Bimodal Expression of Emotion by Face and Voice

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
Adjunct Faculty, Robotics Institute
Carnegie Mellon University
jeffcohn@vms.cis.pitt.edu

Gary S. Katz
Department of Psychology
California State University at Northridge
gk45683@email.csun.edu

ABSTRACT

A goal of research in human-computer interaction is computer systems that can recognize and understand nonverbal communication. In a series of studies, we developed semi-automated methods of discriminating emotion and para-linguistic communication in face and voice. In study 1, three computer-vision based modules reliably recognized FACS action units, which are the smallest visibly discriminable changes in facial expression. Automated Face Analysis demonstrated convergent validity with manual coding for 15 action units and action unit combinations central to the expression of emotion. In study 2, prosodic measures discriminated pragmatic intent in infant-directed speech with accuracy ranging from 61-65% in test samples. In study 3, facial EMG and prosodic measures combined discriminated between negative, neutral, and positive emotion with accuracy ranging from 47-79% in test samples. These results support the feasibility of human-computer interfaces that are sensitive to the full range of human nonverbal communication

Keywords

Human-computer interaction, para-linguistic communication, facial expression recognition, FACS, prosody, emotion

1. INTRODUCTION

The design and implementation of computer systems that can understand and respond appropriately to the full range of human communication is an important objective of research in human-computer interaction. We present recent progress in the development of computer systems that discriminate between subtle changes in facial expression and recognize a speaker's communicative intent and felt-emotion.

2. FACIAL EXPRESSION

Facial expression communicates emotion and pain, regulates interpersonal behavior, and is computationally related to speech (Cohn, Zlochower, Lien, & Kanade, In press; Ekman & Rosenberg, 1977). In the past five years, significant progress has been made in computer-vision based approaches to discriminating facial expression. Most of these approaches attempt to discriminate between a small set of emotion expressions (Black & Yacoob, 1997; Essa & Pentland, 1994; Mase, 1991; Otsuka & Ohya, 1998). This focus follows from the work of Darwin (1872) and more recently Ekman (Ekman & Rosenberg, 1997) who proposed that "basic emotions" (i.e., joy, surprise, anger, sadness, fear, and disgust) each have a universally recognized facial expression, involving changes in multiple facial regions, which facilitates analysis. In daily life, these prototypic expressions occur relatively infrequently. Emotion more often is communicated by small changes in facial features, such as furrowing of the brows to convey negative affect (e.g., Carroll & Russell, 1997). Consequently, a system that describes only emotion expressions is of limited use. In

Figure 1. (A) Feature-Point Tracking. (B) High-Gradient Component Detection. (C) Dense-Flow Tracking.

addition, emotion labels lack standard meaning. Expressions with the same label often refer to different facial displays (Oster, Hegley, & Nagel, 1992).

To represent the full range of facial expression, we (Cohn, Zlochower, Lien, and Kanade, In Press; Lien, Kanade, Cohn, & Li, 1998) developed a computer-vision based system that automatically recognizes individual FACS facial action units (AUs) or AU combinations and estimates expression intensity. FACS action units (Ekman & Friesen, 1978) are the smallest visibly discriminable changes in facial expression. Three modules are used to extract facial expression information in digitized image sequences: (A) facial feature point tracking, (B) high gradient component detection, and (C) dense flow tracking. Examples of the output of each module are shown in Figure 1. To control for variation in head position, orientation, and scale, image data are automatically aligned by affine or more recently by perspective transformation prior to analysis. In image sequences from 100 subjects, the average discrimination rate in test image sequences was greater than 80% for each of the three modules. Feature-point tracking, which was most extensively tested, had high concurrent validity with manual coding of 15 FACS action units (Figure 2) in all facial regions.

3. VOCAL EXPRESSION

Prosodic features of speech, which include vocal fundamental frequency (f_o), intensity, and rhythm, carry a significant portion of a speaker's meaning (e.g., Cooper & Sorenson, 1981) and felt emotion (Frick, 1985; Murray & Arnott, 1993; Scherer, 1986). For example, by varying the placement of f_o excursions within an utterance, a speaker can alternatively communicate doubt ("Bev loves BOB?") or conviction ("Bev LOVES Bob"). Temporal parameters of speech, such as the duration of turn-taking pauses in dyadic interaction are sensitive to emotion. Zlochower and Cohn (1996), for instance, found that depressed mothers were slower and more variable in responding to their child's vocalizations. Of the prosodic features, f_o may be the most important (e.g., Fernald, 1989; Frick, 1985), but it has been difficult to find reliable mappings between quantitative measures of f_o and communicative intent and emotion. Most studies have used only aggregate measures of f_o or qualitative description of f_o contours (e.g., Fernald, 1989; Stern et al. 1982). For automated analysis, quantitative measurement of both summary and dynamic measures is needed.

3.1 Communicative Intent in Infant-Directed Speech

In an initial study (Katz, Cohn, & Moore, 1996; Moore, Cohn, & Katz, 1994), we focused on the modulation of vocal f_o during mothers' speech to their 4-month-old infants. Forty-nine mothers were instructed to use their voice to get their infant's attention, to communicate approval, and provide comfort. Vocal f_o from 621 utterances was extracted using a Computerized Speech Laboratory (CSL Model 4300 from Kay Elemetrics Corp.) and custom software (Moore, Cohn, & Katz, 1994). Dynamic features were measured by quantitative modeling of f_o shape. The choice of equations was informed by their ecological validity. They model physiologic systems and have been described qualitatively in related research (e.g., Fernald, 1989; Stern et al., 1982). They were 16 functions grouped into seven classifications of curve type (linear, power, transfer, decay, exponential, bell-shaped, and sinusoidal functions). An example of curve fitting can be seen in Figure 3. Summary features were vocal f_o mean, standard deviation, and duration. Dynamic and summary measures each contributed to the discrimination of communicative intent; a discriminant classifier based on both sets of measures demonstrated a moderate to high level of

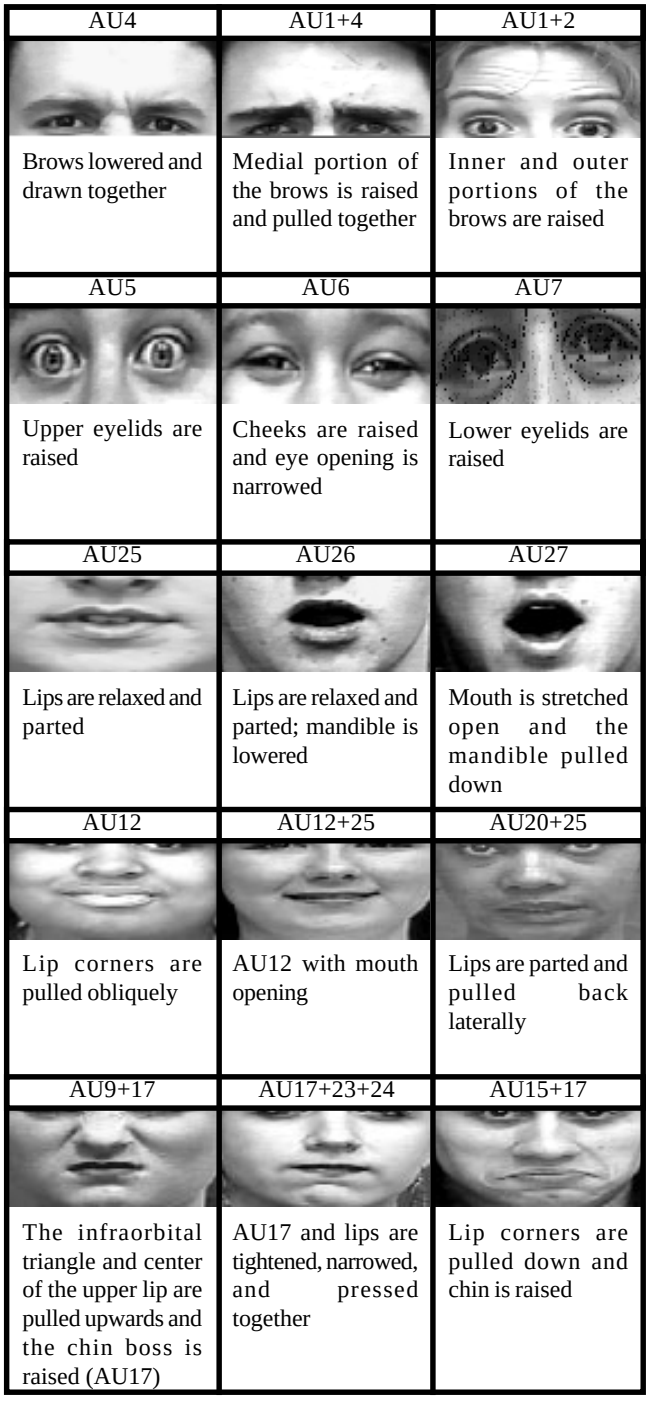

Figure 2. FACS Action Units Discriminated by Automated Face Analysis

agreement between measures of f_o and communicative intent (see Table 1).

3.2 Expression of Emotion in Adult-Directed Speech

Because prosody is exaggerated in infant-directed speech, we (Katz & Cohn, 1998) also investigated whether prosodic parameters were sensitive to felt emotion in adult-directed speech in a context with fewer demands to communicate. Subjects were 112 young adults. The International Affective Picture System

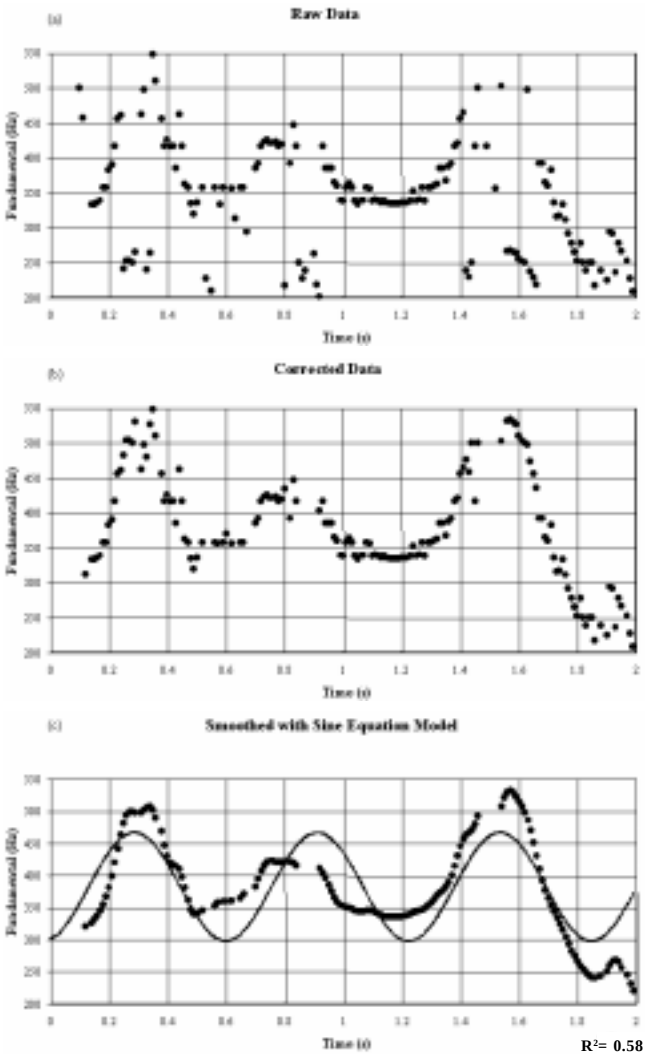

Figure 3. Quantitative Modeling of Fundamental Frequency Contours. The initial contour exhibits discontinuities that result from the extraction algorithm (top panel). Many of these discontinuities can be resolved by correction for half-frequency errors (middle panel). Quantization error is reduced by Fast Fourier Transform to yield the signal detection subjected to modeling (bottom panel). Sixteen functions for seven classifications are fit for each contour. Results for a sinusoidal function are shown in the bottom panel. A sinusoidal curve is fit to the contour in the bottom panel and goodness of fit quantified by R^2. (Adapted from Moore et al., 1994).

Table 1. Cross-Classification Between Actual and Predicted Utterances (Test Sample)

		Predicted Utterance		
	N	Attention	Approval	Comfort
Actual Condition				
Attention	97	0.63	0.10	0.27
Approval	207	0.11	0.61	0.28
Comfort	186	0.18	0.17	0.65

(IAPS: Lang, Bradley, & Cuthbert, 1995) was used to induce seven target-emotion conditions (anger, disgust, fear, happy/calm, happy/ excited, neutral, and sadness). Vocal fundamental frequency was extracted following the method in the previous study. Dependent measures were mean, standard deviation, and duration of fundamental frequency and utterance intensity. To examine the concordance of prosodic measures with facial expression, we assessed facial expression using surface electromyography (EMG). Data were divided into training and test sets and analyzed by discriminant analysis.

In the test set, 33% of emotions were correctly recognized. Disagreements were most common for closely related emotions (e.g., neutral and Happy/Calm). Discrimination was highest for Happy/Excited (63%). Negative emotions were well discriminated from positive from not from each other (i.e., lack of specificity among negative emotions), which was consistent with circumplex models of emotion (Larsen & Diener, 1992). When discrete emotions were aggregated into more molar categories of negative, neutral, and positive emotions, recognition rates increased to 56%, 47%, and 79% for negative, neutral, and positive emotions respectively.

Facial EMG and prosodic measures both discriminated between emotions and showed moderate convergent validity with heart rate and self-reported emotion. The level of discrimination was lower, however, than that found in infant-directed speech. This was not surprising in that the communicative demands were reduced relative to the task of communicating with a preverbal infant and the emotion elicitations were generally mild. Also, the obtained discrimination rate was comparable to that obtained in a study of vocal expression of emotion in professional actors (Base & Scherer, 1996).

The combination of facial and vocal measures was most powerful, which supports the need to use convergent measures of communicative intent and emotion in designing computer systems that understand human behavior. In addition, because of individual differences in how people use para-linguistic cues and express emotion, a system that uses convergent measures of expression and physiology may prove more robust.

4. CONCLUSION

We designed and developed a prototype computer-vision based system that discriminated subtle changes in facial expression and acoustic systems that discriminated communicative intent in infant-directed and felt emotion in adult-directed speech. In the study that combined facial and prosodic measures,

convergence with self-reported emotion was greatest when both types of measures were combined. The use of convergent measures may afford more robust results, especially when communicative demands or emotion intensity are reduced. Although significant problems remain to be solved, computer-user interfaces that are sensitive to the full range of human para-linguistic communication appear feasible.

5. ACKNOWLEDGMENTS

This research was supported by NSF grant BNS-8919711 and NIMH grant R01 MH51435 to Jeffrey F. Cohn. Our collaborators on the projects described include Christopher Moore (University of Washington), James Lien, Takeo Kanade, Wei Hua, Yingli Tian, and Yu-Te Wu (Robotics Institute, Carnegie Mellon University), and Adena Zlochower, Zara Ambadar, and Chelsea Jankel (University of Pittsburgh).

6. REFERENCES

[1] Banse, R. and Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 614-636.

[2] Black, M. and Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision, 25*, 23-49.

[3] Carroll, J. M., & Russell, J. A. (1997). Facial expressions in Hollywood's portrayal of emotion. *Journal of Personality and Social Psychology, 72*, 164-176.

[4] Cohn, J.F., Zlochower, A., Lien, J., & Kanade, T. (In press). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology*.

[5] Cooper, W.E. and Sorenson, J.M. (1981). *Fundamental frequency in sentence production.* NY: Springer-Verlag.

[6] Essa, I.A. and Pentland, A. (1994). A vision system for observing and extracting facial action parameters. *Proceedings of the IEEE Conference on Vision and Pattern Recognition (CVPR'94).*

[7] Ekman, P. & Friesen, W.V. (1978a). *Facial action coding system*. Palo Alto: Consulting Psychologist Press.

[8] Ekman, P. & Rosenberg, E. (1997). *What the face reveals*. NY: Oxford University.

[9] Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants: Is the melody the message? *Child Development, 60*, 1497-1510.

[10] Frick, R.W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97, 412-429.

[11] Katz, G.S., Cohn, J.F., & Moore, C.A. (1996). A Combination of vocal f_0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development*, 67, 205-217.

[12] Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (1995). *International affective picture system (IAPS): Technical Manual and Affective Ratings.* Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

[13] Larsen, R.J. & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M.S. Clark (Ed.), *Emotion. Review of Personality and Social Psychology, 13*, 25-59.

[14] Lien, J.J., Kanade, T., Cohn, J.F., & Li, C.C. (April, 1998). Automated Facial Expression Recognition. *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition (FG'98),* pp. 390-395. Nara, Japan.

[15] Mandel, D.R. & Jusczyk, P.W. (1994). Do 4-5-month-old infants know their names? *Journal of the Acoustical Society of America, 95*, 3015.

[16] Moore, C.A., Cohn, J.F., & Katz, G.S. (1994). Quantitative description and differentiation of fundamental frequency contours. *Computer Speech and Language*, 8, 385-404.

[17] Murray, I.R. & Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human emotion. *Journal of the Acoustical Society of America*, 93(2), 1097-1108.

[18] Oster, H., Hegley, D., & Nagel, L. (1992). Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. *Developmental Psychology*, 28, 1115-1131.

[19] Otsuka, T. & Ohya, J. (1998). Spotting segments displaying facial expression from image sequences using HMM. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, '98*, 442-447.

[20] Stern, D.N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinquistic infants. *Developmental Psychology, 18*, 727-735.

[21] Scherer, K.R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143-165.

[22] Zlochower, A. & Cohn, J.F. (1996). Vocal timing in face-to-face interactions of clinically depressed and nondepressed mothers and their 4-month-old infants. *Infant Behavior and Development*, 19, 373-376.