Sensitive Talking Heads

Thomas S. Huang, Mark A. Hasegawa-Johnson, Stephen M. Chu, Zhihong Zeng, and Hao Tang

Reprinted from IEEE Signal Processing Magazine 26(4):67-72, 2009

Spoken language user interface can dramatically speed up computer use. Unfortunately, if the speech user interface gets in the way too often, the user turns it off. Users are unforgiving: a technology that impairs productivity just once may never get a second chance.

In order to give the user interface a fighting chance, why not give it a certain amount of emotional sensitivity? Users respond better to an avatar that displays appropriate emotional nuance; conversely, if the avatar detects extreme frustration on the part of the user, it can hide in the corner of the monitor until the frustration has passed. A hidden avatar is still present, and can continue to be of service to the user, upon request.

Audio-only speech synthesis and recognition are now sufficiently accurate to be the foundation for a host of application technologies (see, e.g., the May 2008 issue of *Signal Processing Magazine*). Automatic recognition and synthesis of emotionally nuanced speech, on the other hand, are still topics of active research.

This column describes experiments in emotive spoken language user interface. We find that both recognition accuracy and synthesis quality are improved when one takes advantage of multimodal information, synthesizing and recognizing information in both the audio and video modalities.

Speech: Recognition by humans

Human speech perception is a multimodal process, in which one's interpretation of the audio signal is constrained by many types of context information. One important information source that human listeners often use to augment speech recognition is the lip motion of the speaker. Clearly, the visual channel is independent of the types of noise and reverberation suffered by the audio channel, and has the potential to serve as a complementary information source. In fact, Sumby and Pollack found that seeing the talker's face is equivalent to about 15dB improvement in the signal-to-noise ratio (SNR) of the acoustic signal [6].

Exactly how the fusion of audio and visual information takes place in the brain is not well understood. However, a number of interesting studies have shown that the human perception of bimodal speech does not require tight synchronization of the two modalities. Asynchrony arises naturally in audio-visual speech. Sound and light travel at different speeds—a 10m distance between the speaker and the listener will introduce roughly 30ms delay in the audio channel. In addition, the intensity of the stimuli is known to affect the neurological travel time. These suggest that the fusion mechanism must be relatively immune to asynchronies between the two channels. Conversely, Massaro found that humans use certain types of audio-visual asynchronies as multimodal features [4]. For example, the voice onset time (VOT), an important cue to the voicing feature in stop consonants, can be conveyed bimodally by the interval between a visible stop release and the following audible onset of vocal cord vibration. Therefore, a successful artificial audio-visual fusion scheme must be tolerant to asynchrony between the audio and visual cues, yet flexible enough to capture and exploit this bimodal cues.

Automatic audio-visual speech recognition

Because speech relevant visual information is found predominantly in the mouth region, audiovisual automatic speech recognition (AVASR) must begin with a computer vision system capable of tracking the speaker's face, and locating the mouth. Once the mouth location is given, features must be extracted from the image to summarize all speech-relevant information about the shape of the lips. Existing approaches can be grouped into two categories: *shape features* and *appearance features.*

Shape features aim to directly capture the high-level geometric information of the moving lips through straightforward measurements such as height, width, and area, or through the fitting parameters of deformable templates or active contour models (snakes). Shape features are compact, usually including fewer than 8-10 measurements per frame. However, the substantial reduction in data may also result in the loss of speech relevant information. For example, the image intensity of the mouth region may reflect aspects of the 3-D shape of the lips that are not represented in the 2-D geometric features.

Appearance features attempt to comprehensively describe the mouth region using a vector of pixel intensities, projected through a series of feature compression algorithms to reduce the vector dimension. Lowest error rates are usually achieved if the vector is compressed first using class-independent compression methods such as principal component analysis (PCA), discrete cosine transform (DCT), and discrete wavelet transform (DWT), and second using class-dependent methods such as linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT).

After visual features have been extracted, the next step in AVASR is sensory fusion. The name "sensory fusion" has been used to demarcate a general class of problems in pattern recognition; sensory fusion problems arise when multiple channels carry complementary information about the components of a system. In the case of audio-visual speech, the two modalities manifest two aspects of the same underlying speech production process. From an observer's perspective, the audio channel and the visual channel represent two interacting stochastic processes. We seek a framework that can model the two individual processes as well as their dynamic interactions.

Most practical audio-visual ASR systems aim to address the fusion problem within the hidden Markov model (HMM) framework. This can be accomplished by attaching multiple observation variables to the state variable, with each observation variable corresponds to one of the modalities, resulting in a state-synchronous multi-stream HMM (Fig. 1(a)), which is an instance of decision fusion at the state level. To better model asynchrony, a coupled HMM (CHMM) [2] or product HMM (PHMM) [5] can be used. The CHMM and PHMM can be viewed as parallel rolled-out HMM chains coupled through cross-time and cross-chain conditional probabilities, as shown in Fig. 1(b). The state of the system at a certain time slice is jointly

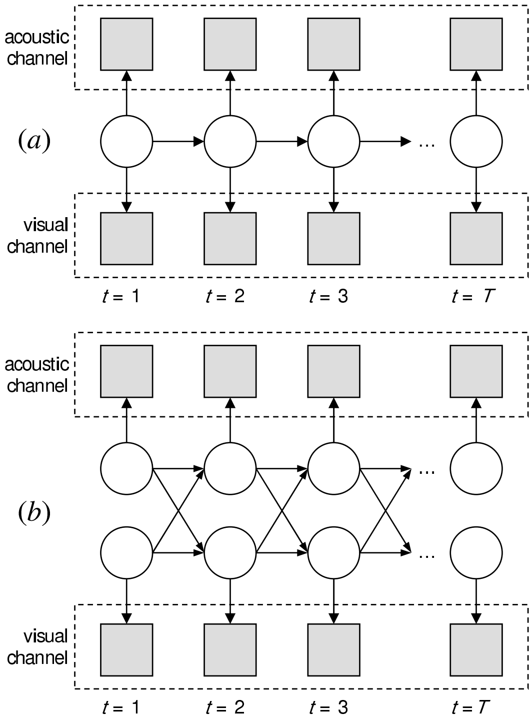

Figure 1: Fusion strategies for audiovisual automatic speech recognition: (a) hidden Markov model, (b) coupled hidden Markov model

determined by the states of these two multinomial variables. Moreover, the state of each state variable is dependent on both of its two parents in the previous time slice. This configuration essentially permits unsynchronized progression of the two chains, while encouraging the two subprocesses to assert temporal influence on each others states.

A representative of the state of the art in AVASR technology is reported in [5]. For visual frontend, the system extracts appearance features using the DCT/LDA/MLLT pipeline. For audio-visual integration, feature fusion, multi-stream HMM, and the product HMM are implemented. It was shown that on a connected digit task, the audio-visual system outperformed the audio-only system across a wide range of SNRs, with the largest gain achieved by the product HMM. For instance, at -2.2 dB SNR, the product HMM gives a 4.1% word error rate (WER), representing a 79% relative error reduction over audio-only ASR, and is 35% better than the best feature fusion method. Overall, incorporating visual information into recognition using the product HMM is equivalent to approximately 10dB gain in SNR. Similarly, on a large-vocabulary continuous speech task, the audio-visual system at 2dB SNR attains the same WER as the audio-only system at 10dB, representing an 8dB equivalent gain.

Emotion: Recognition by humans

A widespread accepted description of emotion consists of multiple components (cognitive appraisal, action tendencies, motor expression, physiological symptoms, and subjective feeling) that represent the different

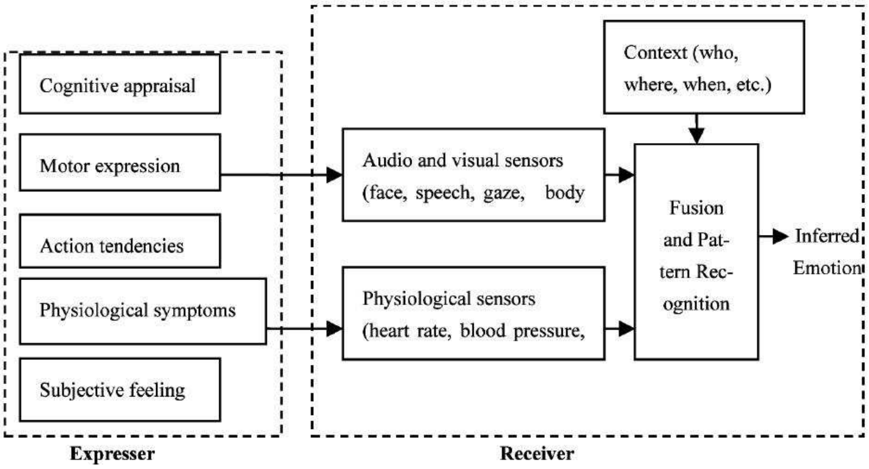
Cognitive appraisal
Motor expression
Action tendencies
Physiological symptoms
Subjective feeling
Expresser
Context (who, where, when, etc.)
Audio and visual sensors (face, speech, gaze, body
Physiological sensors (heart rate, blood pressure,
Fusion and Pattern Recognition
Inferred Emotion
Receiver

Figure 2: An emotion emission diagram from expresser to receiver

respects of emotion. Figure 2 illustrates a diagram of multimodal emotion expression and perception in which the audio and visual sensors of the receiver (e.g., human eyes and ears) capture emotional expressions (facial expressions, speech, and bodily gestures), a physiological sensor tracks the physiological responses, and a fusion and pattern recognition module (brain) integrates all these related information sources in order to label the emotion. In normal human interaction, audio and visual channels are the most important for humans to perceive emotion.

Some basic emotions (e.g., happiness and sadness) seem to be universal, and may be easily inferred from facial expressions and audio prosody; other emotions and nuances are culture-dependent, and difficult to infer without context information. Across all types of emotion (both universal and culture-dependent), humans are able to label an emotion with higher inter-transcriber agreement rates when presented with a facial expression (visual stimulus) than when presented with an audio speech stimulus. The amount of agreement drops considerably, however, when the stimuli are spontaneously displayed expressions rather than posed exaggerated displays. Many studies have theoretically and empirically demonstrated the advantage of integration of multiple modalities in human emotion perception [3].

Automatic audiovisual emotion recognition

The goal of human-centered computing is to bring the computer closer to the human, rather than vice versa. Computerized tutors should initiate new interactions when a user is bored, try to help when a user is confused, and passively guide a user who is actively engaged in a current task. Automatic speech recognition should not fail just because the user's voice is frustrated and angry; the user interface should detect the failure, diagnose its cause, and respond appropriately. Computer-assisted language learning applications should teach culturally appropriate expressions of happiness, sadness, anger, fear, and disgust; avatars in a language learning application should be capable of warning the user when her speech or facial expressions convey situationally inappropriate nuance likely to provoke an undesirable response. Some of these applications are still science fiction, but several are currently possible; for example, the Beckman

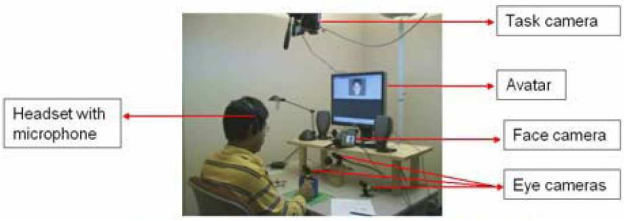

Figure 3: A prototype audiovisual computer-aided learning system

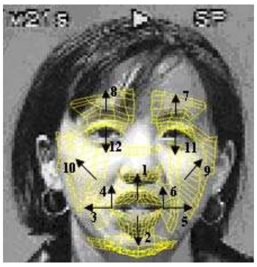

Figure 4: Facial motion units in the face tracker

Institute computerized tutor, illustrated in Fig. 3, is sensitive to displays of boredom or confusion by the user [10]. A survey of the state of the art is available in [8].

Most published systems recognize emotion using either audio or video cues, but a few recent systems integrate cues from the audio and video modalities. The system described in [9], for example, was trained based on the data of 20 subjects (10 female and 10 male). Eleven emotional states were elicited, with three sentences for each emotion; the elicited emotions were neutral, happy, sad, angry, disgusted, afraid, surprised, frustrated, interested, bored, and puzzled. A tracking algorithm (a 3D facial mesh model embedded in multiple Bezier volumes) was used to track both global motion and local deformations. Local deformations were explained in terms of twelve predefined motion units (Fig. 4), and the motion units were used for emotion recognition. Audio information was summarized in two streams: an F0 stream, and an energy stream. Five different systems were tested: two audio-only HMMs (energy and F0), a video-only HMM (whose observation vector included all twelve motion units), an independent-HMMs (IHMM) fusion strategy (modalities are independent given the emotion label), and a multi-stream fusion HMM (MFHMM), in which state residencies of the unimodal HMMs were observed by a supervisor HMM. The results of leave-one-person-out cross-validation experiments are shown in Fig. 5.

Audiovisual speech synthesis

A basic audiovisual speech synthesis system includes at least three components: text processing, audio synthesis, and visual synthesis.

Text processing is the name given to a series of steps taken to convert orthographic text into a prosodically-annotated phone sequence. A preprocessing step organizes an input sentence into a list of words and interprets numbers, abbreviations, acronyms, and other special symbols. A morphological step proposes all the possible part of speech categories for every word, and a contextual analysis step refines them according to the context of the words. A syntactic-prosodic parsing step finds the syntactic structure of the sentence, and a

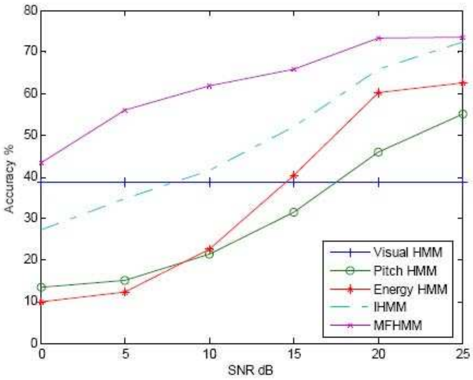

Figure 5: Accuracies of different methods under various audio SNR condition

phonetization step computes the correct pronunciation of each word using a dictionary (if the word is in the dictionary) or letter-to-sound rules (otherwise). Finally, a prosody prediction step infers the degree and type of phrasal prominence to associate with each word, as well as the locations of intonational and intermediate phrase boundaries; from these discrete prosodic labels, real-valued estimates of the target duration and fundamental frequency (F0) are computed for each phone.

The most popular *audio speech synthesis* techniques are based on concatenation of short segments (units) of recorded human speech in a database. One strategy is to record a minimum set of speech units (e.g. diphones) and modify the prosody of the selected units to match a specification through signal processing (e.g., time-domain pitch synchronous overlap-add). Another strategy is to record a large database of natural speech with its content designed to maximize the phonetic and prosodic coverage of the language. During synthesis, variable-length units are selected to minimize a joint function of the target cost and concatenation cost by means of dynamic programming.

Visual speech synthesis is often implemented by visemes and keyframe interpolation. A viseme is the visual counterpart of a phoneme that describes the facial movements of a person when he or she utters a particular sound. The mapping from phonemes to visemes is however not one-to-one. Several phonemes may share the same viseme if they are produced in a similar way. As few as 14 visemes may be enough to cover the 40 phonemes in American English.

Emotive prosody

“Prosody” is an umbrella term, covering the intonation, prominence, and rhythm of natural speech, and their acoustic correlates including F0, energy, duration, voice quality, and hyper-articulation of the phones. Prosody can communicate syntax, semantics, and para-linguistic information such as emotion. Copy-synthesis experiments have proven that highly natural emotive speech can be synthesized by a diphone synthesizer provided that accurate emotive prosody is supplied.

The prediction of emotive prosody from text can be formulated as a transformation that maps a sequence of linguistic contexts into a sequence of prosodic parameters $c_1, c_2, \dots, c_N \rightarrow p_1, p_2, \dots, p_N$ where N denotes the number of basic linguistic units (e.g., phones) in a sentence. The goal is to construct a prosody prediction

Emotion	Pitch mean	Pitch range	Speaking rate
Happy	140%	200%	150%
Sad	60%	80%	70%
Angry	160%	200%	160%
Afraid	180%	180%	160%
Disgusted	80%	80%	80%
Surprised	120%	80%	120%

Table 1: Measured percentage differences between the acoustic correlates of prosody appropriate for emotive vs. neutral speech

model $P_e(E, C)$ given the emotional state E and linguistic contexts C. This model reduces to $P_n(C)$ when predicting neutral prosody.

Statistical model-based algorithms for building $P_e(E, C)$ can be trained using a *direct approach*, i.e., using the same methods that one would use to train a neutral-speech prosody model $P_n(C)$, but based upon speech recorded for emotional state E. The direct approach suffers from an important drawback, however: available databases of emotional speech are usually quite a bit smaller than the available databases of neutral speech. A more robust system can be trained using a two-step *differential model*, $\Delta P_e(E, C)$, constructed so that $P_e(E, C) = P_n(C) + \Delta P_e(E, C)$ [7]. The differential model has several advantages over the direct approach. First, the neutral prosody model $P_n(C)$ has been extensively studied and thus it is beneficial to build the emotive prosody model $P_e(E, C)$ on top of it. Second, the differential model needs far less data to train than the direct model, as a large part of the dependency of $P_e(E, C)$ on the linguistic contexts C has been accounted for by $P_n(C)$. Third, the differential model allows us to render an emotion with a continuum of intensities, using interpolation formulae such as $P_e(E, C) = P_n(C) + \alpha \Delta P_e(E, C)$, where α is a continuous coefficient specifying the intensity of the emotion. For example, $0 < \alpha < 1$ makes the emotion less than full-blown, while $\alpha > 1$ exaggerates the emotion. $\Delta P_e(E, C)$ may be created by learning the difference between a neutral speech database and a corresponding emotive speech database, using a machine learning algorithm such as a classification and regression tree (CART).

Table 1 summarizes some of the prosodic differences between emotive speech and neutral speech.

Emotionally nuanced audiovisual speech: Emotive avatar

Visual speech synthesis without emotional nuance can be performed using viseme-based keyframe interpolation, but two problems arise when one attempts to use the same strategies to synthesize emotional speech. First, speech gestures and emotional gestures interact dynamically—static addition of the speech and emotional gestures results in highly unnatural speech. Second, the co-articulation between successive visemes is also modulated by emotion. There seems to be a lack of theoretical grounding for determining the dynamic relative contributions of speech and emotion to the positions of the lips, chin, cheeks and tongue during

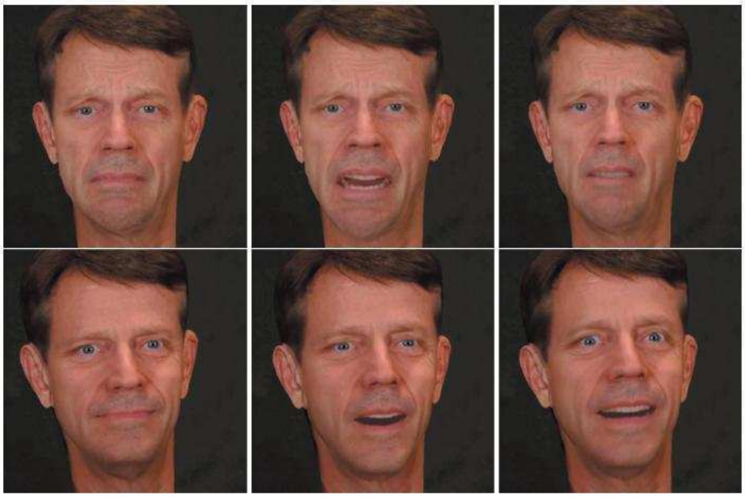

Figure 6: Synthesized visual speech snapshots with emotion variations (Sad and Happy). Figure reprinted with permission from [1].

emotive speech. However, separation of these sources is the key to the successfulness of an emotive avatar.

Heuristic approaches to the above problems include methods that treat speech gestures and facial expressions as separate parts which are combined in either a linear, piecewise linear, or nonlinear way, and methods that invent "emotive visemes" to cover all kinds of combinations of speech gestures and facial expressions despite that the number of such emotive visemes is tremendous. State-of-the-art approaches are based on facial motion capture (mocap) data and machine learning techniques. One such approach derives a generative model that incorporates facial expression control while maintaining accurate speech gestures [1]. The basic idea is to apply independent component analysis (ICA) to provide a semantic decomposition of the mocap data into a speech space and an expression space. In the speech space, a phoneme is represented by an anime $a =< P, C, M, E >$ where P is a phoneme label, C the trajectory of prosodic features, M the compressed anime motion curves, and E the emotion label. A search of the anime graph of the mocap database is performed to select appropriate anime motions M according to the anime specifications $s =< P, C, E >$. Speech gestures are synthesized by proper concatenation of the selected animes. In the expression space, a mapping $T : M_i \rightarrow M_j$ between two motions corresponding to the same speech content but different emotions is learned. This mapping takes as input the motions in one expression space and transforms them into a new expression space. The transformed motions are then combined with the speech gestures to produce the final result. An example of this approach is shown in Fig. 6.

Conclusion: Computers as actors

A professional storyteller infers emotion from the text of her tale, conveys emotion in the telling, and responds to the attachment and involvement of her audience. The characters in a video game are storytellers, but they are storytellers with a peculiar emotional handicap. Computers need not be as emotionally handicapped as they currently are: emotion recognition and emotional speech synthesis are now reliable enough to be useful in limited-domain real-world applications.

Authors

Thomas Huang is William L. Everitt Distinguished Professor of ECE and Director of the Beckman Institute Image Laboratory, University of Illinois at Urbana-Champaign. His research interests cover all aspects of image and video processing.

Mark Hasegawa-Johnson is Associate Professor of ECE, University of Illinois at Urbana-Champaign, and General Chair of the Fifth International Conference on Speech Prosody (Speech Prosody 2010). His research interests include speech production and automatic speech recognition.

Stephen M. Chu is a research staff member at the IBM T. J. Watson Research Center. His research interests include automatic speech recognition and multimedia signal processing.

Zhihong Zeng is a Research Assistant Professor at the Computer Science Department, University of Houston. His research interests include automatic face recognition, facial expression analysis, and multimodal fusion. He was a Beckman Post-Doctoral Fellow.

Hao Tang is a Ph.D. candidate of ECE and research assistant of Beckman Institute Image Laboratory, University of Illinois at Urbana-Champaign. His research interests include audiovisual speech synthesis.

References

[1] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Trans. on Graphics*, 24(4):12831302, 2005.

[2] S. M. Chu and T. S. Huang. Audio-visual speech modeling using coupled hidden markov models. In *Proc. IEEE ICASSP*, pages 2009–2012, Orlando, FL, 2002.

[3] J. A. Harrigan, R. Rosenthal, and K. R. Scherer. *The New Handbook of Methods in Nonverbal Behavior Research.* Oxford University Press, New York, USA, 2005.

[4] D. W. Massaro. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* The MIT Press, Cambridge, 1998.

[5] G. Potamianos, C. Neti, G. Gravier, A. Grag, and A. W. Senior. Recent advances in automatic recognition of audio-visual speech. *Proc. IEEE*, 91(9):1306–1326, 2003.

[6] W. H. Sumby and I. Pollak. Visual contributions to speech intelligibility in noise. *J. Acoustic Society of America*, 26:212–215, 1954.

[7] H. Tang, X. Zhou, M. Odisio, M. Hasegawa-Johnson, and T. Huang. Two-stage prosody prediction for emotional text-to-speech synthesis. In *Proc. Interspeech*, 2008.

[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1):39–58, 2009.

[9] Z. Zeng, J. Tu, B. Pianfetti, and T. S. Huang. Audio-visual affective expression recognition through multi-stream fused hmm. *Journal of Multimedia*, 10(4):570–577, 2008.

[10] T. Zhang, M. Hasegawa-Johnson, and S. E. Levinson. Cognitive state classification in a spoken tutorial dialog system. *Speech Communication*, 48(6):616–632, 2006.