INTERACTION STYLE DETECTION BASED ON FUSED CROSS-CORRELATION MODEL IN SPOKEN CONVERSATION

Wen-Li Wei, Chung-Hsien Wu, Jen-Chun Lin, and Han Li

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
Email: {lilijinjin, chunghsienwu, jenchunlin, 1988lh} @gmail.com

ABSTRACT

In recent years, much attention has been given to dialogue strategy design to achieve intelligent speech-based human-computer interaction. Since speakers generally express their intents in different Interaction Styles (ISs), the responses of a spoken dialogue system should be versatile instead of invariable and planned. This paper presents an approach to automatic detection of a user's IS using a Fused Cross-Correlation Model (FCCM). As IS generally involves high level psychological meaning, cross-correlation among various psychological factors including emotion, personality trait, and IS is thus considered for IS detection modeling. The Bayes' theorem is then used to integrate the cross-correlation into the IS detector for enhancing the IS detection accuracy. Experiments show a promising result of the proposed approach.

Index Terms— Interaction styles (ISs), emotion, personality trait

1. INTRODUCTION

Speech plays an important role in communication, perception, memory, and social contact. Understanding the latent meaning of informative speech signal for perception of the psychological meaning such as emotion, personality trait, dialogue act, or Interaction Style (IS) is beneficial to provide harmonious interactions or communication between computers and humans [1-11]. In recent years, researchers have seen increasing attention in creating a harmonious spoken dialogue system such as iPhone 4S Siri. Toward flexible and versatile responses in spoken dialogue systems, providing ability to computers with perception and recognition of a person's mental state is desirable.

Recently, Berens noted that understanding the natural IS will help us obtain satisfactory communication results in social contact [2]. Accordingly, four fundamental IS patterns including *Chart-the-Course*, *Behind-the-Scenes*, *In-Charge*, and *Get-Things-Going* were proposed and characterized by two dynamics: *Directing*/*Informing* and *Responding*/*Initiating*. Berens further demonstrates that different ISs are often reflected to different emotional tendencies and personality traits. For example, *Get-Things-Going* is often accompanied with the tendencies of sanguine emotion and extraversion personality trait. Based on these analyses, detecting a person's IS provides a new research direction for creating a harmonious spoken dialogue system.

At present, various groups of researchers have worked with human communication style analysis; each group has its own taxonomy. For example, Walker et al. introduced the Linguistic Style Improvisation (LSI) by avatars [12]. The LSI concerns the choices that the speakers make according to the semantic content, syntactic form, and acoustic realization of the spoken utterances. Pentland took steps toward quantifying social context in human communication [13], [14]. Mairesse et al. reported experimental results for recognition of the Big Five personality traits, in both conversation and text, utilizing both self and observer ratings of personality [9]. Jurafsky et al. proposed the concept of social meaning in spoken conversations [15]. The social meaning comprises three types: awkward, friendly, and flirtation. They employed regularized logistic regression to classify each social meaning using prosodic, lexical, dialogue act, and disfluency features. As the review mentioned above, major research focused on exploring the correlations between the social meaning and fundamental linguistic, prosodic or acoustic features, while lacked consideration on the mental factors.

Considering that the IS involves high level psychological meaning [2] and to narrow down the gaps between the low level features and the high level IS, intermediate psychological factors, such as emotion and personality trait are therefore considered in this study. Accordingly, this study presents a multi-modal fusion approach named Fused Cross-Correlation Model (FCCM) to detect IS in spoken conversation. The block diagram of the proposed IS detection is shown in Fig. 1. First, the prosodic information-based emotion recognition scheme is adopted to recognize a user's emotional state (i.e., containing four emotional states with high and low arousal characteristics), and the recognition result is then applied to select a better speech recognizer for later linguistic feature extraction. The

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

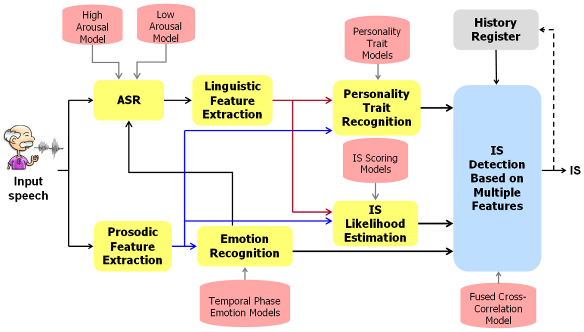
High Arousal Model
Low Arousal Model
Personality Trait Models
History Register
ASR
Linguistic Feature Extraction
Personality Trait Recognition
Input speech
IS Scoring Models
IS Detection Based on Multiple Features
IS
IS Likelihood Estimation
Prosodic Feature Extraction
Emotion Recognition
Temporal Phase Emotion Models
Fused Cross-Correlation Model

Fig. 1. The proposed framework for interaction style detection.

extracted linguistic feature is then paired with prosodic feature for personality trait recognition and IS likelihood estimation, respectively. Finally, using the estimated results of emotion, personality trait, IS likelihood, and the history of IS, the proposed FCCM is employed for final IS decision.

The rest of the paper is organized as follows. Section 2 details the derivation of the proposed fused cross-correlation model. Section 3 shows the experimental results. Section 4 offers a conclusion.

2. FUSED CROSS-CORRELATION MODEL

2.1. Model Derivation

The IS detection task considers four ISs, *Chart-the-Course*, *Behind-the-Scenes*, *In-Charge*, and *Get-Things-Going*, represented by $IS_k \in \left\{IS_c, IS_b, IS_i, IS_g\right\}$. Given the observed prosodic and linguistic features $O = \left(O^{pro}, O^{lin}\right)$, and the history of the detected interaction style IS_h, the probability of an interaction style IS_k in a current conversational turn can be estimated using (1), where IS^* represents the IS detection result by maximizing the *a posteriori* probability $P(\lambda_{IS_k} | O, IS_h)$.

$$IS^* = \underset{IS_k \in \left\{IS_c, IS_b, IS_i, IS_g\right\}}{\arg\max} \; P(\lambda_{IS_k} | O, IS_h) \tag{1}$$

In view of the IS involving psychological meaning, for IS detection, the emotion E and personality trait PT are further included. Thus, $P(\lambda_{IS_k} | O, IS_h)$ can be described as follows:

$$P(\lambda_{IS_k} | O, IS_h) = \sum_{E,PT} P(\lambda_{IS_k}, E, PT | O, IS_h) \tag{2}$$

where E denotes four emotional states, consisting of happy, neutral, angry, and sad, represented by $E \in \{H, N, A, S\}$, and PT represents extraversion and introversion, represented by $PT \in \{ET, IT\}$ in this study. For IS detection, $P(\lambda_{IS_k} | O, IS_h)$ can be further approximated by selecting an optimal state of emotion and personality trait, which maximizes $P(\lambda_{IS_k}, E, PT | O, IS_h)$ as follows:

$$P(\lambda_{IS_k} | O, IS_h) \approx \max_{E,PT} P(\lambda_{IS_k}, E, PT | O, IS_h) \tag{3}$$

The *a posteriori* probability $P(\lambda_{IS_k}, E, PT | O, IS_h)$ can be further decomposed using Bayes' rule as follows:

$$\begin{aligned} & P(\lambda_{IS_k}, E, PT | O, IS_h) \\ & = P(\lambda_{IS_k} | E, PT, O, IS_h) P(E | PT, O, IS_h) P(PT | O, IS_h) \end{aligned} \tag{4}$$

where $P(\lambda_{IS_k} | E, PT, O, IS_h)$ in (4) can be divided into four parts according to some simplification assumptions and Bayes' rule as follows:

$$\begin{aligned} P(\lambda_{IS_k} | E, PT, O, IS_h) &= \frac{P(O, E, PT, IS_h | \lambda_{IS_k}) P(\lambda_{IS_k})}{P(E, PT, O, IS_h)} \\ &\propto P(O, E, PT, IS_h | \lambda_{IS_k}) \\ &= P(O | E, PT, IS_h, \lambda_{IS_k}) P(E | PT, IS_h, \lambda_{IS_k}) \\ &\quad P(PT | IS_h, \lambda_{IS_k}) P(IS_h | \lambda_{IS_k}) \end{aligned} \tag{5}$$

Considering the limited corpus size, the probabilities are approximated as (6), by simplifying the given conditions.

$$\begin{aligned} & P(\lambda_{IS_k} | E, PT, O, IS_h) \\ & \approx P(O | \lambda_{IS_k}) P(E | \lambda_{IS_k}) P(PT | \lambda_{IS_k}) P(IS_h | \lambda_{IS_k}) \end{aligned} \tag{6}$$

where $P(O | \lambda_{IS_k})$ denotes the IS likelihood. The probabilities $P(E | \lambda_{IS_k})$, $P(PT | \lambda_{IS_k})$, and $P(IS_h | \lambda_{IS_k})$ represent the cross-correlation among emotion, personality trait, IS history and the current IS, and can be estimated based on the co-occurrence probabilities of the psychological factors. Thus, (6) can be further represented by (7).

$$\begin{aligned} & P(\lambda_{IS_k} | E, PT, O, IS_h) \\ & \approx P(O | \lambda_{IS_k}) P(E | IS_k) P(PT | IS_k) P(IS_h | IS_k) \end{aligned} \tag{7}$$

Similarly, the probabilities $P(E | PT, O, IS_h)$ and $P(PT | O, IS_h)$ in (4) can be re-written as (8) and (9), respectively.

$$P(E | PT, O, IS_h) \approx P(O | \lambda_E) P(PT | E) P(IS_h | E) P(E) \tag{8}$$

$$P(PT | O, IS_h) \approx P(O | \lambda_{PT}) P(IS_h | PT) \tag{9}$$

where $P(O | \lambda_E)$ and $P(O | \lambda_{PT})$ denote the emotion and personality trait likelihoods, respectively. The probabilities $P(PT | E)$, $P(IS_h | E)$, and $P(IS_h | PT)$ represent the cross-correlation among personality trait, emotion and IS history, and can also be estimated by the co-occurrence probability. $P(E)$ is the *a priori* probability, and can be modeled by the emotional temporal phase transition probability. Substituting (7), (8), and (9) into (4), we arrive at (10)

$$\begin{aligned} & P(\lambda_{IS_k}, E, PT | O, IS_h) \\ & = P(O | \lambda_{IS_k}) P(E | IS_k) P(PT | IS_k) P(IS_h | IS_k) P(O | \lambda_E) \\ & \quad \times P(PT | E) P(IS_h | E) P(E) P(O | \lambda_{PT}) P(IS_h | PT) \end{aligned} \tag{10}$$

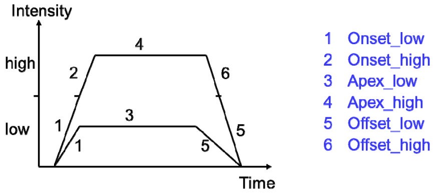

Fig. 2. An example of the emotional expression with different temporal phases.

Finally, combining (10) and (3) into (1) yields (11) for IS detection using FCCM:

$$\begin{aligned} IS^{*} = \underset{IS_k \in \{IS_c, IS_b, IS_i, IS_g\}}{\arg\max} \ \max_{E,PT} \{ & P(O \mid \lambda_{IS_k}) P(E \mid IS_k) \\ & P(PT \mid IS_k) P(IS_h \mid IS_k) P(O \mid \lambda_E) P(PT \mid E) \\ & P(IS_h \mid E) P(E) P(O \mid \lambda_{PT}) P(IS_h \mid PT) \} \end{aligned} \quad (11)$$

Clearly, the proposed FCCM not only models the low level features but also considers the cross-correlation among psychological factors by exploring their statistical dependency to enhance the IS detection accuracy.

2.2. Likelihood Estimation

For the estimation of IS likelihood, $P(O \mid \lambda_{IS_k})$ in (11) can be further approximated by the assumption of linear combination of likelihoods from the prosodic and linguistic models as follows:

$$\begin{aligned} P(O \mid \lambda_{IS_k}) &= P(O^{pro}, O^{lin} \mid \lambda_{IS_k}) \\ &\approx P(O^{pro} \mid \lambda_{IS_k}^{pro})^{\beta} P(O^{lin} \mid \lambda_{IS_k}^{lin})^{1\text{-}\beta} \end{aligned} \quad (12)$$

where $P(O^{pro} \mid \lambda_{IS_k}^{pro})$ and $P(O^{lin} \mid \lambda_{IS_k}^{lin})$ denote the likelihoods from the prosodic and linguistic models, respectively. β is a weighting factor.

Similarly, the personality trait likelihood $P(O \mid \lambda_{PT})$ in (11) can also be approximated as (13).

$$\begin{aligned} P(O \mid \lambda_{PT}) &= P(O^{pro}, O^{lin} \mid \lambda_{PT}) \\ &\approx P(O^{pro} \mid \lambda_{PT}^{pro})^{\alpha} P(O^{lin} \mid \lambda_{PT}^{lin})^{1\text{-}\alpha} \end{aligned} \quad (13)$$

where α is a weighting factor.

Since the results of Automatic Speech Recognition (ASR) on affective speech may be unsatisfactory [10], in this study, the emotion recognition task is performed first, and the recognition result is then applied to select a better speech recognizer which is trained in different affective conditions (i.e., the conditions of high and low arousal). Accordingly, for estimation of emotion likelihood, only the prosodic features are considered. In order to manage the complex temporal course of emotional expression in natural conversation, a Hidden Markov Model (HMM)-based emotion recognition scheme is constructed in terms of the temporal phases of low or high intensity of onset, apex, and offset [16-19] for audio signal stream. Fig. 2 shows an example to describe an emotional expression which can have different expressivity depending on the manner and intensity across time [17]. By integrating an emotion language model, which considers the emotional temporal phase transition probability $P(E)$, the proposed approach of the temporal course modeling can further provide a soft constraint on allowable temporal structures to determine an optimal emotional state. In this study, a bigram language model is adopted to construct the emotion language model as $P(E) = P(e_1, e_2, ..., e_M) = \prod_{k=1}^{M} P(e_k \mid e_{k\text{-}1})$, where M denotes a total number of temporal phases. Hence, $P(O \mid \lambda_E)P(E)$ in (11) is further approximated by $P(O^{pro} \mid \lambda_E)P(E)$ for emotion recognition using temporal course modeling.

3. EXPERIMENTS

3.1. Experimental Setup

This study evaluates the performance of the proposed method based on the data sets collected from Multimedia Human-Machine Communication (MHMC) Laboratory, to construct the conversation-based affective speech corpus which was provided by 53 students of both genders in the National Cheng Kung University, Taiwan. During the recording, toward naturalistic conversation, a conversation topic was first selected by each paired participants, and for each topic, the participants spoke as they like rather than navigating a pre-design script. In total, 2,120 utterances were collected in the MHMC conversation-based affective speech corpus.

For labeling the recorded data, since personality trait is reflected with an intrinsic temperament, in this study, the International Personality Item Pool (IPIP) [20] was used for measuring the personality trait for each participant. In terms of the IS and emotional expression labeling, subjective tests were performed for the recorded data. Hence, three subjects were recruited from our laboratory, and each of them was asked to give an opinion on IS and emotion labels for the recorded data. After the labeling process, each labeled utterance was then evaluated based on the opinions of all subjects. If less than two subjects reached an agreement, the utterance was not included in the experiment. Finally, a total of 1,114 data (i.e., 273 utterances for *Chart-the-Course*, 346 utterances for *Behind-the-Scenes*, 257 utterances for *In-Charge,* and 238 utterances for *Get-Things-Going*), which passed the evaluation, were regarded as the ground truth data for the experiments.

For IS and personality trait modeling, the global features of the prosody including the speech rate, duration of the pause and the minimum, mean, and maximum of the pitch, energy and formant (F1-F5) were considered, and the Support Vector Machine (SVM) [21] with radial basis kernel function was used. In terms of linguistic modeling, the Latent Semantic Analysis (LSA) was applied to extract the meaningful linguistic features for modeling the IS and personality trait using SVM. For emotion modeling, the left-to-right topology of the HMM structure with three hidden states was applied for modeling each temporal phase of emotional expression, and the local features [22] of the pitch, energy and formants F1-F5 in each speech frame were used. A bigram language model was adopted to construct the emotion language model which estimates the temporal phase transition probability. The speech recognizer was constructed by tri-phone models with the characteristics of the high and low arousal of affective conditions through HTK toolkit [23]. For each speech frame, 12 Mel Frequency Cepstrum Coefficients (MFCCs) plus a normalized energy parameter, and the first and second derivatives were estimated to form a 39-D feature vector.

3.2. Experimental Results

For performance evaluation, 80% of the ground truth utterances were randomly selected from the MHMC conversation-based affective speech corpus for training, and the remaining utterances were selected for testing. In the experiments, compared with the traditional (universal) speech recognizer, the proposed speech recognizer considering the affective conditions of low and high arousal characteristics achieved 6.12% word accuracy improvement (i.e., traditional speech recognizer achieved 68.12% word recognition accuracy, and the proposed achieved 74.24%). The result confirmed that affective factor has significant effect on the recognition result of the traditional speech recognizer. For the average emotion recognition accuracy, compared with the traditional HMM recognition scheme (eight hidden states were used) [24], the proposed temporal course modeling method achieved 23.32% improvement for four emotional states; the proposed method achieved the 79.82% average emotion recognition accuracy. The result demonstrates that considering the complex temporal structure of emotional expression in natural conversation is useful. Thus, the proposed method can provide a better ability for providing emotion information to support IS detection. In addition, for low and high arousal categories, the proposed method achieved 89.69% average emotion recognition accuracy. Accordingly, the proposed emotion recognition scheme can provide the ability for selecting an appropriate speech recognizer. In terms of the personality trait recognition, the average recognition accuracy of the extraversion and introversion achieved 91.03% when the weighting factor α was set to 0.6.

Based on the sub-components recognition, we first evaluated the contributions of each sub-component for detecting the IS using the proposed FCCM as shown in Fig. 3. According to the observation, compared with that using only low level features (i.e., IS likelihood only), the result shows that the psychological factors, emotion and personality trait, are helpful for improving the IS detection accuracy. The result strengthened the analyses of the mentioned research; that is, different ISs are often reflected to the different emotional tendencies and personalities. For the estimation of IS likelihood, in this study, the weighting factor β was set to 0.75. In addition, to further demonstrate the efficiency of the proposed FCCM, we also compared the FCCM to the score level fusion approach. For score level fusion [25], in this study, the feature vectors were obtained by concatenating the recognition likelihoods resulted from emotion, personality trait, and IS detection models. The obtained feature vectors were then used for IS detection using an SVM classifier (radial basis kernel function was used). The average detection accuracy for two approaches is shown in Table 1. The results in Table 1 indicate that the FCCM achieved a better performance than that of the score level fusion approach. These findings show that considering the cross-correlation among various psychological factors in the proposed FCCM is of great help for IS detection.

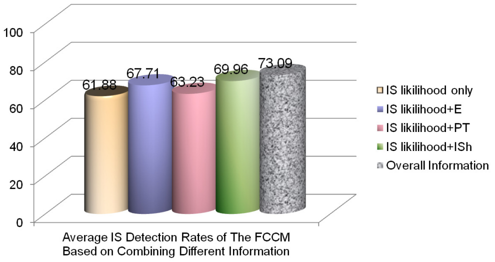

Fig. 3. Average IS detection rates by combining different information using the proposed FCCM.

TABLE 1
AVERAGE IS DETECTION RATES (%) FOR DIFFERENT MODELS

	Score level fusion	FCCM
Average detection rates	69.06	73.09

4. CONCLUSION

This paper presented a novel statistical model to detect the interaction style under spoken conversation. Two findings are summarized from our experiments. First, modeling the complex temporal structure of emotional expression is helpful for improving the emotion recognition accuracy in natural conversation. Second, modeling the relationships among psychological factors to different interaction styles in the proposed FCCM is useful for enhancing the detection accuracy of the interaction style. Compared to the current fusion approach, experimental results demonstrate the effectiveness of the proposed FCCM. Since interaction style contains rich semantic meaning, future research to effectively judge the grammatical and semantic features such as imperative sentences and affirmatives would be beneficial for interaction style detection.

5. REFERENCES

[1] R.W. Picard, *Affective Computing*. MIT Press, 1997.

[2] L.V. Berens, *Understanding Yourself and Others: An Introduction to Interaction Styles*. Telos Publications, 2008.

[3] W.B. Liang, C.H. Wu, C.H. Wang, and J.F. Wang, "Interactional style detection for versatile dialogue response using prosodic and semantic features," in *Proc. INTERSPEECH*, pp. 1345–1348, 2011.

[4] C.H. Wu and W.B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affective Computing*, vol. 2, no.1, pp. 1–12, 2011.

[5] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[7] C.P. Chen, C.H. Wu, and W.B. Liang, "Robust dialogue act detection based on partial sentence tree, derivation rule, and spectral clustering algorithm," *EURASIP Journal on Audio, Speech, and Music Processing,* 2012:13, 2012.

[8] C.H. Wu and G.L. Yan, "Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system," *IEEE Trans. Speech and Audio Processing*, vol.13, no.3, pp.330-344, 2005.

[9] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[10] J.C. Lin, C.H. Wu, and W.L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no.1, pp. 142–156, Feb. 2012.

[11] C.H. Wu, Z.J. Chuang, and Y.C. Lin, "Emotion recognition from text using semantic label and separable mixture model," *ACM Trans. on Asian Language Information Processing*, vol. 5, no. 2, pp. 165–182, Jun. 2006.

[12] M.A. Walker, J.E. Cahn, and S.J. Whittaker, "Improvising linguistic style: social and effective bases for agent," in *Proc. International Conference on Autonomous Agents*, pp. 96–105, 1997.

[13] A. Pentland, "Socially aware computation and communication," *IEEE Computer Society*, vol. 38, no. 3, pp. 63–70, Mar. 2005.

[14] A. Pentland, "Social dynamics signals and behavior," *ICDL*, San Diego, 2004.

[15] D. Jurafsky, R. Ranganath, and D. Macfarland, "Extracting social meaning: identifying interactional style in spoken conversation," in *Proc. NAACL HLT*, pp. 638–646, 2009.

[16] P. Ekman, *Handbook of Cognition and Emotion*. Wiley, 1999.

[17] N. Mana and F. Pianesi, "Modeling of emotional facial expressions during speech in synthetic talking heads using a hybrid approach," *Int'l Conf. Auditory-Visual Speech Processing (AVSP)*, 2007.

[18] M.F. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," *Proc. Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, 2006.

[19] M.F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Systems, Man and Cybernetics–Part B*, vol. 42, no.1, pp. 28–43, 2012.

[20] L.R. Goldberg, J.A. Johnson, H.W. Eber, R. Hogan, M.C. Ashton, C.R. Cloninger, and H.G. Gough, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in Personality*, vol. 40, no.1, pp. 84–96, 2006.

[21] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[22] M.E. Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification, schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[23] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book* (Version 3.4). Cambridge University Press, 2006.

[24] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *Proc. 28th Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP' 03)*, pp. II 1–4, 2003.

[25] Y. Wang, L. Guan, and A.N. Venetsanopoulos, "Kernel cross-model factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia,* vol. 14, no.3, pp. 597–607, 2012.