence of visual effect for tones as well. Regarding lexical tone visual identification, Burnham and colleagues [7] have shown that under certain conditions the visual-only information is used to identify and differentiate the six different Cantonese lexical tones. For Mandarin, Mixdorff and colleagues [8] have shown that the visual information integrated with the auditory information helps native speakers to correctly identify tones but in babble-noise masked conditions only. In a training study, Chen and Massaro [9] have shown that native Mandarin speakers can efficiently exploit visual cues to identify tones when they are taught which visual information to pick up. After visual examination of the videos, the authors noticed that there are visible movements of the neck, the head and the mouth differing according to which tone is produced. However there was no quantitative measurement of that visual difference. One study of Cantonese using OPTOTRAK showed that the F0 values of tones were better correlated with rigid head movements than with non-rigid articulatory face movements and this rigid head movement information was used in the perception of lexical tones as well [10].

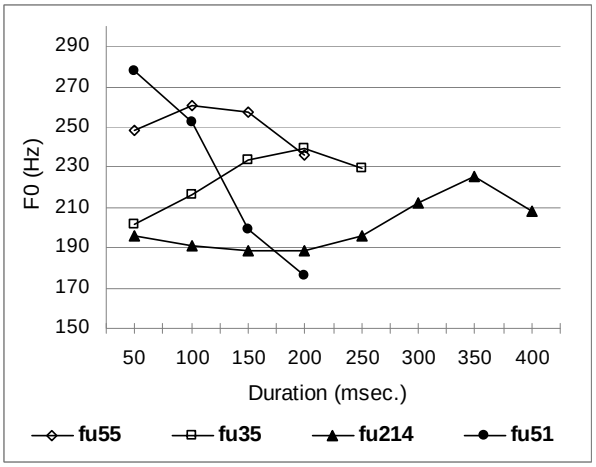

Figure 1: *F0 contours as a function of time for the four Mandarin tones of the word 'Fu'.*

The aim of the present study is to characterize the visual and articulatory features involved in the production of each of the four Mandarin lexical tones. In particular we would like to know if there are visual differences in Mandarin which are related to the properties of each tone, and in particular the F0 variations, and quantify these relationships. It should be noted that the data used in this paper are part of a larger project involving the production of Mandarin tones, Cantonese tones [10], Japanese accents and English stress. Only analyses of Mandarin data are presented here.

2. Methods

2.1. Speaker

A 38-year-old female native speaker of Mandarin was recorded. She had no hearing or vision deficits and had no speech production problems. She was born in Beijing and spent most of her life there. She was a speaker of standard Mandarin and had no noticeable accent from other Mandarin dialects.

2.2. Corpus

Mandarin words differing only in tones (with same segments) were used. The corpus was composed of six different syllables (wei, fen, fu, bao, cai, hui – pinyin annotation) pronounced with the four Mandarin tones giving a total of 24 different Mandarin words. Each word was repeated five times in citation form and in sentential context. For the words in sentences, there were five versions of lists of sentences containing all the 24 words, i.e. there were five sentential contexts for each of the four tones in each of the six segmental contexts.

2.3. Data Recording

The non-rigid face and rigid head motion data were recorded by 21 sensors placed on the face and the head of the speaker (see Figure 2) using two linked Northern Digital OPTOTRAK devices. These motion capture data were recorded at a sampling rate of 60 Hz. The time-aligned acoustic data were recorded and sampled at 16 kHz.

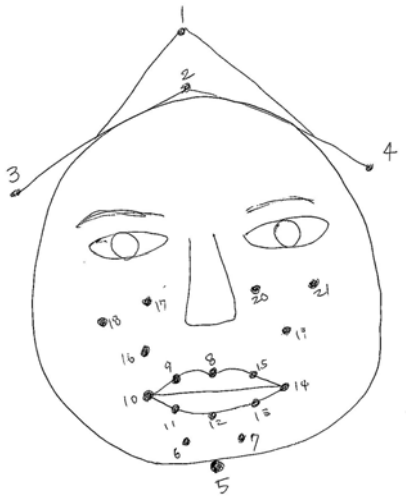

Figure 2: *Position of the 21 sensors (four sensors mounted on a crown for rigid motion estimation and 17 sensors on the face) during the motion capture data recording*

2.4. Data modeling and analysis

Regarding movement, the OPTOTRAK raw data (see the variance in Figure 3) were modelled using Principal Component Analysis (PCA) to estimate elementary rigid head movements. For the non-rigid face movements (lips and jaw), PCA was performed on pertinent subsets of points along specific directions, referred thereafter to as guided PCA [11]. The use of guided PCA for articulatory face movements has been shown to be pertinent for speech (e.g. [12]). However because little is known about head movements in speech and especially for tone languages, we decided to use simple (unguided) PCA for head movements to see what emerges from the data. To avoid taking into account too much rest position in the recording, a quantization that guaranteed a minimum 3-D distance between selected training frames (0.4 mm) was performed on every frame of all the OPTOTRAK data before building the model. An estimation of the centre of rotation of head movements was performed on initial wag trials so that head movements were removed from the rest giving two separate head and face data sets (see the remaining variance for face movements after removing the head motion in Figure 4). A total of 12 Principal Components (PCs) were used to model the data: six for the face movements (corresponding to jaw opening JawO, lip protrusion LPro, lower lip closing LLC, lip raising LR, lip closing LC, jaw advance JawA) and six for the head movements (PC1, PC2, PC3, PC4, PC5 and PC6). Note that because unguided PCA was used for the head movements, the six head PCs could be different components for words in citation form and words in sentences since the identity of these PCs is defined by the data. Once the model was built, each OPTOTRAK file was inverted to determine, for each production, the variation of the articulatory and rigid movement parameters as a function of time.

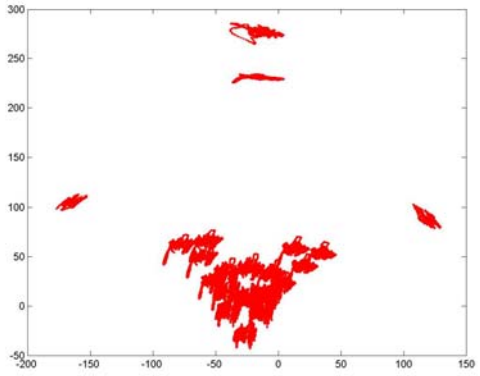

Figure 3. *Variance of the raw data including head motion for all the words in citation form*

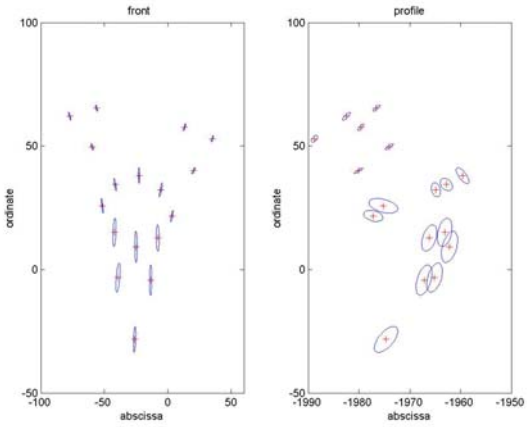

Figure 4. *Dispersion ellipses of the raw data without head motion for front and profile views.*

Analyses were also performed on the acoustic signal. All the words (both in citation and in sentential forms) were labelled using Praat [13]. Note that the phonetic labelling was done by a native Mandarin speaker who checked that the words and tones were correctly produced by the speaker. An additional evaluation by a second native Mandarin speaker made sure that the words kept for the analysis were accurate. A Matlab® function was written to import all the Praat TextGrid annotations into the Matlab environment and perform a series of analysis. Mean duration of words and standard deviations were calculated and compared. The F0 values for each word were estimated using the AMDF (average magnitude difference function) algorithm [14]. Correlations between each movement PC and the corresponding F0 were calculated as a function of the tone and the word produced.

3. Results

3.1. Principal Component Analysis

Table 1. *Percentages of explained variance by each PC for words in citation and sentential forms.*

Face PC	JawO	LPro	LLC	LR	LC	JawA
Words in citation form	66.85	17.99	3.87	4.09	1.08	1.61
Words in sentences	62.46	15.58	3.97	5.12	4.15	2.79
Head PC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Words in citation form	46.72	30.61	13.55	5.55	2.71	0.86
Words in sentences	50.55	19.75	11.75	9.49	5.75	2.71

The six PCs for the non-rigid face movements explain 95.49% of the total variance for words in citation form and 94.07% for words embedded in sentences. Table 1 shows the re-partition of the different percentages for each PC. Overall the differences are small between words in citation form and words in sentences. For example the first PC for the non-rigid movement JawO corresponds to jaw opening; from our data it appears that words produced in sentences involve less jaw rotation movements (62.46% of variance) than words produced in citation form (66.85%). To illustrate the motion involved in JawO, Figure 5 shows spatial variation of JawO modelled on words in citation form. For the rigid head movements, it is more difficult to directly compare the percentages for each PC because the data were modelled using a simple unguided PCA, so it is more likely that PC1 for isolated words does not correspond exactly to PC1 for words in sentences (see Figure 6). Nevertheless, what we can say is that the first rigid (head) predictor, PC1, accounts for slightly more variance (50.55%) for words in sentences compared to words in isolation (46.72%).

As an illustration of the head PCs obtained, Figure 6 shows spatial variation of all the head PCs modelled on words in citation form and embedded in sentences. As noted earlier, we can see that PC1 for words in citation form differs from the one for words in sentences: while it is mainly a back and forth movement of the head for isolated words, for words in sentences the head tends to move upwards. Differences are also noticed for the other head PCs. These differences might be explained by a phenomenon that could be called head motion "coarticulation" that affects the words when they are produced in sentences and probably a more general effect of the prosody of the sentence. Indeed it has been shown that head movements are strongly correlated to speech intonation for both English and Japanese sentences ([15] and [16]).

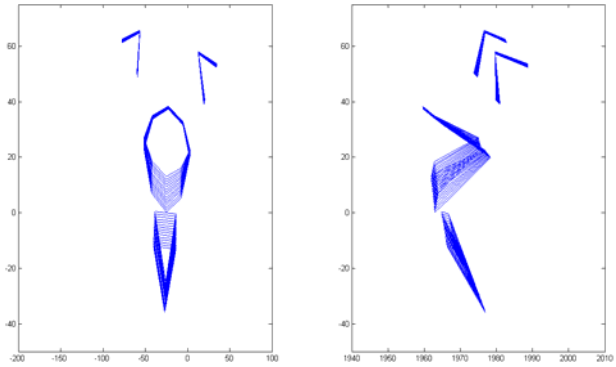

Figure 5. *Front view (left) and profile view (right) of the spatial variation involved in JawO (words in citation form).*

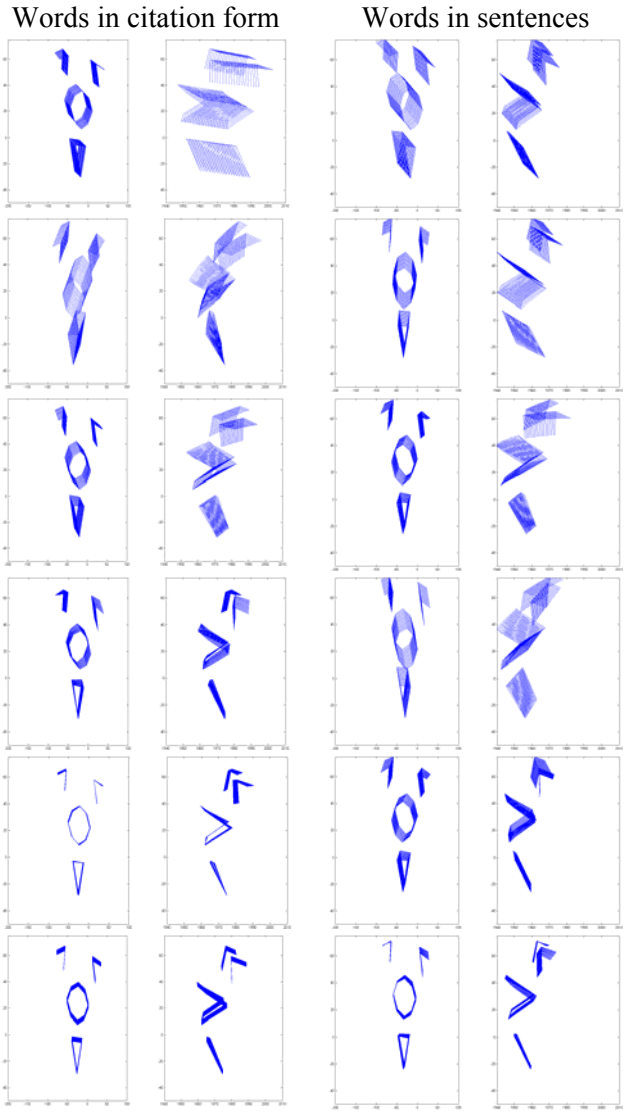

Figure 6. *Front view and profile view of the spatial variation involved in each head PC (PC1 to PC6 from top to bottom) produced in words in citation form (left) and words in sentences (right).*

3.2. Acoustic analysis

3.2.1. Word durations

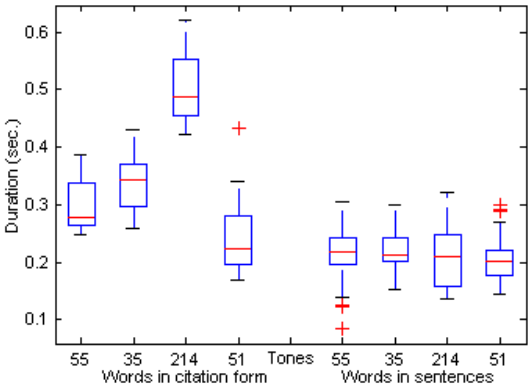

Figure 7. *Boxplots of word durations for words in citation form (left) and words in sentences (right) as a function of the tone*

Word durations were not significantly different as a function of the word produced (tested with a non-parametric Kruskal-Wallis test, $\chi^2(5)=7.41$, $p=0.19$), so the data were merged across segmental words for each of the four tones.

Figure 7 shows the different durations as a function of the tones for the words produced in citation form and in sentences. A two-way analysis of variance (ANOVA) revealed that durations were significantly higher for words produced in citation form (mean=0.3472 sec.) than words produced in sentences (mean=0.2125 sec.) ($F(1,232)=442$, $p<0.01$). Also word durations were significantly different depending on which tone was produced (mean tone55=0.2568; mean tone35=0.2788; mean tone214=0.3587; mean tone51=0.2250) ($F(3,232)=79.19$, $p<0.01$). Finally, there was a significant interaction between tones and word context of production ($F(3,232)=76.83$, $p<0.01$). While the durations seem overall homogenous for words in sentences, they were quite different for words in citation form depending on which tone was produced. In particular, words produced with tone 214 in citation form were longer than the other ones (see Figure 7).

3.2.2. Correlations between the acoustics and the visible movements

The F0 values have been estimated for all the words. So for each word, produced either in citation form or in sentences, we have the six non-rigid face PCs and the six rigid head PCs which were obtained after data modelling (as explained above) time-aligned with the F0 values and the acoustic signal. Figure 8 shows an example of those signals for the word 'fu214'. We can see that there is articulatory face and head movement involved in all the different PCs during the production of the word. Note that all the movements start before and end after the corresponding sound, compared to the F0 which is strictly bounded by the sound boundaries (and specifically by the voiced portion). In order to investigate if there are visual cues for the different tones, we calculated the correlations between the visual PCs and the F0 values as a function of the tone. All the words repeated five times were mixed together to create sets of data per tones. The correlations were calculated on these data sets.

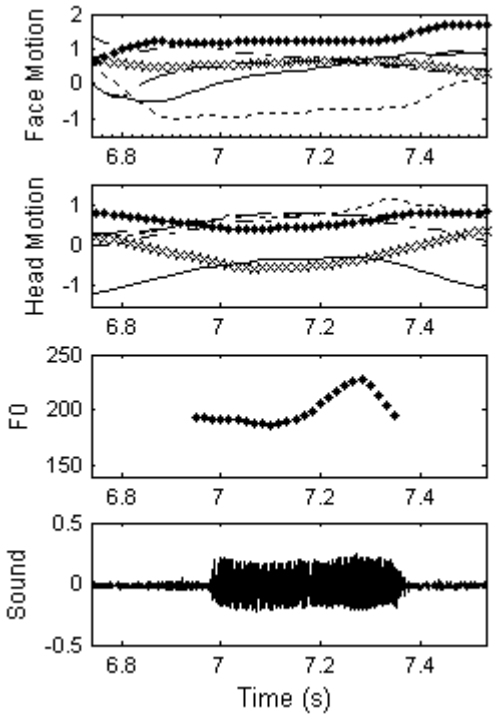

Figure 8. *Plots of the different time-aligned signals as a function of time for the word 'fu214' produced in citation form: (from top to bottom) the 6 face PCs (JawO indicated by dots, LPro: cross, LLC: solid line, LR: dashdot line, LC: dashed line, JawA: dotted line), the 6 head PCs (PC1 to PC6: same order of line conventions), the F0 and the acoustic signal.*

Table 2. *Correlations between PCs and F0 for words produced in citation form (significant correlations are indicated in bold, p<0.01)*

Face PC	JawO	LPro	LLC	LR	LC	JawA
Tone 55	**0.17**	0.12	**0.18**	**-0.28**	**0.22**	**-0.38**
Tone 35	**0.22**	**-0.19**	**0.31**	-0.09	**0.16**	-0.02
Tone 214	**0.27**	0.01	0.09	**-0.19**	**-0.15**	**-0.19**
Tone 51	**0.18**	**-0.17**	0.14	-0.06	**0.41**	**-0.45**
Head PC	PC1	PC2	PC3	PC4	PC5	PC6
Tone 55	-0.14	**0.36**	**-0.31**	0.06	0.06	0.01
Tone 35	**0.18**	0.09	**-0.13**	-0.11	0.03	0.05
Tone 214	**0.13**	-0.09	0.01	**-0.14**	-0.02	**-0.11**
Tone 51	**0.39**	-0.12	0.00	**-0.20**	**-0.32**	**0.22**

Table 3. *Correlations between PCs and F0 for words produced in sentences (significant correlations are indicated in bold, p<0.01)*

Face PC	JawO	LPro	LLC	LR	LC	JawA
Tone 55	**-0.25**	0.01	**-0.24**	**0.2**	**-0.24**	0.16
Tone 35	0.01	-0.14	**-0.18**	0.05	0.00	**0.21**
Tone 214	-0.06	-0.13	**-0.25**	-0.02	0.04	**0.24**
Tone 51	0.09	0.14	0.13	0.07	-0.09	0.06
Head PC	PC1	PC2	PC3	PC4	PC5	PC6
Tone 55	-0.02	0.05	**-0.34**	0.03	**-0.25**	0.10
Tone 35	-0.03	-0.01	0.13	0.14	-0.04	0.07
Tone 214	-0.17	0.08	0.12	-0.11	**0.21**	**0.31**
Tone 51	-0.09	0.02	0.09	0.11	**0.18**	-0.00

Table 2 and Table 3 represent the correlations per tone between each face and head PC with the F0 values for all words confounded. Globally the correlations values are not very high but this might be explained by the fact that the correlations are global correlations for all the words together. There are a certain number of significant correlations which indicates that there are visual cues related to tones acoustics. While some PCs show significant correlations with F0 values whatever the tone is (e.g. for JawO and LC in Table 2) others correlate better with one or two tones only. The fact that JawO and LC, the jaw opening and the lip closing, correlate well with all the tones produced in isolation seems to indicate that these PCs are well related to the F0 in general and so to tones but they may not be really discriminant visual cues for distinguishing among the four lexical tones. It is not surprising however that JawO shows strong correlations with the F0 whatever the tone is since the jaw opening is the main movement involved in speech. So these components should better reflect articulatory variations rather than something specific to each tone (even if they could probably still provide some general information about tones). Other movements are involved in the production of certain particular tones only (e.g. for words in citation form, the lower lips closing movement LLC has significant correlations with F0 for tones 55 and 35 only). If we consider each tone, there is a unique pattern of correlations associated to each tone and involving a combination of several different PCs. For example tone 55 could be defined by its relations with LLC, LR and JawA, suggesting that either the relative height of this tone or its lack of contour modulation are uniquely related to lip raising and jaw advance, while tone 35 is more related to LPro and LLC, suggesting that the rising contour of tone 35 may be uniquely related to lip protrusion. The fact that there are several significant correlations for non-rigid motion (Table 2) could indicate that it is difficult to separate face movements involved in speech articulation from those more related to the tones. But it appears that face articulations are somehow influenced by tones.

As concerns the head movements, there is globally less number of significant correlations compared to the face movements. Each tone seems to be related to a few components, which combine together in a unique way. For example, for words produced in citation form, tones 35, 214 and 51 F0 values are well correlated with PC1 (plus other combinations of PCs), suggesting that the back and forth movement of PC1 may accompany the change in contour modulation, while tone 55 is better characterized by relations with PC2 and PC3, suggesting that the movements involved in these PCs (a sort of head nodding for PC2) may be better related to the production of a high tone (Table 2).

Interestingly there are fewer significant correlations for words produced in sentences compared to words produced in citation form (Table 3). As concerns the face PCs, while tone 55 shows multiple correlations as for isolated words, tone 35 and tone 214 are both related to LLC and JawA in the same way. Tone 51 shows no significant correlation with any of the PCs. So contrary to words in citation form, tones in words produced in sentences are not well related to facial PCs. However the lower lip closure movement LLC seems somehow influenced by the tones as it shows significant correlations with three of the tones (tones 55, 35 and 214). As regards head PCs, as for the above analyses, there are unique combinations of PCs related to each tone for tone 55, tone 214 and tone 51. However this time these combinations do not include the first two main PCs, PC1 and PC2. Moreover tone 35 shows no significant correlation. So it appears that in some cases the face PCs or the head PCs don't allow to define uniquely the tones. However when we look at the overall pattern of correlations, we can see that face and head PCs combine together to define each tone. For example tone 51 is not linked to any of the face movements but is associated to a small up and down movement of the head (PC5). It appears that for words in sentences the tones distinction relies on a tight combination of face and head movements whereas for isolated words, the distinction may depend on multiple associations of face and head motions, probably redundant. The difference in pattern of correlations between the contexts where words were produced may be explained by the fact that in sentences, tones tend to be more similar to each other (e.g. see the durations in Figure 7) and because the words are embedded in sentences there may be a general effect of coarticulation and sentence intonation affecting both motion and acoustics (see for e.g. [17] for modifications of F0 contours in connected discourse).

In all cases, we note a large number of significant correlations between F0 values and PCs which do not explain much of the total variance of the data (e.g. PC5 and PC6), so that would mean that lexical tones are more related to subtle movements of the face and head of the speaker than to main movements. And this is particularly true for words produced in sentences since the first two components for the face and the head (JawO, LPro, PC1 and PC2) were globally not related to the tones (except for JawO and tone 55). It is likely that in sentences the more important head movements better correlate with the general intonation of the sentence [15] rather than with local lexical tones which are better associated with subtle head movements.

4. Discussion and Conclusion

In this paper we presented a study of audio-visual production for Mandarin words in citation form and in sentences that vary on lexical tones. OPTOTRAK motion capture data were modelled using PCA for rigid head movements and guided-PCA for non-rigid face movements and head movements were separated from face movements. For each tone, correlations between F0 values and the different face and head components were calculated.

It was found that the head movements involved in word production in citation form and in sentences were different, probably due to a general effect of visual speech prosody of head movements involved in speech (e.g. [15]). As regards acoustic duration, it was found that words in sentences were shorter and more homogeneous across the tones than words produced in citation form. In particular the duration does not seem to be a discriminant factor of tones for words in sentences. This may be explained by the fact that duration of words in sentences has been shown to be used to organize syllables into groups [18]. A complementary analysis on the F0 values (values for F0 peaks and valleys) of words produced in citation form and in sentences could be interesting to further characterize the different tones. Correlations between the F0 values and the different PCs were more numerous for tones of words produced in citation form compared to in sentences. It should be noted that a similar result using LDA tone prediction category for words in isolation and words in sentences was found for Cantonese [10]. For words in citation form, unique patterns of correlations in both face and head movements were found for each tone indicating that there exist visual cues specific to each tone. These unique patterns of cues are of interest but their psychological utility remains to be determined; perception studies in which particular

components are manipulated independently and in combination are required. In the case of words produced in sentences such unique combinations involved both the face and head movements, and correlations were essentially with PCs which do not explain much of the variance in the data, i.e., PCs involving subtle movements of the face and head. So there were fewer relationships between acoustic and visual features in sentences, probably implying that Mandarin lexical tones would be harder to discriminate audio-visually when produced in sentences compared to in isolation. This outcome should be considered in future studies of Mandarin lexical tones production and perception when generalizing results obtained with words produced in isolation. Note that this difference between contexts of production was already noticed for acoustic studies of tones (e.g. [19])

PCA was used in this paper to model the rigid and non-rigid articulatory speech movements and separate the contribution of each movement. This was done to allow further comparisons between speakers and languages. Even though use of PCA allows a more comprehensive generalization (in particular guided-PCA as it takes into account a set of crucial sensors), it should be noted that the use of raw sensor positions for the calculation of the correlations with the F0 could have led maybe to stronger values. In particular for the lips parameters, specific Optotrak sensors have been used already to examine visible correlates of prosodic focus in French [20].

The correlations presented in this paper were global correlations between F0 values and each PC across segmental word identity. It would be interesting to consider individual words for calculating the correlations to investigate if this general pattern is maintained at the individual word level. Furthermore it would be interesting to separate the visual cues (especially the head related ones) associated with the prosody of the sentence from those specifically associated with tones. In addition we could probably get a clearer pattern as regards head movements in isolated words and words in sentences if we could compare the same components; this could be achieved by comparing directly the pitch, yaw and roll movements of the head for example. Finally, this study would benefit from complementary analyses of the dynamics of the articulatory gestures and head motions as a function of each tone. Considering other F0 parameters such as peak and valley values or F0 slopes could be also of interest for characterizing the visual correlates of lexical tones. Of course a study involving several speakers and additional sensors (e.g. in the eyebrows as their movements have been shown to be related to F0 variations [21], and on the neck around the larynx, as larynx height has been shown to be correlated with F0 values, e.g. [22]) will allow us to generalise the results further, and such studies are planned.

5. Acknowledgements

The authors would like to thank ATR Laboratories, Japan for providing access to OPTOTRAK equipment. The authors would like to thank Dr. Nan Xu for assisting with some of the phonetic labelling and providing expertise in Mandarin, and Mr. Leo Chong and Mr. Johnson Chen for checking the accuracy of the Mandarin productions. This study is funded by two Australian Research Council (ARC) Discovery project grants No. DP0988201 and No. A79917254 to the 4th author and an ARC Discovery project grant No. DP0211947 to the 3rd and 4th authors.

6. References

[1] Sumby, W.H., and Pollack, I., "Visual Contribution to Speech Intelligibility in Noise", JASA, 26(2):212-215, 1954.

[2] Bernstein, L.E., Demorest, M.E., and Tucker, P.E., "What makes a good speechreader? First you have to find one.", in R. Campbell, B. Dodd, and D. Burnham (Eds.): Hearing by eye (II): The psychology of speechreading and auditory-visual speech, 211-228, Psychology Press, 1998.

[3] Benoit, C., Mohamadi, T., and Kandel, S., "Effects of phonetic context on audiovisual intelligibility of French ", J. Speech Hear. Res., 37(5):1195-1203, 1994.

[4] Yip, M.J.W., "Tone", Cambridge University Press, 2002.

[5] Chao, Y.-R., "A system of tone-letters", Le Maitre Phonetique, 45:24-27, 1930.

[6] Blicher, D.L., Diehl, R.L., and Cohen, L.B., "Effects of Syllable Duration on the Perception of the Mandarin Tone-2 Tone-3 Distinction - Evidence of Auditory Enhancement", Journal of Phonetics, 18(1):37-49, 1990.

[7] Burnham, D., Ciocca, V., and Stokes, S., "Auditory-visual perception of lexical tone". Proc. Eurospeech 2001, Aalborg, Denmark, 395-398,2001.

[8] Mixdorff, H., Hu, Y., and Burnham, D., "Visual Cues in Mandarin Tone Perception". Proc. Eurospeech 2005, Lisbon, Portugal, 405 - 408,2005.

[9] Chen, T.H., and Massaro, D.W., "Mandarin speech perception by ear and eye follows a universal principle", Percept. Psychophys., 66(5):820-836, 2004.

[10] Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., and Haszard Morris, R., "The perception and production of phones and tones: The role of rigid and non-rigid face and head motion". Proc. ISSP 2006, 7th International Seminar on Speech Production, 185-192, 2006.

[11] Revéret, L., Bailly, G., and Badin, P., "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation". Proc. International Conference on Speech and Language Processing, Beijing, China, 755-758,2000.

[12] Bailly, G., Bérar, M., Elisei, F., and Odisio, M., "Audiovisual speech synthesis", International Journal of Speech Technology, 6(4):331-346, 2003.

[13] Boersma, P., "Praat, a system for doing phonetics by computer.", Glot International, 5(9/10):341-345, 2001.

[14] Tan, L., and Karnjanadecha, M., "Pitch Detection Algorithm: Autocorrelation Method and AMDF". Proc. the 3rd International Symposium on Communications and Information Technology, Songkhla, Thailand, 541-546,Sept., 3-5 2003.

[15] Yehia, H.C., Kuratate, T., and Vatikiotis-Bateson, E., "Linking facial animation, head motion and speech acoustics", Journal of Phonetics, 30(3):555-568, 2002.

[16] Ishi, C.T., Ishiguro, H., and Hagita, N., "Analysis of inter and intra-speaker variability of head motions during spoken dialogue". Proc. AVSP, 37-42,2008.

[17] Xu, Y., and Keith, B., "Tone in Connected Discourse": Encyclopedia of Language & Linguistics, 742-751, Elsevier, 2006.

[18] Xu, Y., and Wang, M., "Organizing syllables into groups--Evidence from F0 and duration patterns in Mandarin", Journal of Phonetics, 37(4):502-520, 2009.

[19] Shen, X.S., and Lin, M., "Concept of tone in Mandarin revisited: A perceptual study on tonal coarticulation", Language Sciences, 13(3-4):421-432, 1991.

[20] Dohen, M., Loevenbruck, H., and Hill, H., "Recognizing prosody from the lips", in A.W.-C.L.S. Wang (Ed.): Visual Speech Recognition: Lip Segmentation and Mapping 416-438, IGI Global, 2009.

[21] Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., and Espressor, R., "About the relationship between eyebrow movements and F0 variations". Proc. Fourth International Conference on Spoken Language Processing, 2175-2178, 1996.

[22] Honda, K., Hirai, H., Masaki, S., and Shimada, Y., "Role of Vertical Larynx Movement and Cervical Lordosis in F0 Control", Language and Speech:401-411, 1999.