INVESTIGATING DISTINCTIVENESS AND INDIVIDUAL VARIATION IN THE EXPRESSION OF VISUAL PROSODIC ATTITUDES

Jeesun Kim and Chris Davis
The MARCS Institute, Western Sydney University
j.kim;chris.davis@westernsydney.edu.au

ABSTRACT

A talker can communicate different attitudes simply by changing how an utterance is expressed rather than by what is said. Typically, such changes in prosody have been investigated by measuring vocal properties; here we examined the expression of different attitudes by measuring changes in visual ones (Facial Action Units and head motion). Using multinomial logistic regression and a recognition experiment, we determined the extent to which different attitudes can be discriminated, and the variability of expressions within and across production sessions. Ten talkers expressed six attitudes, "warning", "criticism", "doubt", "suggestion", "longing", "neutral" in four within-session trials across four different day sessions. Face/head motion was tracked using a Constrained Local Neural Field model on 2D movies. The regression models and recognition experiment showed that attitudes were discriminable; with some better discriminated than others and some talkers much clearer than others. Within-talker productions were more consistent, both within and across sessions.

Keywords: Visual prosody; prosodic attitudes; expressive speech.

1. INTRODUCTION

In addition to symbolic content, speech can convey expressive information about such things as a talker's emotions and attitudes. Understanding how speech conveys such expressive information is important for ultimately deciphering what a speaker means [1]. In this study, we investigated how talkers intentionally express attitudes, for example, how a talker may express that she is surprised, or doubtful.

Typically research on how attitudes are expressed has taken an auditory perspective, examining change in three basic prosodic properties: speech timing, amplitude and fundamental frequency. Whereas change in fundamental frequency relates only to auditory speech, timing and amplitude can apply also to change in visual properties that are transmitted by talkers. That is, when speakers can see each other, expressive speech information can be conveyed visually, e.g., by non-rigid face motion and rigid head motion.

In this study, we chose to chiefly investigate *visual* prosodic attitudes. We did so partly because studies of visual prosodic expression are scarce, but also because we believe that there are interesting unexplored issues concerning the consistency and reliability of such visual signals.

Early speculation that the auditory expression of attitudes is likely to vary across individuals and be context dependent [2] appears not to have been borne out by research. For example, based on the results from a production study of four talkers producing single word (or nonword) utterances that expressed six different prosody attitudes, it was argued that the prosodic forms associated with attitude expression are highly conventionalized and stable across individuals and can be realized without context [3]. This argument was based on finding that the different attitudes could be clearly distinguished by the pattern of their prosodic features as indicated by discriminant analyses used to predict attitude class.

Although this may be the case for auditory prosodic attitudes, no similar study has been carried out for visual prosody. Why might the expression of auditory and visual prosodic attitudes be different? This idea comes from a proposal by [4;5] that there is a difference in the degree that auditory and visual information is distinctive for attitudes that express a proposition content (e.g., irony, incredulity, etc), and those that express more social attitudes, i.e., those that make references to interpersonal relations, (e.g., politeness, arrogance, etc.). Here, it is claimed that auditory information plays a more important role in the expression of propositional attitudes (more connected with linguistic function), whereas visual information is more crucial for the expression of social attitudes.

The proposal that the visually expressed prosodic attitudes may be less robust than their auditory counterparts (at least for non-social attitudes) illustrates the importance of determining the extent of variation in the production and perception of these expressions. To this end, the current study examined the consistency across individuals and sessions of the visual expression of different prosodic attitudes.

To do this, we videoed ten people (far more than other studies) who each intentionally produced the

word ‘beer’ in six different attitudes (those used in [3; 6]), over four within-session trials across four different sessions (ran at least one day apart). From each frame of the captured videos we extracted face and head motion features (see Method). We then used a multinomial logistic regression model to determine, for each person, how well these features could be used to classify the different attitudes. To examine how consistent a person’s attitude productions were, we trained a regression model on the data from the first two sessions and used it to classify the data from the last two sessions. To determine whether people used a similar pattern of features to signal the different prosodic attitudes we determined whether there was a loss of fit (more errors) when logistic regression was applied to the data from all participants compared to individual fitted models.

A follow-up perception experiment (using the productions of three talkers) tested how discriminable the different prosodic attitudes were for visual only; auditory only; and auditory-visual presentations.

2. METHOD

2.1. Participants

2.1.1. Production experiment

Five female and five male native speakers of Australian English were paid a small amount to take part in the production experiment (mean age 25.8 years, range 22-28). Participants were non-actors as actor renditions may be less representative of typical language use.

2.1.2. Perception experiment

Twenty-two participants (first year students from Western Sydney University) took part in the perception experiment for course credit.

2.2. Production experiment

Color Image sequences were captured at 24 fps (640 x 480, VGA) using a Carmine 1.09 camera. In addition, 3D face and head data were captured from the IR sensor of the Carmine close-range sensor (0.35m - 1.4m). This latter data was only used in the current experiment to check the adequacy of the 2D data capture (which proved accurate).

2.2.1. Production materials

The ten speakers expressed the spoken word “beer” using six different attitudes: criticism, doubt, suggestion, warning, wishful and neutral naming.

2.2.2. Production procedure

Each talker was recorded individually. Talkers were seated in a quiet room with the camera positioned directly in front at face level and at approximately 0.6 metres distance. In the test session, image acquisition was controlled by an operator in a separate control room who ensured that the participants looked at the camera throughout the capture performance. The different prosodic attitudes were elicited using the same procedure as [3, 6]. 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. For each new prosodic attitude, the speaker said aloud 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. Each speaker participated in four sessions that took place at least one day apart.

2.2.3. Image processing

The quantification of speech related face and head movements was carried out by analysing each of the captured videos (that were segmented to show just the production of the key spoken word ‘beer’) using a state-of-the-art program, ‘openface’ designed for continuous head pose estimation and facial action unit recognition [7].

This program uses a Constrained Local Neural Field [8] for facial landmark detection and face tracking. Face appearance features are obtained by extracting Histograms of Oriented Gradients. The model accurately detects facial action units [9] by being trained on seven public face-expression databases using support vector machines and support vector regression for the facial action unit intensity estimate. Facial action units are quantified in the output of the model for presence intensity.

2.2.4. Image features and quantification

The openface program outputs 18 facial action units and provides a binary decision on whether the action unit was present or not and how intense is the action unit was (minimal to maximal). In order to quantify the presence of action units in each video token, we calculated the proportion of frames that the action unit was active. To quantify the intensity of the active facial units, we used the average intensity of non-zero (i.e., active) frames. Therefore, for each video, 36 features were produced. In addition to facial action units, we also used head-pose change measures. Openpose produces six indices of head motion, three translation measures in the x, y, z axes, and three

measures of head rotation, i.e., pitch (Rx, glossed as head nodding), yaw (Ry, glossed as head shaking), and roll (Rz, glossed as the head 'maybe' gesture).

2.2.5. Production analysis

To quantify how well prosodic attitudes could be discriminated using the above features, we used multinomial logistic regression (with ridge estimators [10]) and used 10-fold cross-validation (CV) for individuals and for the group data. We also constructed a multilayer perceptron (MLP) that had a 10-unit hidden layer (5000 trials for training and a learning rate of 0.3 and momentum of 0.2) to investigate whether a non-linear decision boundary improved the model (again with 10-fold CV). As mentioned above a trained regression model was used to determine how consistent a person was in producing prosodic attitudes.

2.3. Perception experiment

2.3.1. Materials

The stimuli for the perception experiment consisted of the videos recorded from the production study. In order to keep the length of the experiment manageable, tokens from only three talkers were used. These stimuli consisted of tokens from the five prosodic attitudes (plus the neutral condition). Also, one token from each of a talkers four sessions was used with three presentation conditions, Audio-Only (AO), Visual-Only (VO) and Audio-Visual (AV), making a total of 180 trials in all.

2.3.2. Procedure

Participants were tested individually in a sound-attenuated booth. The videos were presented using the DMDX display software [11]. Trials were blocked by talker and the talker neutral naming video was presented at the beginning of the block for the purpose of providing a talker-specific calibration.

3. RESULTS

3.1. Production

Table 1 shows the results of the 10-fold cross-validated logistic regression analysis to classify the six prosodic attitudes from visual features for each of the 10 talkers (% correct).

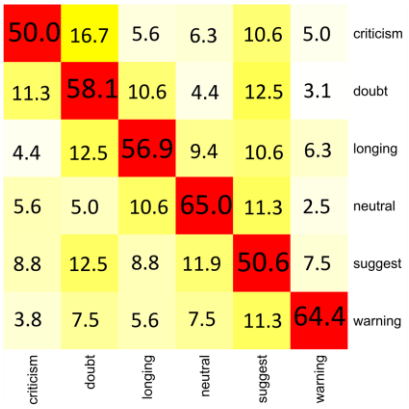
As can be seen, correct classification rates differed considerably across talkers (ranging from 89.5% correct to 36.4% correct). The table also shows the results for the multilayer perceptron. As can be seen, there was a general improvement in classification performance of about 5% (although the level of improvement varied considerably). Table 1 also shows the results from training a regression model on data from a talker's first two sessions to predict classification in the last two sessions.

Table 1: The left column shows results of the 10-fold cross-validated logistic regression analysis classifying the six prosodic attitudes from visual features for each of the 10 talkers; centre column shows results of the multilayer perceptron; the right column shows the results of the predicting the data of the last two sessions (% correct).

Talker	Log Reg	MLP	Prediction
T1	89.5	94.8	87.5
T2	78.1	82.3	68.8
T3	71.9	79.2	66.7
T4	59.4	66.7	68.8
T5	56.3	60.4	64.6
T6	54.2	69.8	58.3
T7	50.0	55.2	30.0
T8	49.0	50.0	37.5
T9	40.6	57.3	31.3
T10	36.4	36.4	41.7

Figure 1 shows a confusion matrix generated by averaging each individual talker's confusion matrix as generated by the regression analysis. Neutral and warning were classified best; followed by doubt and longing; with suggest and criticism the poorest.

Figure 1: Confusion matrix based on the individual talker's regression analyses. The vertical gloss represents the attitude label and the horizontal one the classification result



The average percent correct classification of each individually calculated regression analysis was 58.5%. If the regression analysis was calculated on the entire dataset at once, the performance declined to 48.1%, indicating that different talkers likely used different features in their productions.

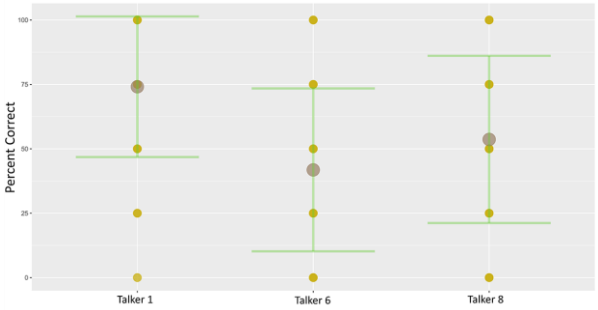
To explore the feature sets used by each talker, we used logistic regression (5-fold CV) to select that top

ten features that best predicted the prosodic attitudes. Across all 10 talkers, 23 features were common in the set of top 10 features. The most common features consisted of a mix of face and head motion (e.g., Brow lowering; nodding; Cheek raising, head motion towards/away from the camera). The three talkers for whose data the regression analysis gave the best results had about half of the same features in their top 10. For the three talkers where the regression analysis gave the worst results, only a quarter of features were shared. This suggests that their use of features to convey the attitudes was somewhat idiosyncratic.

3.2. Perception

Figure 2 shows the mean percent correct attitude recognition for the Visual Only display condition as a function of the production of the three talkers.

Figure 2: Percent correct mean attitude recognition as a function of Talker (mean = large circles, whiskers = SD), VO display.



Correct scores as function of Talker were analysed using a Generalized Linear Mixed Model (using the R lmer package [12]) with random slopes for participants and items; comparison between conditions (Talkers) was conducted using the multcomp package, Tukey contrasts [13]. The analysis indicated that there were significant differences in recognizing the attitudes expressed by each of the three talkers, Talker 1 vs. Talker 6, z-value = 9.776, $p < 0.001$; Talker 1 vs. Talker 8, z-value = 6.471, $p < 0.001$ and Talker 6 vs. Talker 8, z-value = 3.643, $p < 0.001$. It is of interest to note that the pattern of these differences between the Talkers is similar to that shown by the regression analysis.

Table 3 shows the percent correct recognition rates (VO condition) as a function of expressed attitude (note that neutral naming was used as calibration items and so were not tested).

As can be seen in the table, performance was best for the expression of warning and worst for the expression of criticism (the same best and worst scores occurred in the classification results).

Table 2: Mean percent correct VO attitude recognition scores (SE) for the perception experiment and correct classification scores for comparison

Attitude	Perception	Regression
Criticism	51.9 (4.5)	50.0
Doubt	52.3 (3.7)	58.1
Longing	56.8 (4.4)	56.9
Suggest	58.0 (3.2)	50.6
Warning	63.6 (4.4)	64.4

The GLMM analysis on the perception data (using the multcomp package, Tukey) indicated that the only statistically significant differences were between warning and criticism (z-value = 2.931, $p = 0.028$) and warning and doubt (z-value = 2.829, $p = 0.039$).

Table 3. shows the percent correct recognition rates for the VO, AO and AV presentation conditions as a function of expressed attitude.

Table 3: Mean percent correct **for** VO, AO and AV attitude recognition scores (SE)

Condition	%Correct	SE
VO	56.6	2.1
AO	48.5	2.4
AV	67.7	2.2

The GLMM analysis on the perception data (using the multcomp package, Tukey) showed that there was a statistically significant difference between all conditions (VO vs. AO, z-value = 4.258, $p < 001$; VO vs. AV, z-value = 6.103, $p < 001$; AO Vs. AV, z-value = 10.239, $p < 001$).

4. DISCUSSION

The regression results showed that the different prosodic attitudes could be classified from visual features at a rate far better than chance (i.e., average individual scores 57.5% vs 16.7%). Unlike [3] the overall classification was far from ceiling (for [3] classification was 92% correct) and varied over talkers. This suggests that the visual prosodic signal may be more variable than the auditory one; although the perception results, where VO was higher than AO recognition, seems problematic for this interpretation. The AV recognitions results indicate that multimodal presentation was best; our future work will use classification analysis on both visual and auditory signals of attitude.

5. REFERENCES

[1] Kim, J., Bailly, G., & Davis, C. (2018). Introduction to the Special Issue on Auditory-visual expressive speech and gesture in humans and machines. Speech Communication, 98, 63-67

[2] Ohala, J. J. 1996. Ethological theory and the expression of emotion in the voice. *Proceedings of the International Conference on Speech and Language Processing* 3, 1812-1815.

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

[4] Moraes, J. A. D., Rilliard, A., Mota, B. A. D. O., & Shochi, T. (2010). Multimodal perception and production of attitudinal meaning in Brazilian Portuguese. In Speech Prosody 2010-Fifth International Conference.

[5] Moraes, J. A., Rilliard, A., Erickson, D., & Shochi, T. (2011). Perception of attitudinal meaning in interrogative sentences of Brazilian Portuguese. In Proc. of ICPhS.

[6] Kim, J., & Davis, C. (2016). The Consistency and Stability of Acoustic and Visual Cues for Different Prosodic Attitudes. In INTERSPEECH (pp. 57-61).

[7] Baltrušaitis, T., Zadeh, A., Lim, Y. C., Morency, L-P. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit, *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[8] Baltrušaitis, T., Robinson, P., Morency, L-C (2013). Constrained Local Neural Fields for robust facial landmark detection in the wild. IEEE Int. Conference on Computer Vision Workshops, 300 Faces in-the-Wild Challenge, 2013.

[9] Action Units: Facial Action Coding System. https: github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units (accessed 13/12/2018).

[10] Le Cessie, S., Van Houwelingen, J. C. 1992. Ridge estimators in logistic regression. *Applied statistics*, 191-201.

[11] Forster, K. I., Forster, J. C.. DMDX: A Windows display program with millisecond accuracy. *Behavior research methods, instruments, & computers* 35(1), 116-124.

[12] Bates, D., Mächler, M., Bolker, B. Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1-48.

[13] Hothorn, T., Bretz, F., Westfall, P. 2008. Simultaneous Inference in General Parametric Models. *Biometrical Journal* 50(3), 346-363