Auditory and Visual Emotion Recognition: Investigating why some portrayals are better recognized than others

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

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

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

Abstract

To understand the factors that influence auditory and visual emotion recognition performance we examined a perception set of stimuli produced by three talkers that differed in how well people could recognise their emotions. Our proposal was that productions based on a model of prototypical emotion attributes will be more consistent and better recognized. To test this, we trained a classification model on a parallel hold-out set of stimuli by the same talkers and examined the consistency of the emotion portrayals. We found that the emotion stimuli from a talker who produced more consistent emotion portrayals were better recognized than those stimuli that were less consistently produced.

Index Terms: emotion recognition, consistency, logistic regression

1. Introduction

When it comes to the selection of stimulus materials, experimental studies of human emotion recognition face a dilemma. On the one hand, there is the general aspiration for experimental studies to have controlled, unambiguous stimuli; on the other, there is the need that stimuli capture the important properties of the phenomenon under investigation.

In general, research has tended to favour the selection unambiguous stimuli. For instance, investigations of the recognition of facial emotional expressions typically use photographs as stimuli; particularly those that have been shown to produce high recognition rates [1]. This has led to the construction of databases of face emotion stimulus sets that have very high ratings of the intended expression [2].

The situation for the auditory expression of emotions is a little different due to the dynamic nature of the signal. Here, many studies have used brief spoken sentence material as stimuli (e.g., of the 106 studies reviewed by Juslin and Laukka [3], 60 studies used brief sentences); although the use of spoken words, vowels, digits or affect bursts is also common. Indeed, several databases containing non-verbal emotional expressions have been developed [4,5].

The relatively small number of investigations of auditory-visual (AV) emotion recognition have characteristically selected stimuli that have been shown to attract high recognition rates. However, this selection requirement is difficult for spoken AV stimuli, as such often elicit only moderate recognition rates [6]. So studies have used AV stimuli from different sources. That is, one strategy has been to pair a well-recognised photograph of an emotion expression with a simple auditory stimulus that conveys a well-recognized emotion. For instance, Hunter et al [7] used static pictures and short nonlinguistic emotion expressions of the vowel /a/ and Chaby et al [8] used static pictures and nonverbal emotion expressions. Another strategy has been to use spoken sentence AV stimuli, but to only test a few emotions that are very different (thus reducing stimulus confusability) [9]; or to use very short verbal utterances, e.g., Lambrecht et al [10] only examined single words (testing four emotions).

The general use of stimulus materials that depict well-recognised, unambiguous, prototypical emotion stimuli is often motivated by a desire to have the best possible stimulus set to study emotion recognition. However, the use of these materials means that research questions about the factors that modulate emotion recognition performance go largely unaddressed.

To get around this limitation, a recent study combined 23 auditory emotion datasets [11]; the resultant analysis indicated which factors influence emotion recognition. One factor that was identified is the intensity of emotion rendition. Here it was proposed that high intensity portrayals attract higher levels of decoding accuracy than low-intensity ones. It was also suggested that this may hold for facial expression [12].

In these studies, emotion intensity was based on subjective ratings. However, a problem with ratings of emotion intensity is that they can be influenced by the clarity of the emotional expression; as such, relating intensity to recognition accuracy risks circularity. Taking this into account, the current study employed a different way to examine what influences emotion decoding accuracy using a selection of emotion stimuli from a study of Kim & Davis [14] that identified emotion stimuli that varied in recognition accuracy.

Our idea starts with the straightforward notion that emotion portrayals are more likely to be classified correctly when they include prototypical emotion attributes. Of course, specifying what these attributes are, especially in dynamic stimuli, is tricky. Our proposal is to use the consistency of the portrayals as an index of attribute prototypicality. Here, we presume that the consistency of emotion portrayal relates to the prototypicality of the person's emotion model. That is, productions based on a prototypical model of emotion attributes will be more consistent and better recognized.

To examine this, we selected from [14] auditory (A) and visual (V) emotion stimuli from talkers whose emotion renditions were either well recognized, moderately recognized, or poorly recognized. In addition to this set of emotion stimuli (the perception set) we selected a 'hold-out' set of stimuli from the same talkers (producing emotions for the same set of base sentences). Using logistic regression, we built classification models from the auditory and visual attributes of the hold-out stimuli and determined how well these classified the stimuli of the good, moderate and poorly recognized perception sets. Our expectation was that the better

the fit of the hold-out model to the perception set, the better would be the emotion recognition performance on that set.

In summary, we selected auditory and visual emotion stimuli produced by three talkers that differed in how well people could recognise their emotions and built logistic regression models. We determined the relationship between accuracy scores and the accuracy of logistic regression models; and then examined the consistency of emotion portrayal for the three talkers by determining how well a classification model trained on a hold-out set could classify the perception stimuli. In addition, we also compared the magnitudes and variation of the top auditory and visual classification attributes across talkers to determine if those from the more accurate talker differed from those of the other talkers.

2. Method

2.1. Production participants

Three male native speakers of Australian English (in their early twenties) were recruited to record the face and voice stimuli. These speakers were chosen to capitalize on natural variation in people's ability to express emotions (e.g., [13]). These particular speakers were selected based on the results of an emotion recognition task [14]

2.2. Materials

Audio and video recordings consisted of three male native Australian English talkers uttering 10 Semantically Unpredictable Sentences [15]. Talkers portrayed facial and vocal expressions of anger, sadness, disgust, surprise and happiness, as they spoke each sentence. They were instructed to produce these expressions as if they were communicating this emotion to an observer. Recordings were rendered to produce AO, VO, and AV stimuli.

2.3. Perception participants

Fifty-five undergraduate female students (age range 18-23) from Western Sydney University participated in the perception experiment for course credit. All were speakers of Australian English with self-reported normal or corrected normal vision and no history of reported hearing loss. The sex of the perceivers (female) and the producers (male) was kept constant as the results of [16] suggest that the sex of both perceiver and producer influences categorization of AO, VO, and AV expressions.

Audiovisual recordings were made using a Sony TRV19E digital video camera (25 fps) externally connected to a lapel microphone (44.1 kHz, 16-bit stereo) in a well-lit IAC booth. The video recording included the head and top part of the neck. AV recordings of individual emotion expression tokens were extracted from the full recording based on a hand-labelled Praat textgrid and by a tailored Matlab script.

2.4. Perception procedure

At the beginning of the session, each presenter was given the 10 sentences to read to be familiar with them. For the capture session, each sentence was displayed one at a time on a computer monitor (next to the camera) until the participant pressed the space bar to get the next sentence. On each screen was the name of the emotion that he should express when saying the sentence. The emotions to be expressed were blocked so that the participant would say all 10 sentences expressing the same emotion. The whole session was repeated to obtain two sets of materials (perception and hold-out sets).

2.5. Auditory and Visual attributes

2.5.1. Auditory attributes

We used the auditory attributes of the Interspeech 2009 emotion challenge [17]. This consists of 384 attributes that includes 16 low-level descriptors: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and 1-12 mel-frequency cepstral coefficients (MFCC). For each of these descriptors delta coefficients were calculated, as well as the mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range. Also, linear regressions were calculated in terms of offset, slope and mean square error (MSE).

2.5.2. Visual attributes

The quantification of speech-related face and head movements was done by analysing each of the segmented videos using 'openface' [18]. The program, trained on seven public face-expression databases, detects facial action units [19] using support vector machines, and support vector regression for the facial action unit intensity estimate (see below). Facial action units are quantified in the output of the model in terms of the presence of an action unit and the intensity of activation of that unit. That is, the program outputs 18 facial action units and provides a binary decision on whether the action unit was present or not and how intense the action unit was (minimal to maximal).

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. Openface 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). We quantified each of these movements by calculating the coefficient of variation of the motion over frames.

Classification accuracy was determined using logistic regression (LR) with a ridge estimator, see [20]. We also used a three-layered neural network with backpropagation, however, due to space limitations we only report the results of the LR model.

3. Results

Mean correct percent accuracy scores from the 55 participants for the auditory-only (AO) and visual-only (VO) stimuli in the emotion recognition study for each of the talkers are shown in Table 1. Also shown are the logistic regression scores (accuracy) based on the auditory and video attributes for the perception items.

For the AO perception data, there was a difference in accuracy across the six emotions, $F(5,380) = 33.4$, $p < 0.001$. There was an effect of Talker, $F(2,76) = 102.8$, $p < 0.001$ and

there was an interaction between these two effects, $F(10, 380) = 21.9$, $p < 0.001$.

Table 1: Mean percent correct AO and VO emotion recognition for human perceivers (Hum), 95% confidence intervals, and the correct classification scores for the Logistic Regression models (LR)

	AO Hum	95% Confid.	AO LR	VO Hum	95% Confid.	VO LR
Talker1	77.6	71.1-81.0	65.0	94.5	90.3-97.0	90.0
Talker2	55.6	50.6-59.5	51.7	51.6	47.3-54.0	56.7
Talker3	32.2	27.3-36.3	48.3	62.6	57.9-64.6	86.6

For the VO perception data, once again there was a difference in accuracy across the six emotions, $F(5,380) = 33.9$, $p < 0.001$. There was an effect of Talker, $F(2,76) = 204.8$, $p < 0.001$ and there was an interaction between these two effects, $F(10, 380) = 24.5$, $p < 0.001$.

To compare the perception and attribute classification data in more detail, Figure 1 shows the confusion matrices for the AO perception data and those generated by the logistic regression models.

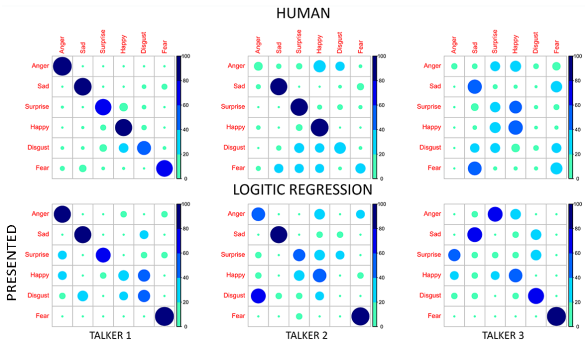

Figure 1: *AO confusion matrices for the human perception and the LR models. The larger and darker the circles the more accurate the score. The presented emotion is vertical and the classification horizontal*

As can be seen, the similarity between the confusion matrices for Talker 2 and 3 is rather weak. A similar result can be seen in comparison of the VO human and LR data shown in Figure 2.

In general, the logistic regression accuracy scores tracked those of human recognition, i.e., the order of the classification performance of the talker's stimuli (e.g., those Talker 1 produced the best classification performance) had a similar order to the perceived accuracy of the talker's renditions (e.g., the stimuli produced by Talker 1 were recognised best). However, the pattern of classification confusion matrices did not match the pattern of the perceptual confusions (except for the stimuli produced by Talker 1).

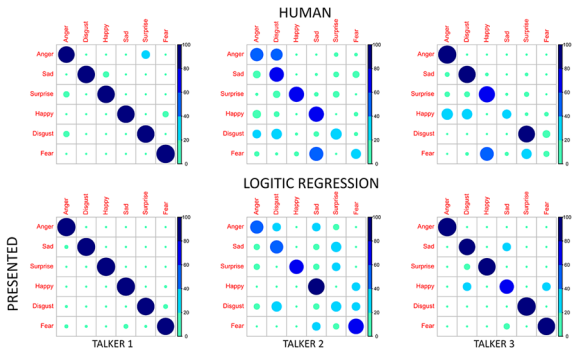

Figure 2. *VO confusion matrices for the human perception and the LR models. The larger and darker the circles the more accurate the score. The presented emotion is vertical and the classification horizontal*

What distinguished the productions of Talker 1 from the other talkers? As mentioned above, Juslin and colleagues [11] suggest that renditions that have a high-intensity lead to higher levels of decoding accuracy than do low-intensity ones. To examine this, we determined for the perception stimulus set of each talker the attributes that had merit in the logistic regression model and used greedy hill-climbing augmented with a backtracking facility to select the best of these. As an example, Table 2 shows for each talker the means, range and standard deviation (averaged across the six emotion types) for the four best attributes for the AO stimuli of Talker 1.

Table 2: *Mean and SD of the best auditory classification attributes from Talker 1 (averaged over emotion types) for all three talkers.*

	pcmRMS energy sma_min	pcm_fft Magmfcc sma[4] amean	pcm_fftMa gmfcc sma[7] amean	pcm_fft Magmfcc sma[9] amean
Talker1				
(mean)	0.001	-2.323	-4.355	-2.594
(range)	0.002	13.74	10.285	10.58
(SD)	0.001	3.616	2.23	2.58
Talker2				
(mean)	0.001	-5.00	-5.411	0.769
(range)	0.006	12.27	11.05	10.875
(SD)	0.001	3.166	2.42	2.538
Talker3				
(mean)	0.002	2.723	-1.187	-1.583
(range)	0.003	7.52	8.4	8.14
(SD)	0.001	1.83	1.93	1.714

Although the range of two of the attributes is greater for Talker 1, it is not clear that the auditory attributes of Talker 1 were that different from those of the other talkers. Means, range and standard deviation (averaged over emotion types) for the two best visual attributes of Talker 1 for all talkers are shown in Table 3.

Table 3: *Mean and SD of the best FACS classification attributes from Talker 1 (averaged over emotion types) for all three talkers (note: the range is not meaningful for the presence data)*

	AU 4 Brow lowerer Presence	AU 7 Lid tightener Intensity
Talker 1		
(mean)	0.465	2.352
(range)	-	4.07
(SD)	0.474	1.177
Talker 2		
(mean)	0.826	1.382
(range)	-	3.5
(SD)	0.375	1.106
Talker 3		
(mean)	0.409	1.012
(range)	-	3.4
(SD)	0.321	0.943

Once again, there is some (limited) support for the stimuli of Talker 1 being more extreme (larger range) than those of the other talkers. That is, one of the visual attributes that support classification for Talker 1 (FACS AU 7) had a larger mean and greater range than that produced by the other talkers.

The final analysis was to examine the fit of the logistic model trained on the hold-out set to the perception set. Classification accuracy for the hold-out LR model on the perception test set for the three talkers is shown in Table 4.

Table 4: *Percent correct classification performance in classifying the perception set on the LR model trained on the hold-out set.*

	AO	VO
Talker1	76	100
Talker2	48	84
Talker3	48	50

As can be seen in the table, classification performance was highest for Talker 1. This likely indicates that both the auditory and visual attributes produced by Talker 1 were similar in the hold-out and perception stimulus sets. It is tempting to suggest that there is a relationship between this greater consistency in emotion portrayals across the two production sessions and the better human recognition of the emotion conveyed by these portrayals. Further, it may be that these emotion portrayals are more intense than those of the other talkers (although a more extensive examination of the attributes of the talkers is needed).

4. Discussion

Emotion stimuli that vary in recognition accuracy provide an opportunity to determine how auditory and visual attributes affect emotion recognition performance. We found that the stimuli from a talker who produced more consistent emotion portrayals were better recognized than those stimuli that were less consistently produced. This result is consistent with the idea that consistency of production and the recognition of emotion portrayals is governed by the extent that expressions are driven by the activation of prototypical attributes. Further, it may be that prototypical portrayals are more intense, given the suggestion of Juslin and colleagues [11] that low intensity emotion expressions are more different to distinguish than higher intensity ones.

5. Acknowledgements

The authors acknowledge the support of an Australian research Council (DP grant DP150104600).

References

[1] N. Tottenham, J. W. Tanaka, A. C.Leon, T. McCarry, M. Nurse, T. A. Hare, ... and C. Nelson, "The NimStim set of facial expressions: judgments from untrained research participants," *Psychiatry research*, vol. 168, no. 3, 242–249, 2009.

[2] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. D. Van Knippenberg, "Presentation and validation of the Radboud Faces Database," *Cognition and emotion*, vol. 24, no. 8, 1377–1388, 2010.

[3] P. N. Juslin, and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, vol. 129, 770–814, 2003.

[4] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The Montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing," *Behavioral. Research Methods,* vol. 40, 531–539, 2008.

[5] C. E. Parsons, K. S. Young, M. G. Craske, A. L. Stein, and M. L. Kringelbach, "Introducing the Oxford Vocal (OxVoc) Sounds database: a validated set of non-acted affective sounds from human infants, adults, and domestic animals," *Frontiers in psychology*, vol. 5, 562, 2014.

[6] A. Battocchi, F. Pianesi, and D. Goren-Bar, "A first evaluation study of a database of kinetic facial expressions (DaFEx)," in G. Lazzari & F. Pianesi (Eds.), *Proceedings of ICMI '05* (pp. 214–221). New York, NY: ACM Press, 2005.

[7] E. M. Hunter, L. H. Phillips, and S. E. MacPherson, "Effects of age on cross-modal emotion perception," *Psychology and Aging*, vol. 25, 779–787, 2010.

[8] L. Chaby, V. Luherne-du Boullay, M. Chetouani, and M. Plaza, "Compensating for age limits through emotional crossmodal integration," *Frontiers in Psychology*, vol. 6, 691, 2015.

[9] C. Wieck, and U. Kunzmann, "Age differences in emotion recognition: A question of modality?" *Psychology and Aging*, vol. 32, 401–411, 2017.

[10] L. Lambrecht, B. Kreifelts, and D. Wildgruber, "Age-related decrease in recognition of emotional facial and prosodic expressions," *Emotion*, vol. 12, 529–539, 2012.

[11] P. N. Juslin, p. Laukka, and T. Bänziger, "The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion," *Journal of nonverbal behavior*, vol. 42, no. 1, 1–40, 2018.

[12] L. G. Tassinary, and J. T. Cacioppo, "Unobservable facial actions and emotion," *Psychological Science*, vol. 3, 28–33, 1992.

[13] H. G. Wallbott, and K. R. Scherer, "Cues and channels in emotion recognition." *Journal of Personality and Social Psychology*, vol. 51, 690-699. 1986.

[14] J. Kim, and C. Davis, "Perceiving emotion from a talker: How face and voice work together," *Visual Cognition*, vol. 20, no. 8, 902–921, 2012.

[15] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, 381–392, 1996.

[16] O. Collignon, S. Girard, F. Gosselin, D. Saint-Amour, F. Lepore, and M. Lassone, “Women process multisensory emotion expressions more efficiently than men,” *Neuropsychologia*, vol. 48, 220–225, 2010.

[17] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *Tenth Annual Conference of the International Speech Communication Association,* 2009.

[18] T. Baltrušaitis, P. Robinson, and L-C(Morency, “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.

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

[20] S. le Cessie, and J. C. van Houwelingen, “Ridge Estimators in Logistic Regression,” *Applied Statistics,* vol. 41, no. 1, 191–20, 1992.