Emotion Recognition of Conversational Affective Speech Using Temporal Course Modeling-Based Error Weighted Cross-Correlation Model

Jen-Chun Lin*, Wen-Li Wei†, Chung-Hsien Wu†, and Hsin-Min Wang*
*Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.
E-mail: jenchunlin@gmail.com; whm@iis.sinica.edu.tw
†Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.
E-mail: chunghsienwu@gmail.com; lilijinjin@gmail.com

***Abstract*— A complete emotional expression in natural face-to-face conversation typically contains a complex temporal course. In this paper, we propose a temporal course modeling-based error weighted cross-correlation model (TCM-EWCCM) for speech emotion recognition. In TCM-EWCCM, a TCM-based cross-correlation model (CCM) is first used to not only model the temporal evolution of the extracted acoustic and prosodic features individually but also construct the statistical dependencies among paired acoustic-prosodic features in different emotional states. Then, a Bayesian classifier weighting scheme named error weighted classifier combination is adopted to explore the contributions of the individual TCM-based CCM classifiers for different acoustic-prosodic feature pairs to enhance the speech emotion recognition accuracy. The results of experiments on the NCKU-CASC corpus demonstrate that modeling the complex temporal structure and considering the statistical dependencies as well as contributions among paired features in natural conversation speech can indeed improve the speech emotion recognition performance.**

I. INTRODUCTION

Emotions play an important role in human intelligence, rational decision making, social interaction, perception, memory and more [1]. Since perception and experience of emotion is vital for communication in the social environment, understanding emotions becomes indispensable for the day-to-day functioning of humans. Technologies for processing daily activities including speech, language and facial expression have expanded the interaction modalities between humans and computer-supported communicational artifacts, such as robots, iPad, and mobile phones. With the growing and varied uses of human-computer interactions, emotion recognition technologies provide an opportunity to promote harmonious interactions or communication between computers and humans [2], [3]. Hence, constructing a high-performance emotion perception and recognition system from speech signal is highly desirable.

Although various studies in emotion recognition from speech have shown the benefits using different features and classifiers [4], [5], to recognize emotion that occurs during conversation with high accuracy, the dynamic aspects of emotional expression available for model training and recognition are critical. In general, when the temporal course of emotional expression is complex, the temporal information could be lost owing to inappropriate model structures; therefore, the classification result will be unsatisfactory. Previous research [1], [6]-[10] has demonstrated that a complete emotional expression can be divided into three sequential temporal phases, namely, onset (application), apex (release), and offset (relaxation), considering the manner and intensity of an expression. In most hidden Markov model (HMM)-based emotion recognition schemes, a single left-to-right HMM was used to model a specific emotional state [11]-[14] and was demonstrated to facilitate the modeling of a signal stream (i.e., audio or visual) to describe the temporal courses of complete emotional expressions. However, a single left-to-right HMM may be invalid for recognizing emotions of individual utterances in a natural conversation. As shown in Fig. 1, when the emotional state (i.e., happiness) of Speaker 1 was evoked in the conversation, her Utterance 1 covered only the onset temporal phase with low intensity, and her Utterance 2 covered the remaining apex and offset with low intensity phases.

To handle the aforementioned problem, a temporal course modeling (TCM) method [8]-[10] was recently proposed to characterize the evolution of temporal expression involved in an emotional state that occurs in an isolated utterance. In the TCM method, the emotional sub-states were defined to represent the temporal phases (i.e., the onset, apex, or offset with low or high intensity) of an emotional expression, and an HMM was used to characterize a specific single emotional sub-state, rather than the entire emotional state. The emotional sub-state language model was then constructed to provide a constraint on allowable temporal structures (i.e., the transition between emotional sub-states) to determine an optimal emotional state for an isolated utterance. Although experimental results have demonstrated that the TCM method could substantially improve the apparent performance of speech emotion recognition in conversational environment, it did not explore the contributions of multiple features and consider the correlations between features for recognition, which have recently demonstrated greatly helpful for audio-

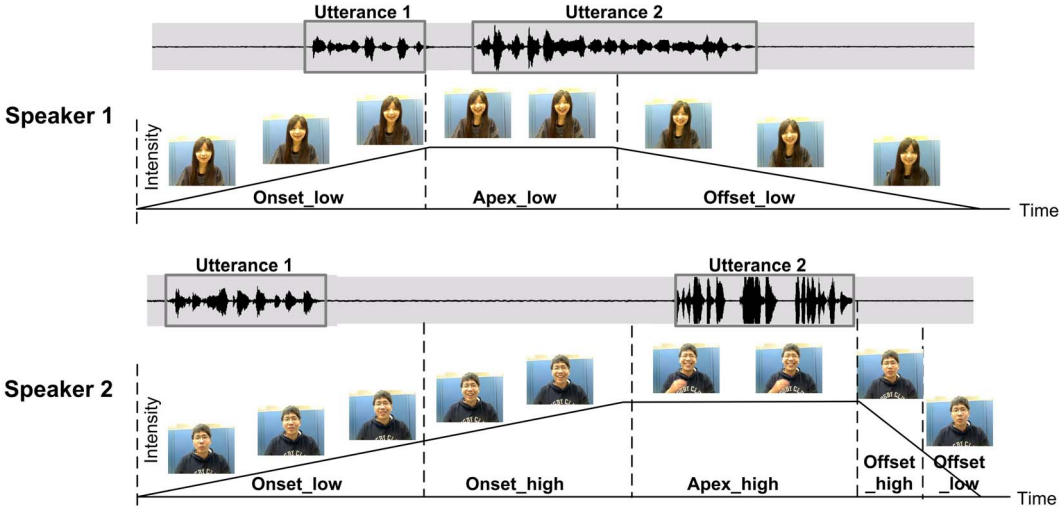

Fig. 1. An example of various temporal phases of happy emotional expression for the utterances in a real conversational environment [10].

visual emotion recognition and facial action unit (AU) or emotion recognition [13], [15]-[18].

To enhance recognition performance, in this study, an extended model named temporal course modeling-based error weighted cross-correlation model (TCM-EWCCM) is proposed. Different from the TCM method, TCM-EWCCM further considers the correlations between acoustic and prosodic features and explores the contributions of paired acoustic-prosodic features to enhance the speech emotion recognition accuracy. The architecture of the TCM-EWCCM method is shown in Fig. 2. A TCM-based cross-correlation model (CCM) is first used to not only model the temporal evolution of the extracted acoustic and prosodic features individually but also construct the statistical dependencies among paired acoustic-prosodic features in different emotional states for recognition. Then, a Bayesian classifier weighting scheme named error weighted classifier combination [19]-[21] is adopted to explore the contributions of the individual TCM-based CCM classifiers for different acoustic-prosodic feature pairs to make the final recognition decision.

The remainder of this paper is organized as follows. Section II briefly outlines the speech feature extraction part. Section III details the derivation of the proposed temporal course modeling-based error weighted cross-correlation model. Section IV shows the experimental results. Section V offers a conclusion.

II. Speech Feature Extraction

An important issue in speech emotion recognition is the extraction of features that can efficiently characterize the emotional content of speech. Accordingly, several speech features, such as the prosodic features, acoustic features, and voice quality features, have been discussed over the years [4], [5]. Among them, the prosodic and acoustic features have been proven useful for characterizing the emotional content and widely used in speech emotion recognition [8], [10], [22]-[26]. In this study, from each speech frame, 12-dimensional mel-frequency cepstrum coefficients (MFCCs) were extracted as the acoustic features, while the pitch, energy, and formants F1-F5 were extracted as the three types of primary prosodic features. Praat [27] and the hidden Markov model toolkit (HTK) [28] were used for prosodic and acoustic feature extraction, respectively.

III. Temporal Course Modeling-Based Error Weighted Cross-Correlation Model

A. Model Derivation

In this study, the goal is to classify a paired acoustic and prosodic observation (O^a, O^p) into an emotional class w out of K classes by combining the decisions determined by the classifiers from C types of acoustic features and D types of prosodic features. Two sets of observations, $O^a = \{O^a_{i=1}, O^a_{i=2} ... O^a_{i=C}\}$ and $O^p = \{O^p_{j=1}, O^p_{j=2} ... O^p_{j=D}\}$ represent multiple independent acoustic and prosodic feature channels, respectively, where $O^a_i = o^a_{i,f=1}, o^a_{i,f=2}, ..., o^a_{i,f=F}$ and $O^p_j = o^p_{j,f=1}, o^p_{j,f=2}, ..., o^p_{j,f=F}$ are the sequences of the i-th and j-th types of features of O^a and O^p, respectively. The probability of w given (O^a, O^p) can be computed by

$$\begin{aligned} P(w \mid O^a, O^p) &= \sum_{i=1}^{C} \sum_{j=1}^{D} P(w, \Lambda^a_i, \Lambda^p_j \mid O^a, O^p) \\ &= \sum_{i=1}^{C} \sum_{j=1}^{D} P(w \mid O^a, O^p, \Lambda^a_i, \Lambda^p_j) P(\Lambda^a_i, \Lambda^p_j \mid O^a, O^p), \end{aligned} \tag{1}$$

where $P(w \mid O^a, O^p, \Lambda^a_i, \Lambda^p_j)$ is the probability of w given by the TCM-based CCM classifier $(\Lambda^a_i, \Lambda^p_j)$ with the paired input of (O^a_i, O^p_j), and $P(\Lambda^a_i, \Lambda^p_j \mid O^a, O^p)$ is the prior weight assigned to the classifier, representing the confidence of the decision of the classifier [21]. This weight can be

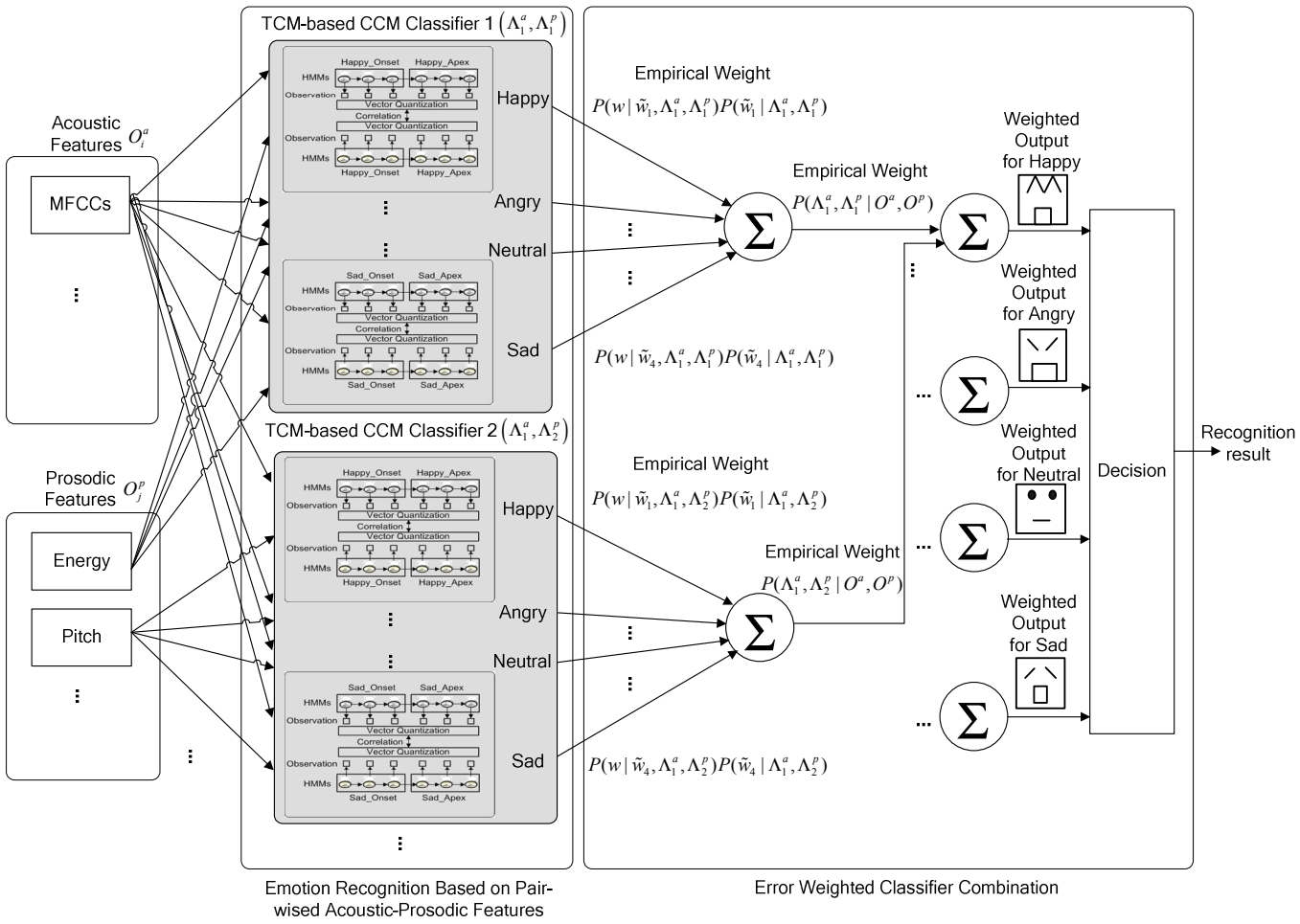

Fig. 2. The architecture of the proposed temporal course modeling-based error weighted cross-correlation model. This figure illustrates the example for recognizing emotional state from the paired acoustic-prosodic features.

computed from the confusion matrix of classifiers $(\Lambda_i^a, \Lambda_j^p)$, i=1,…,C, j=1,…,D.

Since recognition of emotional class w can be made based on the combination of the outputs of all recognized emotional classes $\tilde{w}_k$, k=1,...,K, from individual TCM-based CCM classifiers, $P(w|O^a, O^p, \Lambda_i^a, \Lambda_j^p)$ can be further represented as

$$\begin{aligned} P(w|O^a, O^p, \Lambda_i^a, \Lambda_j^p) &= \sum_{k=1}^{K} P(w, \tilde{w}_k|O^a, O^p, \Lambda_i^a, \Lambda_j^p) \\ &= \sum_{k=1}^{K} P(w|O^a, O^p, \Lambda_i^a, \Lambda_j^p, \tilde{w}_k) P(\tilde{w}_k|O^a, O^p, \Lambda_i^a, \Lambda_j^p), \end{aligned} \tag{2}$$

where the conditional error distribution $P(w|O^a, O^p, \Lambda_i^a, \Lambda_j^p, \tilde{w}_k)$ can be approximated by its projection $P(w|\tilde{w}_k, \Lambda_i^a, \Lambda_j^p)$, which can be simply obtained from the confusion matrix of the corresponding classifier, and can be regarded as an empirical weight [21]. Therefore, (2) can be approximated as

$$\begin{aligned} &P(w|O^a, O^p, \Lambda_i^a, \Lambda_j^p) \\ &\quad \approx \sum_{k=1}^{K} P(w|\tilde{w}_k, \Lambda_i^a, \Lambda_j^p) P(\tilde{w}_k|O^a, O^p, \Lambda_i^a, \Lambda_j^p). \end{aligned} \tag{3}$$

By substituting (3) into (1), we arrive at (4)

$$\begin{aligned} P(w|O^a, O^p) \approx \sum_{i=1}^{C}\sum_{j=1}^{D}\sum_{k=1}^{K} & P(w|\tilde{w}_k, \Lambda_i^a, \Lambda_j^p) \\ & P(\tilde{w}_k|O^a, O^p, \Lambda_i^a, \Lambda_j^p) P(\Lambda_i^a, \Lambda_j^p|O^a, O^p). \end{aligned} \tag{4}$$

By using the Bayes rule, the probability $P(\tilde{w}_k|O^a, O^p, \Lambda_i^a, \Lambda_j^p)$ can be further decomposed as

$$\begin{aligned} &P(\tilde{w}_k|O^a, O^p, \Lambda_i^a, \Lambda_j^p) \\ &\quad = \frac{P(O^a, O^p|\Lambda_i^a, \Lambda_j^p, \tilde{w}_k) P(\tilde{w}_k|\Lambda_i^a, \Lambda_j^p)}{P(O^a, O^p|\Lambda_i^a, \Lambda_j^p)} \\ &\quad \propto P(O^a, O^p|\Lambda_i^a, \Lambda_j^p, \tilde{w}_k) P(\tilde{w}_k|\Lambda_i^a, \Lambda_j^p), \end{aligned} \tag{5}$$

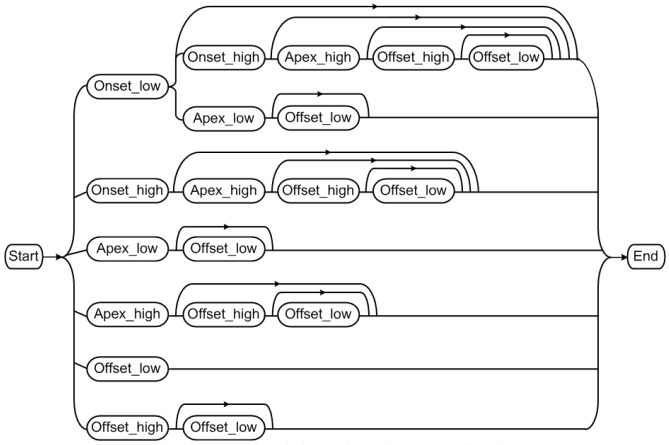

Fig. 3. Recognition network based on the predefined grammar for characterizing an emotional state expressed in an isolated utterance [8], [9].

where the probability $P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p)$ is identical for all possible emotional classes $\tilde{w}_k$, and thus can be omitted. $P(\tilde{w}_k \mid \Lambda_i^a, \Lambda_j^p)$ denotes the probability of the emotional class $\tilde{w}_k$ given the classifier $(\Lambda_i^a, \Lambda_j^p)$. This probability can also be approximated from the confusion matrix of $(\Lambda_i^a, \Lambda_j^p)$ as an empirical weight [21].

As shown in Fig. 3 [8], [9], $\Lambda_i^a = \lambda_{i,m=1}^a, \lambda_{i,m=2}^a, ..., \lambda_{i,m=M}^a$ and $\Lambda_j^p = \lambda_{j,m=1}^p, \lambda_{j,m=2}^p, ..., \lambda_{j,m=M}^p$, where M ranges from 1 to 5 considering all possible emotional sub-state (i.e., temporal phase) transitions in a predefined grammar, denote the acoustic and prosodic emotional sub-state HMM sequences for O_i^a and O_j^p in the TCM-based CCM classifier $(\Lambda_i^a, \Lambda_j^p)$. In order to provide a constraint on allowable temporal structures, the condition of temporal phase (i.e., emotional sub-state) sequence T is introduced in $P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k)$ as follows,

$$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k) = \sum_T P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T) P(T). \quad (6)$$

Then, by selecting the best temporal phase sequence that maximizes $P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T)P(T)$, the probability $P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k)$ can be further approximated as

$$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k) \approx \max_T P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T) P(T), \quad (7)$$

where $P(T) = P(t_1, t_2, ..., t_M)$ is the a *priori* probability of observing temporal phase sequence T, which can be estimated using the emotional sub-state language model [8], [9]. In this study, we use a bigram language model as follows

$$P(T) = P(t_1, t_2, ..., t_M) = P(t_1) \prod_{k=2}^{M} P(t_k \mid t_{k-1}). \quad (8)$$

$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T)$ is approximated by considering the co-occurrence dependencies between features which not only model the temporal evolution of the extracted acoustic and prosodic features individually but also construct the statistical dependencies among paired acoustic-prosodic features in different emotional states for recognition. Based on this assumption, the probability $P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T)$ in TCM-based CCM is approximated by

$$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T) \approx P(O^a, O^p \mid \Lambda_i^a, \tilde{w}_k, T) P(O^a, O^p \mid \Lambda_j^p, \tilde{w}_k, T), \quad (9)$$

where $P(O^a, O^p \mid \Lambda_i^a, \tilde{w}_k, T)$ can be divided into two parts using the chain rule, i.e., $P(O^a \mid \Lambda_i^a, \tilde{w}_k, T)$ and $P(O^p \mid O^a, \Lambda_i^a, \tilde{w}_k, T)$. $P(O^a \mid \Lambda_i^a, \tilde{w}_k, T)$ is the acoustic model likelihood of the *i*-th type of acoustic features given the corresponding emotional sub-state HMM sequence Λ_i^a for the specific emotional class $\tilde{w}_k$ with temporal phase sequence *T*. $P(O^p \mid O^a, \Lambda_i^a, \tilde{w}_k, T)$ is the probability of the co-occurrence dependency between the *i*-th type of acoustic features and the *j*-th type of prosodic features. $P(O^a, O^p \mid \Lambda_j^p, \tilde{w}_k, T)$ can also be derived in a similar way. Therefore, (9) can be re-written as

$$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T) \approx P(O^a \mid \Lambda_i^a, \tilde{w}_k, T) P(O^p \mid O^a, \Lambda_i^a, \tilde{w}_k, T) \times P(O^a \mid O^p, \Lambda_j^p, \tilde{w}_k, T) P(O^p \mid \Lambda_j^p, \tilde{w}_k, T). \quad (10)$$

However, the co-occurrence dependencies $P(O^p \mid O^a, \Lambda_i^a, \tilde{w}_k, T)$ and $P(O^a \mid O^p, \Lambda_j^p, \tilde{w}_k, T)$ are difficult to obtain, because the observations are continuous values, and there is no sufficient training data to construct the statistical dependencies between the acoustic and prosodic features under various joint conditions associated with T and $\tilde{w}_k$. In this study, we simplify the joint conditions for estimation of the co-occurrence dependencies by only considering the emotional state $\tilde{w}_k$, i.e., we only estimate $P(O^p \mid O^a, \tilde{w}_k)$ and $P(O^a \mid O^p, \tilde{w}_k)$. In addition, a vector quantization (VQ) technique is employed to discretize the observations O^a and O^p into discrete codewords denoted as β^a and α^p, respectively. Therefore, (10) can be re-written as

$$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T) \approx P(O^a \mid \Lambda_i^a, \tilde{w}_k, T) P(\alpha^p \mid \beta^a, \tilde{w}_k) \times P(\beta^a \mid \alpha^p, \tilde{w}_k) P(O^p \mid \Lambda_j^p, \tilde{w}_k, T). \quad (11)$$

Because each type of acoustic or prosodic features may include a different number of feature dimensions (e.g., 12-dimensional MFCCs and 5-dimensional formants are considered in this study), we extract the so-called global features [5] from each dimension of the acoustic and prosodic feature sequence extracted from an utterance for VQ. The global features include the mean, maximum, minimum, standard deviation, and ratio of the numbers of rising and descending feature types (i.e., calculated by the statistic of the difference of feature values of every two consecutive speech frames). The k-means clustering algorithm is employed for codebook generation, and the number of clusters is determined based on the analysis of variance (ANOVA) test. Therefore, (11) can be re-written as.

$$P(O^a, O^p \mid \Lambda_i^a, \Lambda_j^p, \tilde{w}_k, T) \approx P(O^a \mid \Lambda_i^a, \tilde{w}_k, T) \prod_{l=1}^{L} \prod_{q=1}^{Q} \left(P(\alpha_l^p \mid \beta_q^a, \tilde{w}_k) \times P(\beta_q^a \mid \alpha_l^p, \tilde{w}_k) \right) P(O^p \mid \Lambda_j^p, \tilde{w}_k, T), \quad (12)$$

where β_q^a and α_l^p denote the codewords of the q-th and the l-th dimensions for the i-th acoustic and the j-th prosodic features, respectively. Finally, combining (12), (7), and (5) into (4) yields (13) for emotional class w recognition using TCM-EWCCM:

$$P(w \mid O^a, O^p) \approx \sum_{i=1}^{C} \sum_{j=1}^{D} \left\{ \sum_{k=1}^{K} P(w \mid \tilde{w}_k, \Lambda_i^a, \Lambda_j^p) \max_T \left[P(O^a \mid \Lambda_i^a, \tilde{w}_k, T) \prod_{l=1}^{L} \prod_{q=1}^{Q} \left(P(\alpha_l^p \mid \beta_q^a, \tilde{w}_k) P(\beta_q^a \mid \alpha_l^p, \tilde{w}_k) \right) P(O^p \mid \Lambda_j^p, \tilde{w}_k, T) P(T) \right] P(\tilde{w}_k \mid \Lambda_i^a, \Lambda_j^p) \right\} P(\Lambda_i^a, \Lambda_j^p \mid O^a, O^p). \quad (13)$$

In this study, one type of acoustic features (C=1) and three type of prosodic features (D=3) are considered for recognizing four emotional classes (K=4), including happy, angry, neutral, and sad. Therefore, the TCM-EWCCM classifier consists of 3 TCM-based CCM classifiers, namely $(\Lambda_1^a, \Lambda_1^p)$, $(\Lambda_1^a, \Lambda_2^p)$, and $(\Lambda_1^a, \Lambda_3^p)$.

IV. Experimental Results

A. Corpus Descriptions and Experimental Setup

TABLE I
THE NUMBERS OF GROUND TRUTH UTTERANCES OF FOUR EMOTIONAL STATES USED IN THE EXPERIMENTS.

	Happy	Angry	Sad	Neutral
# (Utterances)	199	236	214	465

The National Cheng Kung University Conversation-based Affective Speech Corpus (NCKU-CASC) [10] was used in the speech emotion recognition experiments. The speech data were provided by 53 students, including 46 males (87%) and 7 females (13%). To facilitate fluent and natural conversation during recording, each paired participants were allowed to select a conversation topic and spend approximately 10 minutes to discuss the selected conversation topic before recording, instead of using a pre-designed script. In this way, the dialogue turns were recorded according to the contents discussed by the paired participants. A total of 2,120 utterances were collected.

Then, to evaluate the correctness of emotional expression of the recorded data, subjective tests were performed [10]. The kappa statistic was applied to evaluate the inter-rater agreement in labeling the utterances. The values of the kappa statistic κ for the emotional states (i.e., the categorical emotion labels) and the temporal course labels were 0.395 and 0.360, respectively, which indicated only a fair agreement among the raters [29], [30]. Therefore, we had to select the representative utterances based on majority voting as the ground truth data to be used in the experiments. As a result, a total of 1,114 utterances with higher inter-rater agreement were selected. The numbers of ground truth utterances of four emotional states are summarized in Table I.

Three methods were compared in the experiments, namely, the conventional HMM method, the TCM method [8], [10], and the proposed TCM-EWCCM method. For the conventional HMM method, a left-to-right HMM with eight hidden states was used because this setting outperformed other settings in our preliminary experiments. For the TCM and TCM-EWCCM methods, a left-to-right HMM with three hidden states was used to model each temporal phase. We performed 5-fold cross-validation in the experiments. For each fold, 80% of the ground truth utterances were used for training, while the remaining 20% utterances were used for testing.

B. Experimental Results

To demonstrate whether combining acoustic and prosodic features is useful for speech emotion recognition, the first experiment is conducted using the following two feature sets: (i) the prosodic features only (Pro only), and (ii) direct concatenating prosodic and acoustic features (Pro+Aco). The conventional HMM and TCM methods are evaluated. The results in the average recognition accuracy are shown in Table II. From the table, it is clear that combining the acoustic and prosodic features (either for the conventional HMM or the TCM method) is helpful for improving the performance of speech emotion recognition. The results inspire us thinking how to more effectively combine acoustic and prosodic features to enhance the recognition performance. Besides, the

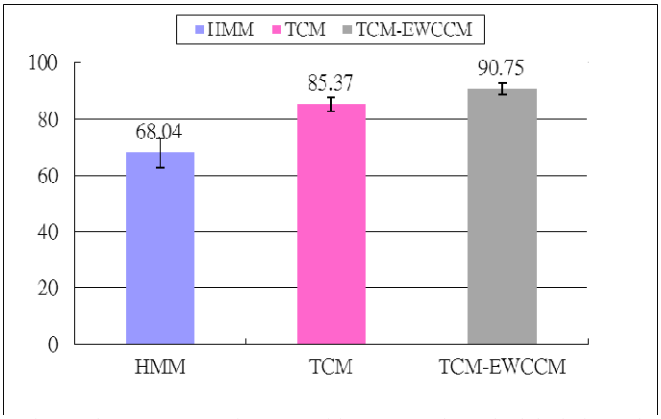

Fig. 4. The average emotion recognition rates and standard deviations of three methods in the 5-fold cross-validation experiments.

result in Table II also demonstrates that the TCM method obviously outperforms the conventional HMM method. The result confirms that considering the temporal phases and the emotional sub-state language model can better describe the complex temporal structure of emotional expression in natural conversation. The finding also indicates that the complex temporal structure of emotional expression has significant negative impact on the recognition result of the conventional HMM method.

As mentioned above, TCM-EWCCM proposed in this paper not only models the complex temporal structure of emotional expressions but also explores the statistical dependencies as well as contributions among paired acoustic-prosodic features for recognition. The average recognition rates and standard deviations of all the three methods compared in this paper are shown in Fig. 4. Among them, TCM-EWCCM achieved the best recognition accuracy as shown in Fig. 4. The Chi-squared test shows that the performance differences among the three methods are statistically significant with $\chi^2(2) = 207.133, P < .0001$. In addition, using post-hoc pairwise comparisons for Chi-squared test, the crosstab (also known as Contingency Table) based on the z-test with Bonferroni corrections is shown in Table III [31], [32]. The results show that the correct recognition results of the proposed method (C) are statistically more significant than those of the other models (i.e., A and B). The statistical analysis verified the effectiveness of the proposed method.

V. Conclusions

This paper has presented a novel temporal course modeling-based error weighted cross-correlation model (TCM-EWCCM) for recognizing the emotional state in a conversational speech signal. Three findings are summarized from our experiments. First, acoustic and prosodic features are complementary to each other in speech emotion recognition, and the integration of both types of features indeed improves the recognition accuracy. Second, modeling the complex temporal structure in the TCM method is greatly helpful for improving the speech emotion recognition accuracy. The conventional HMM method lacks the ability to describe the complex temporal structure of emotional expression in utterances under natural conversation. Third, the experimental results demonstrate that considering the statistical dependencies and exploring the contributions among paired acoustic-prosodic features (cf. the TCM-EWCCM method) can further enhance the speech emotion recognition accuracy.

There are still several issues that need to be further explored in the future. First, the conversation-based emotion language model which characterizes temporal expression evolution between emotional states should be considered for conversation-based speech emotion recognition. Second, more acoustic and prosodic features should be considered such as linear predictor coefficients (LPC) and zero crossing rate (ZCR). In addition, future research to recognize an expanded set of emotion categories is envisioned that may be useful in other applications and contexts.

Table II
Average Emotion Recognition Rates of Four Emotional States in Different Feature Sets

Methods	Conventional HMM	TCM
Pro only	62.93%	83.75%
Pro+Aco	68.04%	85.37%

Table III
A Crosstab of Post-Hoc Pairwise Comparisons for Chi-Squared Test Based on The z-Test with Bonferroni Corrections

		MODEL		
		HMM (A)	TCM (B)	TCM-EWCCM (C)
Significance Comparison	Correct		A	AB
	Error	BC	C	

References

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

[2] 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 Processing Magazine*, vol. 18, no. 1, pp. 33–80, 2001.

[3] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, pp. 389–405, 2005.

[4] 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.*, vol. 31, no. 1, pp. 39–58, 2009.

[5] 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, 2011.

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

[7] 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.

[8] J. C. Lin, C. H. Wu, and W. L. Wei, "Emotion recognition of conversational affective speech using temporal course modeling," *INTERSPEECH*, pp. 1336–1340, 2013.

[9] C. H. Wu, J. C. Lin, and W. L. Wei, "Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course," *IEEE Trans. on Multimedia*, vol. 15, no. 8, pp. 1880–1895, 2013.

[10] W. L. Wei, C. H. Wu, J. C. Lin, and H. Li, "Exploiting psychological factors for interaction style recognition in spoken conversation," *IEEE Trans. Audio, Speech and Language Processing*, vol.22, no.3, pp. 659–671, 2014.

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

[12] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, no. 10-12, pp. 1913–1920, 2008.

[13] 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.

[14] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 116–125, 2012.

[15] J. C. Lin, C. H. Wu, and W. L. Wei, "Facial action unit prediction under partial occlusion based on error weighted cross-correlation model," *Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 3482–3486, 2013.

[16] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 29, no. 10, pp. 1683–1699, 2007.

[17] Y. Tong, J. Chen and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 32, no. 2, pp. 258–274, 2010.

[18] Y. Li and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Trans. Image Processing*, vol. 22, no. 7, pp. 2559–2573, 2013.

[19] A. Kapoor, R. W. Picard, and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest," *Proc. Int'l Conf. Pattern Recognition*, vol. 3, pp. 969–972, 2004.

[20] Y. Ivanov, T. Serre, and J. Bouvrie, "Error weighted classifier combination for multi-modal human identification," Technical Report CBCL paper 258, Massachusetts Institute of Technology, Cambridge, MA, 2005.

[21] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," *Proc. Int'l Symposium on Multimedia*, pp. 250–257, 2008.

[22] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: multi-modal fusion of different timescale features," *IEEE Workshop on Multimedia Signal Processing*, pp. 48–51, 2007.

[23] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[24] 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.

[25] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Trans. Affective Computing*, vol. 4, no. 3, pp. 280–290, 2013.

[26] O. Rudovic, S. Petridis, M. Pantic, "Bimodal log-linear regression for fusion of audio and visual features," *In Proceedings of the 21st ACM international conference on Multimedia*, pp. 789–792, 2013.

[27] P. Boersma and D. Weenink, Praat: doing phonetics by computer. http://www.praat.org/. 2007.

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

[29] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[30] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[31] H. M. Cooper, L. V. Hedges, and J. C. Valentine, *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, NY 2009.

[32] L. E. Toothaker, *Multiple Comparison Procedures*. Sage Pubns, 1992.