EXAMINING VARIABILITY IN THE PRODUCTION AND PERCEPTION OF PROSODIC ATTITUDES

Jeesun Kim, Kate Raue and Chris Davis
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University
j.kim@westernsydney.edu.au, k_alimon@hotmail.com, chris.davis@westernsydney.edu.au

ABSTRACT

We tested the extent that expressions of prosodic attitude have been conventionalised by examining their variability and recognition accuracy. We used acoustic and perceptual measures of recordings of 10 speakers (five females) expressing six different prosodic attitudes 16 times (4 trials x 4 sessions). Recognition scores from 46 listeners were better than chance but generally poor. The top 10 acoustic attributes from classification models showed most speakers used some form of F0 variation to convey different attitudes, but few attributes were common across speakers. Productions from speakers whose use of prosody was more consistent across trials, i.e., productions better classified by four-fold cross-validation models, were better recognised. Further, these consistent models led to better classification when applied to the productions of other speakers with relatively high than low recognition scores. We suggest different attitudes can be appropriately signalled by prosodic forms, but these are only partly conventionalised.

Keywords: prosody; prosodic attitudes; expressive speech

1. INTRODUCTION

Speech conveys information through word meaning and word ordering, and via prosody, i.e., the melodies of pitch variation and the rhythm of speech timing. Traditionally, prosodic information has been classified into two broad categories: linguistic and paralinguistic information. Linguistic prosody refers to prosody that serves a linguistic function such as word stress, sentence focus, segmenting the speech into phrases and signalling their broad pragmatic categories. Given this role in language processing, it is presumed that knowledge and use of linguistic prosody has been conventionalised. Paralinguistic prosody, on the other hand, signals such things as a speaker's affect and their intention or attitudes and as such does not have a linguistic function per se. The current study examined the extent to which the knowledge and use of paralinguistic prosody have been conventionalised, specifically focussing on prosodic attitudes.

Traditionally, it has been argued that for attitudes to be appropriately communicated, they need to be presented in context [1]. That is, context, including the speaker and receiver relationship and the accompanying conversation, makes a large contribution to the success of perceiving the attitude [2]. However, more recent research has suggested that the prosodic forms associated with the expression of attitude have been conventionalized and can be appropriately realized even without context [3].

In [3]), the conventionalisation of prosodic forms was tested by determining how similar speaker productions of the different attitudes were, and how well listeners could identify the intended attitudes. To do this, they had four speakers' express 6 different attitudes (criticism, doubt, suggestion, warning and naming) when saying the German word "Bier" (beer). To assess production, the study used a set of 7 acoustic features (e.g., stimulus duration, mean intensity, harmonics-to-noise ratio (HNR), mean fundamental frequency (f0), the difference between offset and onset f0, spectral centre of gravity and the standard deviation of the spectrum). The results of a discriminant analysis showed the correct attitudes were classified with 92% accuracy. This high level of discrimination indicated the distinctiveness of the prosodic patterns that speakers used to express the different intentions. To assess how well the attitudes were recognised, a forced-choice task was used and mean correct recognition measured. Correct recognition accuracy was 82%. This high recognition rate along with the production discrimination result was taken as strong evidence that the prosodic forms for the expression of attitudes had been conventionalised.

An issue with [3], however, is that the speakers used were voice coaches who would have had expert knowledge and use of paralinguistic prosody. Indeed, another study [4] that used a similar method as [3] but tested 2 speakers who had no voice training, found considerable variation in how speakers produced the different attitudes and much lower recognition scores. This result suggests that the findings of [3] may not generalise.

One way that the results of the above two studies could be reconciled is if prosodic attitudes have only been partly conventionalised. That is, the extent of knowledge and use of the conventionalised forms vary across individuals. Given this, several predictions can be made. First, when a speaker has knowledge of and uses a more conventionalised prosodic pattern, their productions should be more consistent over trials and better recognised than those of speakers who did not have such knowledge. Second, speaker productions that are better recognised should be more like each other than those of speakers that were not well recognised. The current study aimed to test these predictions.

In the current study, we asked 10 speakers with no voice-training to produce the two-word phrase “first class” to express six different attitudes. We first determined how variability productions were by using classification models. As a first step, we identified the top 10 acoustic attributes used in each speaker’s trained classification model and examined how common these were across speakers. We then tested the first prediction above by determining the association between classification performance and recognition score. Finally, we tested the second prediction by examining the extent that classification models of speaker productions that were well recognised could classify other well recognised productions compared less well recognised ones.

2. METHOD

2.1. Production study

2.1.1. Participants

Ten native Australian English speakers (5 females, mean age = 25.8 years, range 22-28) participated in the study and given a small reimbursement. The speakers had no voice training.

2.1.2. Recording

Auditory recordings were made using an externally connected lapel microphone, (an AT4033a audio-technica microphone) in 44.1 kHz, 16-bit mono.

2.1.3. Procedure

Each talker was recorded individually. Talkers were seated in a quiet room with a camera positioned directly in front at face level at approximately 0.6 metres distance. Recording was controlled by an operator in a control room who ensured that the participants looked at the camera throughout the capture. The different prosodic attitudes were elicited using the method outlined in [3]. That is, the speaker read short scenarios that described a situation in which she/he would say the carrier phrase ‘first-class’ to the interlocutor using each of 6 different attitudes (criticism, doubt, longing, suggestion, warning and naming). For each scenario, the speaker freely vocalised until ready to say the target phrase with the intended attitude. In each session the phrase was said four times in each prosodic attitude. Each speaker participated in four sessions that took place at least one day apart.

2.2. Perception study

2.2.1. Participants

Forty-six (38 females) first year students (mean age = 21 years; range = 17-43 years) participated in the study for course credit. Participants reported normal or corrected-to-normal hearing and vision and were native speakers of Australian English.

2.2.3. Procedure

Each participant was tested in a sound-attenuated booth and wore Sennheiser HD 660S headphones. Participants were told that they would hear different speakers (presented one at a time) expressing five different attitudes using the phrase ‘first class’. That is, the experiment was blocked and randomised by speaker and the order of trials was randomised within each speaker block. The participant was told that at the start of each speaker block, three audio recordings of the speaker saying the phrase ‘first-class’ in neutral naming mode would be presented to provide speaker specific calibration. Following this, the remaining five attitudes (criticism, doubt, warning, longing and suggestion) would be presented in random order. After each of these trials, they would see a screen with boxes containing the names of the attitudes. Their task was to choose the attitude they thought was being expressed. Participants were given a practice trial that used audio recordings by a speaker who was not among the ten speakers used for subsequent analysis. The participants were not given feedback on their responses. The display of trials and recording of responses was done by the DMDX system [5].

3. RESULTS

3.1 Behavioural results

The results were analysed by fitting a logistic mixed model to predict percent Correct recognition with the variables Attitude and Speaker as fixed factors (formula: Correct recognition ~ Attitude * Speaker). We used the AFEX R package [6] to obtain p-values; this package uses effect coding. The model included Participant and Item as random effects (formula: list(~1 | Participant, ~1 | Item)). The model's total explanatory power was low to moderate (conditional R^2 = 0.18) with the part related to the fixed effects alone (marginal R^2) = 0.10. The results of this analysis are shown in Table 1.

	Df	Chisq	Chi Df	P-value
Speaker	48	219.14	4	.0000
Attitude	43	174.50	9	.0000
Attitude x Speaker	16	339.93	36	.0000

Table 1: Results of the regression model for the effect of Speaker, Attitude and their interaction.

There were significant main effects of Attitude and Speaker and a significant interaction between these variables. To illustrate the overall results, mean percent correct recognition scores are plotted by Attitude (Figure 1) and by Speaker (Figure 2).

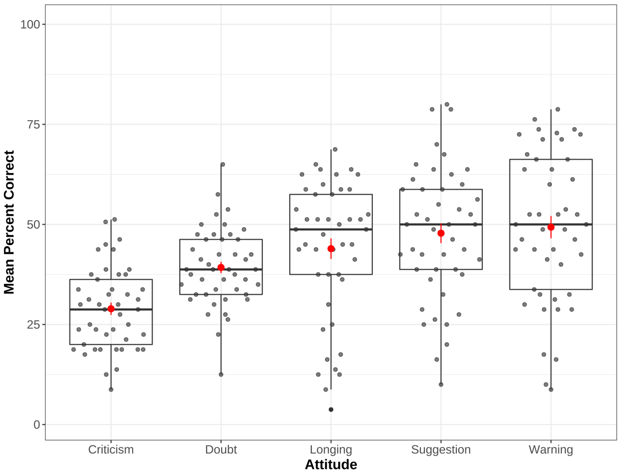

Figure 1: Mean percent correct recognition of each attitude (Tukey-style boxplot).

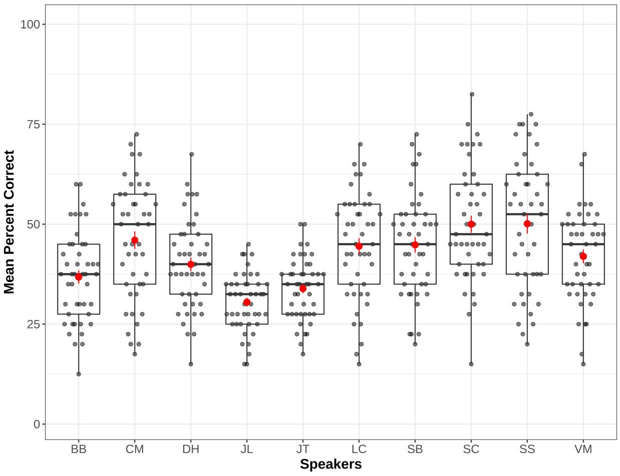

Figure 2: Mean percent correct recognition of each speaker (Tukey-style boxplot).

As can be seen in Figure 1, although recognition levels are better than chance (20%), they are well away from ceiling. The overall percent correct recognition was 41.8% (SE 1.2%). Figure 2 shows that there was considerable across-speaker variability in recognition scores.

3.2 Classification results

3.2.1. acoustic attributes

To represent the auditory productions, we used the 384 acoustic attributes of the 2009 Interspeech emotion challenge [7] via the openSMILE program{ref}. These attributes consisted of three types of low-level descriptors, broadly similar to those used in [3], i.e., zero-crossing rate, root mean square energy, pitch frequency, harmonics-to-noise ratio and 12 mel-frequency cepstral coefficients (numbers 1-12). Also, openSMILE's pitch autocorrelation function was used to calculate a voiced probability measure that represents the probability of a frame representing voiced speech. For each of these descriptors, delta coefficients (i.e., changes in the descriptor over time) were calculated, as well as statistical measures such as the mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range. In addition, linear regression was performed to calculate the offset, slope, and mean square error of the audio signal. Acoustic attributes were calculated for the utterances of each speaker. These data were combined, the results standardised and then each speaker's data was saved separately for classification.

3.2.2. Common acoustic attributes

A statistical classifier was used determine how well the expressed attitude of each speaker could be classified (see 3.2.3) and to determine the top 10 acoustic attributes used in each speaker model. We used the C4.5 classifier (a statistical classifier that uses a decision tree algorithm) with 4-fold, cross validation (see 3.2.3). The C4.5 uses normalized information gain as a splitting criterion and a pruning procedure to mitigate overfitting. We used C4.5 rather than a linear discriminant analysis (as used by [3]) because the acoustic attributes used in classification are clear.

Figure 3 is a chord diagram illustrating the distribution of the connections between each speaker and their top 10 acoustic attributes. The main point of the figure is to show the diversity of the acoustic attributes used in each speaker's classification model. A short code was used, rather than the names of each attribute as these could not be graphed. The attribute that 6 speakers had in common (A13), the quadratic error of linear F0 estimate; 5 speakers had A2 in common (the probability of voicing); the remaining common features were related to estimates of F0.

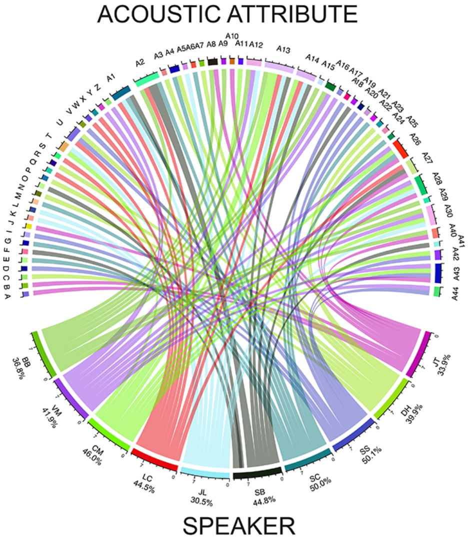
ACOUSTIC ATTRIBUTE
A B C D E F G I J K L M N O P Q R S T U V W X Y Z
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11 A12 A13 A14 A15 A16 A17 A18 A19 A20 A21 A22 A23 A24 A25 A26 A27 A28 A29 A30 A40 A41 A42 A43 A44
JT 33.9%
DH 39.9%
SS 50.1%
SC 50.0%
SB 44.8%
JL 30.5%
LC 44.5%
CM 46.0%
VM 41.9%
BB 36.8%
SPEAKER

Figure 3: A chord diagram showing the connection between each of Speaker (shown with percent correct) and the top 10 acoustic attributes of 4-fold cross-validated classification models (using arbitrary coding, see text)

3.2.3. Model performance and recognition accuracy

Four-fold cross-validation was used to provide a measure of consistency of the calculated classification model for each speaker's data. That is, the performance of a machine learning model was evaluated by dividing the data into four equal-sized folds, or subsets. Four-folds were used because the data were collected over four sessions at least one day apart. The model was trained on three of the folds and tested on the fourth, and this process was repeated four times, with each fold serving as the test set once. If there is consistency in what properties are used to distinguish the different attitudes, then the model created on three folds should produce a good test result on the remaining fold. This entire process was conducted 10 times (each with a random selection of folds) and the average performance taken. This was done for each speaker.

Pearson correlation was used to determine the strength of the association between the performance of each speaker model and participant perception results. The results are shown in Figure 4. As can be seen the figure there was a significant positive correlation between the model and percent correct recognition (r = .63, p = .0049). This suggests that those speakers whose prosodic attitudes were better recognised had a more consistent use of auditory attributes.

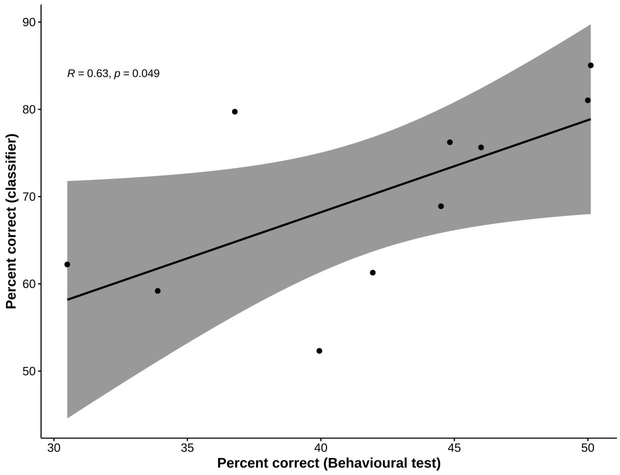

Figure 4: A depiction of the correlation between the behavioural recognition scores and classifier performance.

3.2.4. Model transfer

If prosodic patterns used to express different attitudes were more common (conventional) for speakers whose attitudes were better recognised, then the classification models for these speakers should be more transferable than those of speakers who attitudes were not well recognised. To assess this, we classified the auditory attributes of each speaker (using 10 4-fold cross validated classification runs) using the trained model of each other speaker. We grouped the speakers into those whose attitudes had high or low recognition scores and calculated the mean classification scores for these groups. The results showed that models trained and applied across the high group had significantly better classification (28.9 % correct, SE = 1.9%) than models trained and applied across the low group (19.5% correct, SE = 1.9%), t = 3.412 < p = 0.0373.

4. DISCUSSION

Overall, listeners recognised attitudes from prosody better than chance, but well below the level shown in [3]. The productions of speakers whose prosodic attitudes were better recognised than others, tended to be better classified and these classification models were better at classifying other well recognised productions. These results are consistent with some speakers having a partly conventionalised model of the prosodic forms that represent different attitudes. It seems plausible that with training a speaker could refine such a model to make each attitude more distinctive and hence better recognised (a kind of super–normal stimulus), perhaps like the voice coaches in the [3] study. Yet such highly distinctive models do not appear to occur in speakers with no specific training.

5. RFERENCES

[1] 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.

[2] Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

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

[5] Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior research methods, instruments, & computers, 35, 116-124.

[6] Singmann, H., Bolker, B., Westfall, J., Aust , F. & Ben-Shachar, M. (2022). afex: Analysis of Factorial Experiments. https://afex.singmann.science/, https://github.com/singmann/afex

[7] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.