Prosody for the Eyes: Quantifying Visual Prosody using Guided Principal Component Analysis

Erin Cvejic, Jeesun Kim, Chris Davis, Guillaume Gibert

MARCS Auditory Laboratories, University of Western Sydney, Milperra, Australia

e.cvejic@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au, g.gibert@uws.edu.au

Abstract

Although typically studied as an auditory phenomenon, prosody can also be conveyed by the visual speech signal, through increased movements of articulators during speech production, or through eyebrow and rigid head movements. This paper aimed to quantify such visual correlates of prosody. Specifically, the study was concerned with measuring the visual correlates of prosodic focus and prosodic phrasing. In the experiment, four participants' speech and face movements were recorded while they completed a dialog exchange task with an interlocutor. Acoustic analysis showed that prosodic contrasts differed on duration, pitch and intensity parameters, which is consistent with previous findings in the literature. The visual data was processed using guided principal component analysis. The results showed that compared to the broad focused statement condition, speakers produced greater movement on both articulatory and non-articulatory parameters for prosodically focused and intonated words.

Index Terms: prosody, visual speech, speech production, guided principal components analysis, inter-speaker variation.

1. Introduction

Visual cues available from the face of a speaker can signal information not only about what has been said (phonemic content), but also how it has been said (i.e., speech prosody). While visual cues to speech content are closely linked to articulatory movements in oral regions [1], visual cues to prosody have been shown to be distributed across wider face areas (including mouth, eyebrow and head movements) [2-8].

However, although the acoustic correlates of prosody are fairly well understood, the same cannot be said for visual cues. This could be because such visual cues are less directly coupled to speech production and therefore show less consistent patterns across tokens or speakers [9-11]. However, it should be noted that most studies of visual prosodic cues have been limited in a number of ways. For example, the size of speech corpus has typically been small [11] making generalization of results problematic. Also, visual prosody was typically analysed based on a single produced token and examination of only an initially stressed syllable (rather than examining properties across an entire word or sentence), again raising a question concerning the generalization of results. Moreover, local rather than the whole face and head movements are often measured, missing the potential relationship between movements across face areas.

Given the above, the current study examined visual prosodic cues by measuring speakers' overall face and head movements for 30 different sentences across a range of prosodic contrasts, elicited in an interactive dialog task.

2. Method

2.1. Participants

Four male native speakers of Standard Australian English (M_{Age}= 23 years) participated in the data capture sessions. All reported having no known speech or hearing deficits.

2.2. Materials

The materials consisted of 30 non-expressive sentences drawn from the IEEE Harvard Sentence list [12] describing mundane events with minimal emotive content. Each sentence was recorded in one of three prosodic conditions: as a *broad focused* statement, a *narrow focused* statement, and as an *echoic question*.

Acoustically, narrowly focused sentences are characterised as having longer syllable durations, greater intensity and higher fundamental frequency (F_0) than the same words produced in a broad focused context [13]. Broad focused statements can be characterised as having a steadily falling F_0 contour and ending with a sharp, definitive fall signaling finality, whereas the opposite pattern is observed for echoic questions. The former also tend to have shorter final syllable durations, and steeper final intensity falls relative to the same sentences uttered as questions [14].

To elicit these conditions in the study, a dialog exchange task was used [2, 3] requiring the speaker to interact with an interlocutor, and either repeat what they heard the interlocutor say (broad focused statement), make a correction to an error made by the interlocutor (narrow focused statement) or question an emphasized item within the sentence produced by the interlocutor (echoic question).

2.3. Apparatus

2.3.1. Motion Capture

A Northern Digital Optotrak 3020 machine was used to record the visual speech movements from 38 markers positioned on the head and face of the speaker (see Figure 1). These positions were chosen to reflect non-rigid movements of the jaw, lips, cheeks and brows, as well as rigid rotations and translations from the centre of rotation. The three-dimensional marker positions were captured at 60Hz.

2.3.2. Sound and Video Capture

In addition to the motion capture, auditory data was synchronously captured using a Behringer C-2 condenser microphone connected to an Optotrak Data Acquisition Unit II (Northern Digital Inc.) through a Eurorack MX602A mixer, sampled at 44.1 kHz, digitized mono. Video was also recorded using a Sony TRV19E digital video recorder.

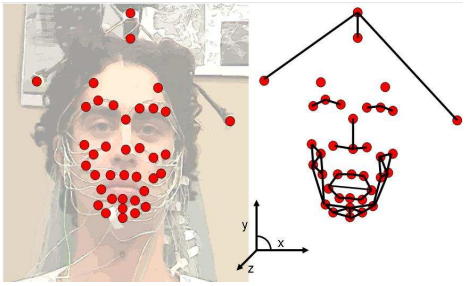

Figure 1. *Location of optical markers (with size exaggerated for clarity) on the face of the speaker are shown on the left.34 markers are placed on the face, with 4 markers positioned on a head band to measure rigid movements around the centre of rotation. The right image depicts the marker positions with "bones" added. Also shown are the directions of the X, Y and Z axes.*

2.4. Motion Capture Procedure

Each session began with the placement of the movement sensors on the face of the speaker in the configuration shown in Figure 1. Each speaker was recorded individually while seated in an adjustable dentist's chair within a double-walled, sound insulated booth (see Figure 2). Participants were instructed to direct their speech towards the interlocutor, who was located approximately 2.5 meters in front of them while engaging in the dialogue exchange task outlined in Section 2.2. Two repetitions of each sentence were recorded in the three prosody conditions. The total motion capture sessions lasted approximately 120 minutes (including occasional breaks). In total, 180 sentences (30 sentences x 3 prosodic contrasts x 2 tokens) were recorded for each speaker.

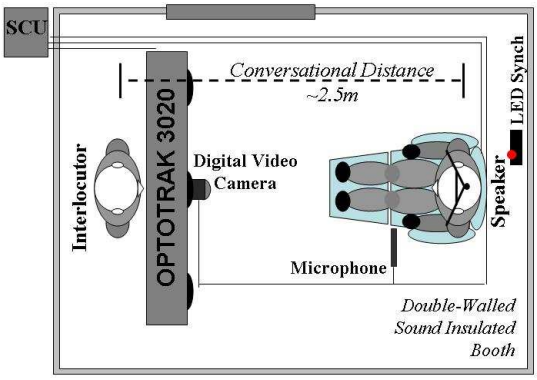

Figure 2. *The experimental setup used in the motion capture sessions. Recording takes place in a double-walled, sound insulated booth, with the system control unit located outside to minimise extraneous noise.*

2.5. Data Processing

Motion capture data was processed for each speaker using so-called guided principal component analysis (gPCA) [15-20] to reduce the dimensionality of the data sets. "Standard" principal component analysis (PCA) delivers optimal orthogonal factors explaining the maximum data variance within a minimal number of components. In contrast, guided PCA consists of linear decomposition to generate a set of components that are interpretable in terms of articulatory control parameters (e.g., jaw opening and jaw protrusion are extracted as separate components), at the cost of sub-optimal variance explanation and minor correlations between derived components. Typically, six components can explain most articulatory data [19], with several additional components used to describe eyebrow and expressive movements [20].

In data processing, it is important to minimise the over-representation of particular marker configurations (e.g., the static pose at the beginning and end of an utterance). For this, a training database of unique movements was first generated, from which the center of rotation was estimated and used to separate out rigid rotations and translations (Figure 3).

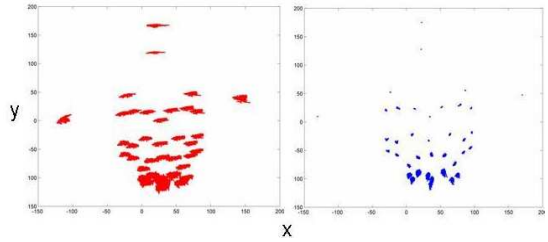

Figure 3. *Marker displacements on the X and Y axes within the training database with (left) all movements and (right) after removal of rigid motions.*

The non-rigid movements within the training database were then subjected to gPCA, while the separated rigid movements undergo standard PCA. Table 1 outlines the guided parameters for extraction using gPCA. The original recordings were then re-projected into component space as a function of time, reducing the dimensionality of the data from 114 data points to 14 interpretable principal components (PCs) per captured frame. All processing was conducted in Matlab (The MathWorks) using in-built and custom functions.

Table 1. *A priori parameters used to guide the gPCA and assigned labels for the extracted rigid parameters using standard PCA.*

Principal Component	Movement Parameter	Axes of Movement
Non-Rigid Parameters (from gPCA)		
1	Jaw Opening	Y
2	Mouth Opening	Y
3	Lower Lip Mvmt.	Y
4	Upper Lip Mvmt.	Y
5	Lip Spreading	X Y Z
6	Jaw Protrusion	Z
7	Brow Raising	Y
8	Brow Pinching	X Y
Assigned Rigid Parameters (from PCA)		
9	Pitch Rot.	X
10	Roll Rot.	Z
11	Yaw Rot.	Y
12	Fwd/Bwd Trans.	Z
13	Left/Right Trans.	X
14	Up/Down Trans.	Y

Audio was manually transcribed in Praat [21] and used to temporally locate the critical word (i.e., the word that received narrow focus or question intonation) within each sentence movement data. Note that although the impact of having prosodic focus and phrasings typically extends beyond the boundaries of a single word within an utterance, to simplify the analysis, only the data associated with the critical word is examined and reported.

In order to compare the visual movements for the critical sections across repetitions, speakers, sentences and prosodic conditions the visual parameters for the critical sections of each utterance were time normalised using linear spline interpolation in Matlab, and projected onto a new time series

(see Figure 4). As can be seen, the normalization changes the overall length but not the characteristic "shape" of the components in time, so that comparisons made are based on the differences in shape.

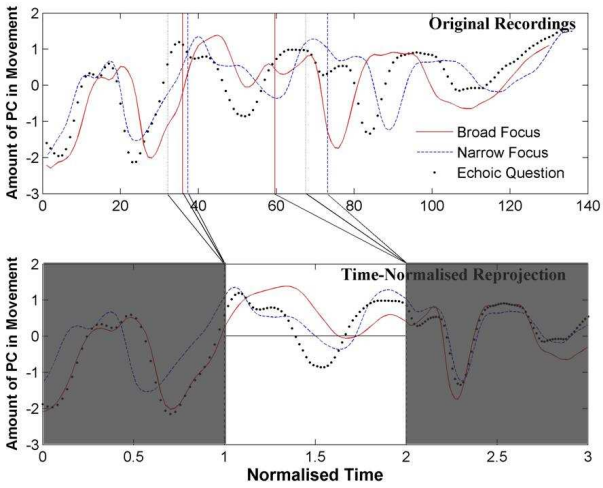

Figure 4. *The original recordings were reprojected onto time normalised space using linear spline interpolation. Only the critical section of each sentence is analysed.*

3. Results and Discussion

3.1. Auditory Analysis

The acoustic properties of the critical item within each utterance was analysed using Praat [21]. The values for duration, mean F_0, F_0 range, mean relative intensity and intensity range of the critical items were compared for each speaker (see Table 2) using analysis of variance (ANOVA), with speech condition as the independent variable [2]. The main effect of prosodic speech condition was significant, F (2, 119) = 2209.07, $p < .0001$.

Post-hoc comparisons showed that words within narrow focus statement and echoic question contexts were produced with increased durations, with greater mean F_0 and employed a greater intensity range, relative to the same words produced within a broad focused context. The results of this acoustic analysis are consistent with previous findings [13,14] confirming that the acoustic properties vary as a function of the prosodic speech conditions.

Table 2. *The mean values (listed by speaker) for each property in the broad focus condition are given for reference. The values for the narrow focus and echoic question renditions are reported as proportions of the average value for the broad focused items (analysed per speaker and sentence).*

	Dur. (ms)	Mean F_0 (Hz)	F_0 Range (Hz)	Mean Relative Int. (dB)	Int. Range (dB)
Mean Values for Broad Focused Renditions					
Spk. 1	365.14	126.91	29.62	74.37	15.69
Spk. 2	356.39	120.13	32.41	58.32	18.17
Spk. 3	303.06	96.14	14.08	49.47	11.11
Spk. 4	359.27	100.22	14.15	51.16	15.57
Values expressed are proportions of Broad Focused Values					
Narrow Focus	1.51	1.21	2.01	1.01	1.94
Echoic Question	1.48	1.17	3.51	1.00	1.75

3.2. Visual Analysis

Figure 5 shows the accounted variance of non-rigid and rigid movements by PCs for each speaker. With only eight non-rigid PCs, in excess of 91% of the variance of face movement for each speaker was recovered. It should be noted that although the sentences and prosodic conditions were identical across speakers and repetitions, there appears to be differences across speakers in the amount of variance accounted for by each PC.

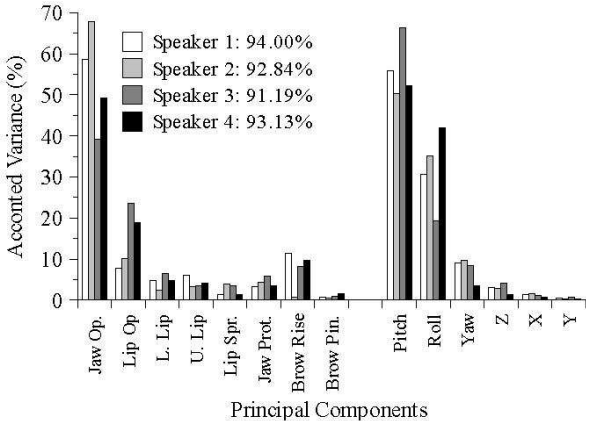

Figure 5. *Variance accounted for by each component, separated by speaker. The eigenvalues from the gPCA are shown in the left half of the figure, with the rigid components from PCA displayed in right half.*

In order to parameterise the differences in visual movement (information) in the critical section between prosodic conditions, the absolute value of the area under the normalized curve for each PC was calculated using trapezoidal estimation, and was used as a measure of PC strength over time. These values were then used in a series of paired samples t-tests (with a Bonferroni adjusted $\alpha = .001$) to determine the PCs that differed as function of prosodic condition. In the analysis that follows, we focus on the first 7 non-rigid parameters (see Table 1) and the rigid parameters corresponding to rotations around the X, Y and Z axes.

Figure 6 shows the mean differences in absolute area under the curve for each of the measured PCs between the *narrow focus/ echoic question* and broad focused conditions. In terms of these selected PCs, critical words can be characterised as being produced with significantly greater jaw and lip openings, as well as jaw protrusions in narrow focus and echoic question conditions in comparison to the same words being produced as a broad focused rendition. Following on from the acoustic analyses in Section 3.1, these motion differences are expected (being a consequence of the articulation required to shape the vocal tract to produce the observed acoustic differences).

Of particular interest here however, are the non-articulatory movement differences observed across the prosodic conditions for brow movement and the pitch and roll head rotations. Broad focused renditions were often devoid of substantial brow motion or head rotation. In comparison, there was a marked increase in such movements for words produced in narrow focus and echoic question contexts. Further analyses of the auditory and visual data together (which are currently in progress) will reveal the temporal relationship between the auditory and visual cues as well as indicating their potential function. For example, theses additional visual markings of focus and intonation may occur ahead of the auditory signal thus cuing perceivers to better use the auditory prosodic cues and enhance understanding of the spoken message [5].

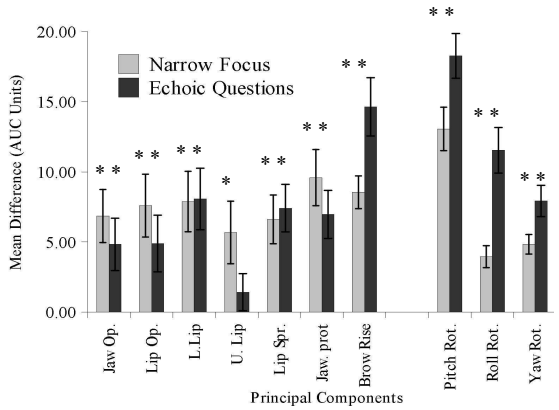

Figure 6. *Mean change in the area under the curve for non-rigid PCs 1-7 and rigid rotations in the narrow focus and echoic question conditions, relative to the broad focus condition. * indicates p<.001 (df=119).*

Furthermore, it should be noted that the differences reported pertain to results collapsed across speakers. Upon further examination of these items by speaker, it was apparent that not all speakers displayed the same degree of visual cues in accompany with acoustic prosody. For example, jaw opening for echoic questions was enhanced for only three of the four speakers with the fourth speaker showing a much greater increase in rigid motion, particularly pitch rotations, for narrow focus and echoic questions relative to broad focused renditions. Similarly, only two of four speakers produced increased lip spreading for narrow focus and echoic questions. As such, these results indicate that although a general increase in measured parameters may be observed, the systematic use of such cues to enhance the acoustic signal varies greatly across speakers [9-11].

4. General Discussion

An in depth study was conducted on visual prosodic cues by examining how the overall face and head movements of four talkers changed as a function of prosodic speech conditions. To reduce problems in generalization, the study induced prosodic contrasts in a sentence context by use of an interactive task and included a relatively large speech corpus. The results showed (1) the presence of robust visual prosodic cues conveyed by not only movements in mouth regions (closely related to articulation) but also brow and head movements (not directly related to articulation); (2) a greater increase in rigid rotations movements for echoic questions than narrow focus renditions relative to broad focused productions; (3) individual variation in regards to which movements were prominent.

The determination of the temporal relationship between acoustic and visual properties of prosody is currently under way, the results of which may illuminate the communicative function of visual prosody. Although previous research has shown correlations between pitch and brow movements, these events may not be strictly "time-locked", i.e., visible gestures may serve as pre-articulatory cues to an upcoming acoustic event. Also, to better understand inter-speaker differences, more speakers are currently being recorded.

5. Acknowledgements

The authors thank Catherine Gasparini and the four speakers for their time and patience during the recording procedure. The authors acknowledge support from the Australian Research Council (DP0666857 & TS0669874).

6. References

[1] Summerfield, Q., "Lip-reading and audiovisual speech perception", Phil. Trans.: Bio. Sciences, 335: 71-78, 1992.

[2] Cvejic, E., Kim, J. and Davis, C., "Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion", Speech Commun., 52: 555-564, 2010.

[3] Cvejic, E., Kim., J. and Davis, C., "It's all the same to me: Discriminating prosody across face areas and speakers", Speech Prosody 2010, 100893: 1-4, 2010.

[4] Swerts, M. and Krahmer, E., "Facial expressions and prosodic prominence: Effects of modality and facial area", J. Phonetics, 36: 219-238, 2008.

[5] Swerts, M. and Krahmer, E., "Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions", J. Phonetics, 38:197-206, 2010.

[6] Lansing, C.R. and McConkie, G.W., "Attention to facial regions in segmental prosodic visual speech perception tasks", J. Speech, Lang. & Hearing Res., 42: 526-539, 1999.

[7] Yehia, H.C., Kuratate, T. and Vatikiotis-Bateson, E., "Linking facial animation, head motion and speech acoustics", J. Phonetics, 30: 555-568, 2002.

[8] Cavé, C., Gua, I., Bertrand, R., Santi, S., Harley, F. and Esesser, R., "About the relationship between eyebrow movements and F0 variations", Int. Conf. on Speech and Lang. Proc., 2175-2178, 1996.

[9] Dohen, M., Lœvenbruck, H. and Hill, H., "Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features?", in A.W.-C. Liew and S. Wang [Eds], Visual Speech Recognition: Lip Segmentation and Mapping, 416-438, IGI Global, 2009.

[10] Dohen, M. and Lœvenbruck, H., "Interaction of audition and vision for the perception of prosodic contrastive focus", Lang. & Speech, 52: 177-206, 2009.

[11] Scarborough, R., Keating, P., Mattys, S. L., Cho, T. and Alwan, A., "Optical phonetics and visual perception of lexical and phrasal stress in English", Lang. & Speech, 52: 135-175, 2009.

[12] IEEE Subcommittee on Subjective Measurements, "IEEE recommended practices for speech quality measurements", IEEE Trans. Audio Electroacoust., 17, 227-246, 1969.

[13] Krahmer, E. and Swerts, M., "On the alleged existence of contrastive accents", Speech Commun., 34: 391-405, 2001.

[14] Eady, S.J. and Cooper, W.E., "Speech intonation and focus in matched statements and questions", J. of the Acoust. Soc. of America, 80: 402-415, 1986.

[15] Badin, P., Bailly, G., Revéret, L, Baciu, M., Segebarth, C. and Savariaux, C., "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images", J. Phon., 30: 533-553, 2002.

[16] Badin, P., Borel, P., Bailly, G., Revéret, L, Baciu, M. and Segebarth, C., "Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images", Proc. 5th Seminar on Speech Production: Models and Data, 2000.

[17] Maeda, S., "Face models based on a guided PCA of motion-capture data: Speaker dependant variability in /s/-/z/ contrast production", ZAS Papers in Linguistics, 40: 95-108, 2005.

[18] Beautemps, D., Badin, P. and Bailly, G., "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling", J. Acoust. Soc. America, 109(5): 2165-2180, 2001.

[19] Bailly, G., Govokhina, O., Elisei, F. and Breton, G., "Lip-synching using speaker-specific articulation, shape and appearance models", EURASIP J. on Audio, Speech and Music Processing, 2009: 1-11, 2009.

[20] Bailly, G., Elisei, F., Badin, P. and Savariaux, C., "Degrees of freedom of facial movements in face-to-face conversational speech", Int. Workshop on Multimodal Corpora, 2006.

[21] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer", Version 5.1.05 [Computer Program], Retrieved from <http://www.praat.org>, 2009.