Affective Speech Design: Emotional I/O

Logan T. Hale

Abstract Emotional states are in many cases crucial to the usability, likability, and effectiveness of an interactive system. An ideal system must be able to both efficiently analyze emotional states of its users and believably convey emotions back to its users, as well as altering its behaviors to best fit the needs of its users. Noninvasive, natural means of ascertaining emotion are essential for users to accept emotion-detection in a system. Analyzing speech patterns is one such means. Speech signals can be collected using a microphone with little to no physical contact with the user and contain a plethora of information for determining emotional baselines and temporary states. A smart, adaptive system could use this emotional information to maximize positive sentiment and affect of the user during and after its use, with system manipulation and the use of emotionally expressive agents.

Keywords Adaptive systems · Affect · Emotion · Human-system interaction · Speech spectrum analysis · Speech synthesis

1 Emotion

The study of emotions and mood (affect) has increased exponentially because of advancements in technology, namely visual, auditory, and brain-imaging technologies. While emotional states are apparently subjective, at least six emotions seem to be persistent across all subsets of humanity. These six emotions are joy, anger, fear, surprise, disgust, and sadness.

Various studies have been done on emotion in speech, or emotion's alter egos/proxies—affect, sentiment, stress, valence, arousal, and so on. An important distinction is made between emotion and affect, in that emotion is temporary and

L.T. Hale (✉)
Heimstra Human Factors Laboratories, Department of Psychology, University of South Dakota, South Dakota Union, 414 East Clark Street, Vermillion, SD 57069, USA
e-mail: Logan.Hale@coyotes.usd.edu

W. Chung and C.S. Shin (eds.), *Advances in Affective and Pleasurable Design*,
Advances in Intelligent Systems and Computing 483,
DOI 10.1007/978-3-319-41661-8_36

object-oriented whereas affect is prolonged and general [1]. The emotions usually studied are the six primary emotions; however, sometimes the distinction is only dichotomous: positive and negative emotions.

Gluck et al. [2] describe emotion as a combination of three elements: physiological responses, overt behaviors, and conscious feelings. Physiological responses of emotion correspond primarily to the sympathetic and parasympathetic divisions of the autonomic nervous system. The sympathetic nervous system is closely associated with a process known as the fight-or-flight response (or sometimes the fight-flight-or-freeze response). The sympathetic nervous system causes increases in physiological systems that aid in short term decision-making and survival while causing simultaneous decreases in physiological systems that facilitate long-term decision-making and survival. Foveal vision increases at the cost of peripheral vision. Blood pressure, flow to motor muscles, and glucose level all increase while blood flow to the digestive system and other nonessential short-term organs decreases. Notably, the sympathetic nervous system also has a pronounced effect on the organs that produce and manipulate sound through the vocal pathway.

While overt behaviors such as smiling or frowning can be utilized in the context of social interactions in order to convey emotion, humans possess the capacity and drive to manipulate their overt emotional states for social gain or compliance. The same can be said for speech, as one can overtly sound happy but be covertly sad or angry. It is, however, virtually impossible to completely disguise actual felt emotions, as unconscious conveyances such as micro-expressions, micro-utterances, and other subtle cues surface.

While there is a distinction between overt and covert emotional states, there is also a distinction between conscious and unconscious states. A person can perceive himself or herself, through self-report, as experiencing one emotion while actually experiencing an entirely different, or even opposite emotion. Studies have shown that participants exposed to sad, meaningful music report sadness but low valence happiness (blithe) [3, 4]. Alternatively, though, these findings could represent the state of experiencing sadness in terms of emotion and happiness in terms of affect.

In addition, aside from the six primary emotions discussed above, humans can express various subtle, contextual, combination, and pseudo-emotions. To name a few: sarcasm, skepticism, aloofness, snootiness, impersonation, sing-songiness, condescension, confusion, excitement, boredom, engagement, disappointment, compassion, frustration, and fun. Detection and differentiation of all these emotional states and expressions is important for the future of Human Factors, as the desire for emotionally receptive and responsive systems increases.

2 Emotion in Human-System Interaction

A person's emotion can have a profound impact on his or her interaction with a system. Specifically, on attention to, memory of, performance with, and assessment of a system [1]. Contextual emotional state temporally surrounding system interaction affects user sentiment regarding the system. These interactions between human and system closely resemble interactions between human and human or between human and animal [5]; therefore, system responses to user emotional stimuli have the potential to greatly influence future interactions with and sentiments for the system.

In order for a system to interact effectively with humans in a socially ideal manner, it must be able to perceive human emotions and to convey emotions believably back to its users. There is still progress that needs to be made, however, on the measurement of emotional states in humans, if systems are ever to be able to read them accurately. Humans have always seen themselves as being able to recognize emotions within themselves with accuracy, but recent studies suggest that that might not be the case all the time [3, 4], so the validity of self-report measures is questionable and they should only be used for comparison between subjective and objective measures.

Physiological measures vary in their intrusiveness. One minimally intrusive study utilized a small bracelet to record skin conductance levels in call center workers as a proxy for stress level during calls [6]. Overly intrusive or self-report measures of emotion can themselves affect the emotional state of the user or detract from their interactive experience with the system by siphoning cognitive, sensational, and psychomotor attentional resources [7]. An example of a noninvasive system that incorporates speech, as well as facial and gestural, analysis is the AutoTutor [8]. This interactive tutor uses these three emotion-filled signals to adjust actions taken by an interactive school tutoring agent.

Progress must also be made on systems' ability to adequately convey emotions. Systems incorporating artificially intelligent agents are seen as more likeable and trustworthy when the agent conveys empathic emotion towards the user [9]. This could be, for example, a face that smiles at the user when he or she is happy or, a voice that encourages the user when he or she is frustrated. Agents that appear to have their own emotions, however, run the risk of entering the uncanny valley or increasing user expectations beyond what the system is capable of; therefore, non-human or humanoid agents can be employed to reduce these effects [5].

Chiefly in regards to Human Factors, the incorporation of accurate emotional modeling into speech recognition, modification, and synthesis would narrow the gap between human and artificial intelligence, facilitating communicative interaction in both directions. A system's ability to detect emotion in speech could allow it to track a user's change in emotional state for health reasons or to modify itself in order to better serve a user in a heightened emotional state. A user's ability to recognize emotion from a system's synthetic voice could increase the user's perceptual humanization of the system, removing the human element while maintaining the human-human level of social interaction.

3 Speech Spectrum Analysis

Briefly put, speech is a controlled expulsion of air from the lungs, through the glottis (the space between the vocal folds), and finally out the mouth and nose, which can be shaped by minute changes to this pathway into recognizable sounds for means of vocal communication.

Speech breathing has an aperiodic component and a periodic component [10]. The aperiodic component is the aeroacoustic noise from air being pushed through the vocal pathway, which can be increased or decreased by altering the amount of pressure on the lungs (subglottal pressure). To shape this speech breathing into specific sounds, laryngeal muscles open and close vocal folds in a complex manner to alter the shape of the glottis, while the vocal folds oscillate at desired frequencies. Additional shaping is done by placement of the tongue and the lips, and by redirecting some or all air flow through the nasal passage. All this shaping that takes place is the periodic component of speech.

The combination of periodic and aperiodic aspects of human speech, like any auditory signal, forms a complex pressure wave that can be visualized through an oscilloscope. This wave can be broken down into fundamental sine waves through a process known as Fourier analysis. Fourier transforms of a speech signal allow for detailed analyses of sound components in speech, such as pitch, tone, timbre, intonation, and prosody, which all vary in unique ways when the speaker is experiencing any sort of emotion [11]. Precise modification of speech signals is also possible, as evidenced by a karaoke system that alters only periodic speech signals to mimic a target singer’s voice while maintaining the aperiodic speech signature of the karaoke singer [12].

When the sympathetic nervous system is active, the diaphragm pushes more air out of the lungs during speech and other non-speech exhalations, and muscles controlling the tongue and many laryngeal muscles exhibit quicker, more forceful movements. These changes lead to a pronounced speech pattern indicative of “high valence” emotions—joy, anger, fear, surprise, and others [13]. Speech during high valence emotional states is louder, faster, and more enunciated as well as having a higher average pitch, more energy in the high-frequency range, and a broader general range of pitch [14]. These patterns in speech have been used as a proxy for sympathetic nervous system activation in many studies, while some researchers refer to it as “stress” or “arousal,” depending on the study. Speech is additionally altered depending on the specific high valence emotion. In observing only one high valence emotion such as stress, joy, or anger, overall sympathetic nervous system activation can been used as a proxy for the observed emotion; however, when differentiating between various high valence emotions, additional semantic and prosodic evidence must be acquired.

On the other side of the coin, low valence emotions—primarily sadness, but also contentment (or blithe), romantic love, and other less studied emotions—are associated with activation of the parasympathetic nervous system and its subsequent effects on the vocal pathway. These effects are a mirror image of those found with sympathetic nervous system activation: lower average pitch, less energy in the high-frequency range, and a narrower range of pitch.

Humans use these auditory spectral cues during social interactions in order to make guesses as to the emotional or affective state of others. Consequently, humans have developed the ability to mimic the spectral signature of emotional speech in order to feign emotion for survival or social status. A fascinating byproduct of this is our unique ability to create music that also mimics emotional speech, causing, in theory, a transference of that emotional state onto the listener. If conveyance of complex emotion states is possible through a completely nonverbal stimulus such as music, then it is also possible to convey emotion through modifications of sounds and artificial speech utilized in systems.

4 Emotionally Adaptive Systems

As system automation becomes more and more prominent in our society, the need for these systems to adapt to the psychological state of the user becomes more and more important. In order for a system to perform at maximum efficiency, it must recognize its users' norms on a variety of psychophysical and assumed psychological measures. Psychologists recognize the similarities among humans in regards to these measures as the environment changes, but should also recognize that every individual is affected differently. An ideal automated system should detect certain states of the user and adapt in order to maximize its usability, performance, and likeability.

With automated cars, spaceships, workstations, and so on, that users interact with every day, it is possible to establish a baseline of affective states of particular users. If the system, then, can assess the emotional state of the user at a particular point and tell that that user is experiencing a high level of stress or anger or other state relative to their norm which could impair effective use of the system, the system could then shut down or adapt to that state in order to ensure maximum safety or usage. Airplane autopilots could engage if a pilot is experiencing a panic attack or other debilitating condition. Healthcare or diet systems could adjust a user's regimen based on their emotion. Artificial intelligences could adjust their techniques of human interaction, say in computerized social communication, based on what the user would most benefit from hearing or seeing given their psychological state.

Speech is an ideal candidate for system assessment and conveyance of emotion. Recording speech from users requires only a microphone placed within their vicinity, as noise can be reduced or eradicated from the sound signal [15]. Speech signals can be analyzed almost instantaneously to assess emotional state by comparing current state to baseline rates and known specific emotional changes on the speech spectrum. Conveyance of emotional or emotionally corrective/responsive auditory information can occur instantaneously as well, only requiring the user be in the vicinity of a speaker.

Abuse of emotionally adaptive systems, however, is inevitable. Company systems could potentially coerce users into unhealthy personal choices while in negative emotional states or target happy individuals for profit. Emotion-capable artificial intelligences could manipulate users into acting outside of their comfort zone. In other words, emotionally intelligent systems could become as coercive and subversive as humans themselves currently are.

5 Conclusion

In order for emotionally adaptive systems to become relevant in human factors, much research has to be done on the development of systems for calculating baselines and variations in emotional and affective states. Research on emotional facial expressiveness is currently popular, but extensive research must be done on other non-intrusive indicators of state such as speech.

Each human's speech spectrum is unique to that individual, both in terms of his or her static voice "fingerprint" as well as the contextual changes their voice undergoes in response to variance in long-term affective and short-term emotional states. A process must be developed to extract all these signatures in speech from individuals if speech is to be utilized effectively as a means of relaying important emotional data between users and systems.

For a system to interact efficiently with human users in a socially ideal manner, it must be able to accurately determine human emotions and to express emotions convincingly back to its users. Just as a human builds a dynamic image of another human's affective equilibrium, states, and variations over time while adjusting responses and actions accordingly, systems should be created to build a mathematical model of their users' affective equilibrium, states, and variations to give their users the best possible experience with the system.

References

1. Brave, S., Nass, C.: Emotion in human–computer interaction. Hum. Comput. Interact. 53. Appendix: Springer-Author Discount (2003)
2. Gluck, M.A., Mercado, E., Myers, C.E.: Learning and memory: from brain to behavior. Palgrave Macmillan, New York (2013)
3. Kawakami, A., Furukawa, K., Katahira, K., Okanoya, K.: Sad music induces pleasant emotion. Front. Psychol. **4**(311), 1–15 (2013)
4. Vuoskoski, J.K., Eerola, T.: Can sad music really make you sad? Indirect measures of affective states induced by music and autobiographical memories. Psychol. Aesthet. Creat. Arts **6**(3), 204 (2012)
5. Kostov, V., Fukuda, S.: Emotion in user interface, voice interaction system. In: 2000 IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 798–803. IEEE (2000)

6. Hernandez, J., Morris, R.R., Picard, R.W.: Call center stress recognition with person-specific models. In: Affective Computing and Intelligent Interaction, pp. 125–134. Springer, Berlin (2011)
7. Picard, R.W.: Toward computers that recognize and respond to user emotion. IBM Syst. J **39**(3.4), 705–719 (2000)
8. D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Graesser, A., et al.: AutoTutor detects and responds to learners affective and cognitive states. In: Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems, pp. 306–308, June 2008
9. Brave, S., Nass, C., Hutchinson, K.: Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int. J. Hum. Comput. Stud. **62**(2), 161–178 (2005)
10. d'Alessandro, C., Doval, B.: Voice quality modification for emotional speech synthesis. SPECTRUM (dB) **20**, 1 (2003)
11. Cahn, J.E.: The generation of affect in synthesized speech. J. Am. Voice I/O Soc. **8**, 1–19 (1990)
12. Cano, P., Loscos, A., Bonada, J., De Boer, M., Serra, X.: Voice morphing system for impersonating in karaoke applications. In: Proceedings of ICMC2000, pp. 109–112 (2000)
13. Fulop, S.: Speech spectrum analysis. Springer Science & Business Media, Berlin (2011)
14. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
15. Vijaykumar, V.R., Vanathi, P.T., Kanagasapabathy, P.: Modified adaptive filtering algorithm for noise cancellation in speech signals. Elektronika ir elektrotechnika **74**(2), 17–20 (2015)