INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The consistency and stability of acoustic and visual cues for different prosodic attitudes

Jeesun Kim [1]*and Chris Davis*[1]

[1] The MARCS Institute, Western Sydney University, Australia

j.kim@westernsydney.edu.au, chris.davis@westernsydney.edu.au

Abstract

Recently it has been argued that speakers use conventionalized forms to express different prosodic attitudes [1]. We examined this by looking at across speaker consistency in the expression of auditory and visual (head and face motion) prosodic attitudes produced on multiple different occasions. Specifically, we examined acoustic and motion profiles of a female and a male speaker expressing six different prosodic attitudes for four within-session repetitions across four different sessions. We used the same acoustic features as [1] and visual prosody was assessed by examining patterns of speaker's mouth, eyebrow and head movements. There was considerable variation in how prosody was realized across speakers, with the productions of one speaker more discriminable than the other. Within-session variation for both the acoustic and movement data was smaller than across-session variation, suggesting that short-term memory plays a role in consistency. The expression of some attitudes was less variable than others and better discrimination was found with the acoustic compared to the visual data, although certain visual features (e.g., eyebrow brow motion) provided better discrimination than others.

Index Terms: expressive speech, audiovisual prosody, prosodic attitudes

1. Introduction

Speakers convey much more than the information associated with their words. This is because speech also conveys expressive information, e.g., about emotions and attitudes. This expressive aspect of speech can be transmitted by speech sounds, and when speakers can see each other, also visually (e.g., face and head motion). Understanding how speech conveys emotion and attitude has practical implications for human-machine communication because such can help in deciphering a speaker's message. That is, taking account of expressive speech can help an automatic system determine such things as what the speaker wants to make prominent, or whether the speaker is serious or not. Furthermore, effective auditory-visual speech synthesis requires knowledge of how to best express emotion and attitude.

Research on expressive speech generally tends to consider linguistic prosody "the organizational structure of speech" [2], separately from paralinguistic prosody that concerns emotions and attitudes. Ohala [3] (1996) proposed that the latter two categories should also be considered separately. Emotional prosody, he argued, is grounded in adaptive processes, where either the transmission of a signal has survival value or where a signal 'leaks' from a beneficial physiological state; whereas attitudes "do not confer obvious survival benefit to the signaler and are probably acquired, i.e., learned".

Ohala considered this proposed difference in the aetiology of emotion and attitudes important for how well-established and characteristic the signaling of such will be. Emotional expressions, he suggested, are likely to be found cross-culturally, whereas the expression of attitudes "are likely to vary considerably from culture to culture and perhaps even from one individual to another". Moreover, he proposed that in order for attitudes to be appropriately communicated, they would need to be contextualized.

Somewhat at odds with this proposal are the suggestions of a recent study by Hellbernd and Sammler [1] where it was argued that prosodic forms associated with attitude are highly conventionalized and can be appropriately realized even without context (i.e., as single word or even nonsense word utterances). It should be noted that [1] distinguished prosodic cues for conveying intentions from those related to a speaker's attitude. Here, we use the term prosodic attitude to include the prosodic cues for attitudes as well as those involved in conveying intentions.

The argument that [1] made was in part based on the finding that the acoustic patterns of a single word spoken 8 times in a single session with six prosodic attitudes can be readily distinguished using Linear Discriminant Analysis (LDA). It was suggested that the high accuracy of the classification implies that there was a reasonable consistency of the realization of the prosodic cues across speakers (although this was not explicitly tested).

The current experiment followed up this aspect of [1] by specifically investigating the consistency and stability of speaker prosodic realizations. To do this we used the same stimulus and prosody induction procedures and compared the prosody profiles of a female and male speaker over four within-session repetitions across four different occasions. Unlike [1], we also examined visual (head and face motion) spoken prosody.

2. Method

2.1. Participants

A female and male native speaker of Australian English (both 23 years old) took part as speakers. Similar to [1], our participants were non-actors as these renditions are more representative of typical language use compared to those of actors. That is, actors' prosodic patterns may diverge from those used in standard conversation. Both participants were familiar with making auditory and video speech recordings.

http://dx.doi.org/10.21437/Interspeech.2016-1505

2.2. Equipment

2.2.1. Image/motion capture

3D data was captured and constructed using a Carmine 1.09 close range sensor (0.35m - 1.4m). The spatial x/y Resolution was 640 x 480 (VGA) (2-Sigma Values) at 0.5m = 0.9 mm; the depth Resolution (2-Sigma Values) at 0.5m = 0.1 cm. The depth Image Field-of-View was Horizontal at 0.5 m = 53.6 degrees and Vertical at 0.5 m = 45 degrees. In addition, color Image sequences were captured at 640 x 480 (VGA).

2.2.2. Image/motion registration and processing

Faceshift Studio® 2014 facial motion software was used to register and process the 3D sensor data (see procedure below). Auditory speech was recorded using this software from an AKG C417 PP professional lavalier microphone input to a Roland Duo capture EX soundcard. The recording sessions took place in a test room lit with two Bowens UNI-LITE BW3370 flood fill lights (with semi opaque diffusers).

2.3. Materials

The two speakers to express the spoken word "beer" with six different communicative intentions or attitudes: criticism, doubt, naming, suggestion, warning, and wish.

2.4. Procedure

2.4.1. Recording

Each speaker was recorded individually. Speakers were seated in a quiet room with the Carmine 1.09 close range sensor positioned directly in front at face level and at approximately 0.6 m distance (see Figure 1). Prior to the test session, a custom specific visual expression model for each individual was constructed. This model consists of 51 blend-shapes that are captured as an individual produces different training postures (by posing 23 face postures).

In the test session proper, the Faceshift acquisition was controlled by an operator in a separate control room who ensured that the participants were looking directly sensor throughout the capture performance. The different prosodic attitudes were elicited using the same procedure as [1]. That is, to elicit the prosodic attitudes the speaker was presented with and required to read short scenarios that described a situation in which she/he interacted with an interlocutor (see [1] for details). For each new prosodic attitude, the speaker uttered an initial sentence of the relevant scenario and was encouraged to freely vocalize until she/he felt ready to begin saying the test word. In each session this word was said four times in each prosodic attitude. In addition, two 'wag' trials, where the participant moved her/his head from side-to-side and up-and-down, were performed to establish the centre of head rotation (used in the analysis of the visual prosody). Each speaker participated in four sessions.

2.4.2. Data Processing

The quantification of speech related articulatory movements: Faceshift uses an input device (here a Carmine 1.09 close range sensor) to construct a depth map by analysing a speckle pattern of infrared laser light. Virtual marker positions can be tagged to this depthmap and used to parameterize motion. Here we exported the FaceRobot® virtual marker set in c3d format and selected a subset of markers to use in a data reduction process (see Figure 1).

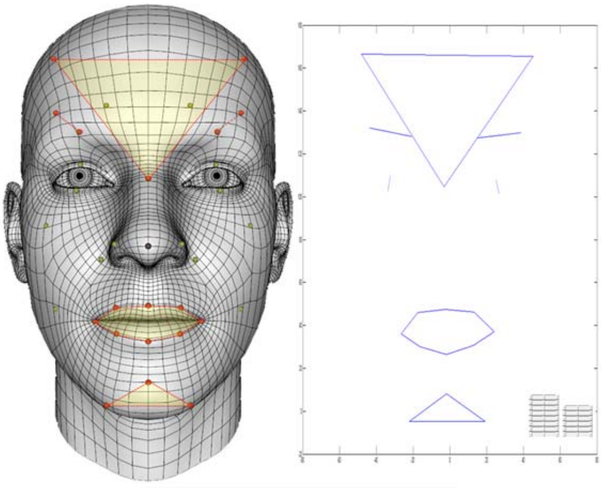

Figure 1: *A depiction of the virtual marker positions exported in c3d format from Faceshift and the subset of markers used for the gPCA (shown in red in the left panel). The right panel shows the tool used to confirm the gPC 3D reconstruction (moving the sliders, bottom right, showed the influence of each PC).*

The data from the three virtual markers on the chin were used to quantify Jaw Opening (constrained to the Z axis); the data from the 8 mouth virtual markers were used quantify Mouth opening (constrained to the Z axis) and Lip rounding (all three axes). The data from the four eyebrow markers were used to define Eyebrow motion (Z axis) and the two markers on the sides of the head and one on nose-bridge were used to define Rigid motion (pitch, yaw and roll rotation and translation).

Given the high dimensionality of the recorded data, dimensionality reduction was performed. Guided principal component analysis (gPCA, [4]) was used for the non-rigid data. This style of PCA employs linear decomposition to extract a set of a priori defined components representing biomechanically plausible articulatory control parameters (six components are typically sufficient to explain the majority of articulatory data [5]).

The shape-normalised (first frame subtracted) motion data was processed using gPCA to reduce the dimensionality of the data set to eight non-rigid components, along with three rigid translations and three rigid rotations (pitch, roll and yaw) of the whole head. To minimise the overrepresentation of particular marker configurations (e.g., the neutral position at the start and end of each utterance), a database of unique movements was generated. Using the 'wag' trials, the six rigid motion parameters around the estimated centre of rotation were determined (using the quaternion method) and extracted from the database. The remaining non-rigid movements were then analysed applying gPCA. The gPCA solutions were inspected in 3D space using a tool in which the influence of each PCA was visualized. Following inspection, the gPCA parameters were output as vectors that quantified the contribution of a gPCA per frame (time).

Quantification of acoustic features: The same features as used in [1] were used, i.e., a measure of stimulus duration, mean intensity, harmonics-to-noise ratio (HNR), mean fundamental frequency (f0), pitch rise and the standard deviation of the spectrum. These data were obtained by using Praat using customized scripts.

Quantification of motion features: From the PCA curves for mouth opening; lip rounding; eyebrow up/down and pitch and yaw rigid head rotation, the following parameters were derived: duration of motion; magnitude of the largest peak; magnitude of the largest trough; time when the peak occurred; time when the trough occurred, the difference between the magnitude of the peak and trough, the highest velocity motion and the coefficient of variation (SD/mean).

2.4.3. *Discriminant analysis*

A linear discriminant analysis (LDA) was employed (as in [1]) using the seven acoustic features as independent variables and the prosodic categories as the dependent variable (class labels). Analyses were cross-validated using a jack-knife procedure. We also used Recursive Partitioning (the RPART package in R that incorporates cross validation) to determine how the data could best be partitioned based on the values of different classes. This analysis was then used for feature selection for an LDA with reduced features.

The same basic analyses were applied to the motion data. Only in this case, an LDA was calculated for each of the five PCA curves (mouth opening; lip rounding; eyebrow up/down; and pitch and yaw rigid head rotation). Following this, the data from all curves was combined and Recursive Partitioning used to select the features that capture most of the variation in the data. These features were used in a reduced feature LDA.

2.5. Results

Acoustic data. The LDA on all the acoustic features classified the correct attitude category for speaker one at 81% correct and for speaker two at 78%. The LDA solutions for the first and second discriminant functions for speaker one and two are shown in Figure 2.

For speaker one, recursive partitioning revealed that pitch rise; mean f0; centre of gravity and HNR best partitioned the data into the attitude categories (when only these data were used discrimination of the LDA was the same). For speaker two, pitch rise; mean f0; HNR and duration best partitioned the data into the attitude categories.

Motion data. An LDA was conducted on the motion data for each speaker and each face/head movement (i.e., mouth, lips, eyebrow, head pitch and head yaw). These single motion feature LDAs produced correct classification performance from 52% to 62% (with rigid head pitch rotation and eyebrow motion better than lip rounding).

Combining all motion features and using recursive partitioning to select features produced better classification performance. For speaker one, the features that measure the magnitude of the largest peak for mouth, lip, eyebrow, rigid head pitch and rigid head yaw, and the duration of motion produced the best classification performance (64% correct).

For speaker two, the difference between the magnitude of the peak and trough for mouth motion, and the same for rigid head pitch, the magnitude of the largest peak for rigid head pitch and the time at which the peak in rigid head yaw produced the best classification performance (78% correct).

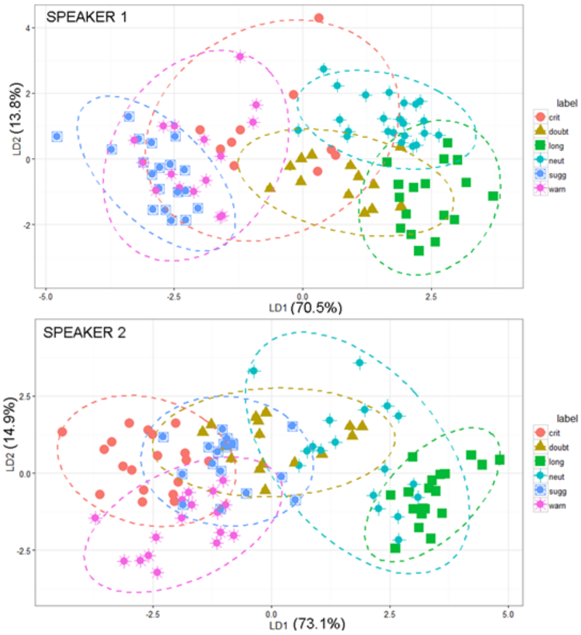

Figure 2: *Acoustic LDA results for both speakers using a reduced set of features as discovered by recursive partitioning.*

Combination of the best acoustic and motion features produced LDA performance similar to the acoustic features alone. For speaker one, the combination of pitch rise; mean F0; HNR; largest peak for rigid head pitch and auditory duration resulted in correct classification performance of 81%. For speaker 2, the recursive partitioning selected only auditory features and these resulted in a classification performance of 78% correct.

In addition to LDA, we also examined the consistency of performance within and across recording sessions for the acoustic and motion data. For the acoustic data, we examined the variation in features' scores within a session compared to between sessions. When collapsed across all acoustic features and both speakers, the within-session variation was smaller than the between-session one, $F(1,82) = 5.61$, $p < 0.05$.

We quantified the degree of similarity/stability of the motion data in two ways. First, to examine temporal changes, we used Dynamic Time Warping (DTW) [6]. Second, to examine amplitude differences, we normalized the durations and then measured variation of the motion from the mean (expressing this as the area of a 1 SD envelope around the mean).

In terms of the first measure, dynamic time warping (DTW) is a procedure that provides a measure of comparison of two series of data points (inherent distance). For example, DTW can expand or compress one time series to resemble another one and by summing the distances of individually aligned elements to produce an inherent distance (cost) between the two.

We compared the warping cost for each of the principal components (PCs) motions curves for all pairs of within-session utterances and then compared this to all between-sessions pairs (time-series were mean-centered to avoid the effect of off-sets). A summary of the mean inherent distance costs is shown in Figure 3.

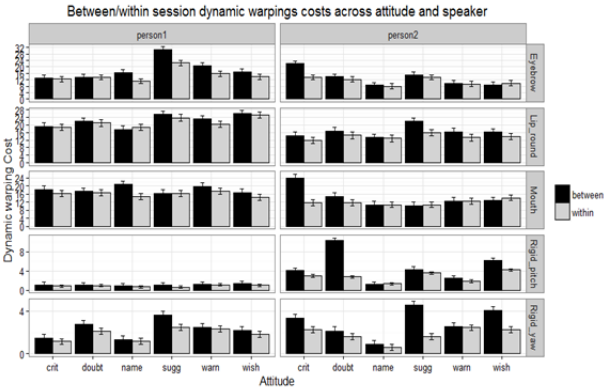

Figure 3: *Mean Dynamic warping cost (inherent distance) for within versus between session utterance pairs for each person.*

As can be seen in Figure 3, warping costs were smaller for within-session utterances compared to between-session ones. Five Bonferroni corrected within- versus between-session repeated measure ANOVAs were conducted (one for each movement type). These were all significant, mouth motion: F1(1,276) = 12.43,, p < 0.05; lip rounding: F(1,276) = 7.21, p < 0.05; eyebrow motion: F(1,276) = 18.65, p < 0.05; rigid head pitch rotation: F(1,276) = 46.72 and rigid head yaw rotation: F(1,276) = 32.74, p <0.05.

The second method we used for quantifying the stability of utterance motions was to examine differences in the amplitude of each principal component curve while normalizing for time. This was done using the following procedure:

a) Normalize the duration of all utterances to a fixed duration of 20 frames (by linear interpolation).

b) Construct an average for each sentence and each speaker.

c) Calculate the area of a one standard deviation (SD) ribbon about the mean. This latter value was then used as an index of variability (see Figure 4 for an example).

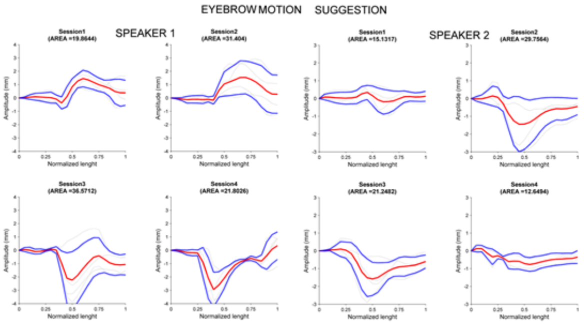

Figure 4: *An example of the variability in the magnitude of the contribution of different PC for each speaker. The red curve depicts the mean; the blue curve shows a 1 SD ribbon around the mean.*

Figure 5 presents a summary of the amplitude variability data for each motion component across the two speakers. As can be seen, there was considerable variation across the speakers.

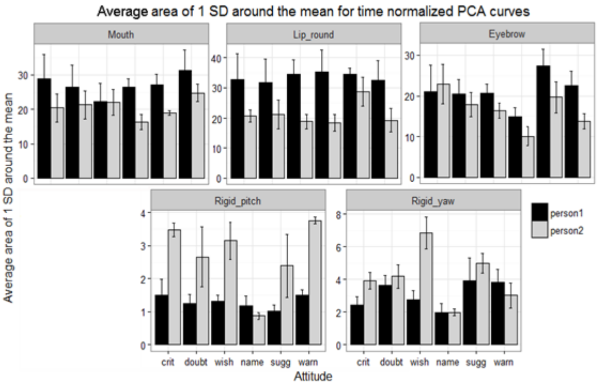

Figure 5: *Average area of 1 SD ribbon around the mean of each of the time normalized PC curves for each person.*

What is interesting is that speaker one had more variability for face motion (i.e., the mouth and to some extent the eyebrows), whereas speaker two showed more variability for head movements (particularly rigid head pitch rotation). An ANOVA comparing mean SD area differences between speakers for all movement types was not significant, F(1,36) = 3.42, p = 0.07; an exploratory analysis comparing area differences by speaker for face versus head movement produced a significant interaction, F(1,36) = 6.54, p < 0.05.

3. Discussion

Following-up a recent study [1] that claimed that the expression of prosodic attitudes has been conventionalized and so are consistent across people, we explicitly examined the production of six attitudes across speaker and within- and across session-variation. We measured acoustic properties along with face and head movements for a single spoken word 'beer" (as used in [1]). Extending the examination to face and head motion is important because prosodic attitudes are expressed both by changes in acoustic features and by changes in face and head motion.

We showed that there was considerable variation in how prosody was realized across two speakers. Also within-session variation for both the acoustic and movement data was smaller than across-session variation, indicating that short-term memory may play a role in consistency.

It should be noted that the current correct LDA classification performance of acoustic data was worse than in [1] where almost perfect classification was achieved. This may have been due to the smaller data set used ([1] had two words and two non-word stimuli and eight repetitions). We are currently collecting data from more speakers and more words. This is important, because there may be reliable difference in how attitudes are expressed by, for example, women and men.

We also need to run a perception study, as differences in how a word is expressed may not necessarily have a one-to-one relationship with what is perceived. For instance, [7] found variable realization of prosody, but [8] good recognition.

4. Acknowledgements

The authors wish to thank Chee Seng Chong for his assistance with data collection and Greg Zelic for his help in data analysis.

5. References

[1] Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. Journal of Memory and Language, 88, 70-86.

[2] Beckman, M. E. (1996). The parsing of prosody. Language and cognitive processes, 11(1-2), 17-68

[3] Ohala, J.J., 1996. Ethological theory and the expression of emotion in the voice. In: Proceedings of the International Conference on Speech and Language Processing. Vol. 3, Philadelphia, USA, pp. 1812-1815.

[4] Maeda, S. (2005). 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.

[5] Badin, P. Bailly, G. Reveret, L. Baciu, M. Segebarth, C. Savariaux, C. (2002). Three-dimensional linear articulatory modelling of tongue, lips and face, based on MRI and video images. J. Phon. 30, 533-553.

[6] Giorgino T. (2009). Computing and Visualizing Dynamic TimeWarping Alignments in R: The dtw Package." Journal of Statistical Software, 31(7), 1{24. URL http://www.jstatsoft.org/v31/i07/

[7] Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. Speech Communication, 57, 317-330.

[8] Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. Cognition, 122(3), 442-453.