Hierarchical Modeling of Temporal Course in Emotional Expression for Speech Emotion Recognition

Chung-Hsien Wu, Wei-Bin Liang, Kuan-Chun Cheng
Dept. of Computer Science and Information Engineering,
National Cheng Kung University
Tainan, Taiwan
{chunghsienwu, liangnet, davidcheng817}@gmail.com

Jen-Chun Lin
Institute of Information Science
Academia Sinica
Taipei, Taiwan
jenchunlin@gmail.com

Abstract—This paper presents an approach to hierarchical modeling of temporal course in emotional expression for speech emotion recognition. In the proposed approach, a segmentation algorithm is employed to hierarchically chunk an input utterance into three-level temporal units, including low-level descriptors (LLDs)-based sub-utterance level, emotion profile (EP)-based sub-utterance level and utterance level. An emotion-oriented hierarchical structure is constructed based on the three-level units to describe the temporal emotion expression in an utterance. A hierarchical correlation model is also proposed to fuse the three-level outputs from the corresponding emotion recognizers and further model the correlation among them to determine the emotional state of the utterance. The EMO-DB corpus was used to evaluate the performance on speech emotion recognition. Experimental results show that the proposed method considering the temporal course in emotional expression provides the potential to improve the speech emotion recognition performance.

Keywords—Temporal course; speech emotion recognition; hierarchical correlation model

I. INTRODUCTION

In human-human communication, much paralinguistic information (e.g., speaker's emotions and intents) is exchanged through speech. The abilities in recognizing, interpreting, and processing the underlying information from affective speech are important in realizing an affective and intelligent system [1], [2]. In the past, most speech emotion recognition approaches focused on utterance-level recognition where an emotion class label was assigned to the target utterance by considering the global (so called static) features from the whole utterance [3]. However, in conversational speech, the emotional expression of an utterance can be described by different temporal phases, including onset, apex and offset phases as shown in Fig. 1 [4]–[6]. Understanding the issue of temporal courses is critical to specify the cognitive system devoted to speech emotion recognition, since speech expressions are ubiquitously dynamic, dictated by their temporal structure. Thus, recognizing emotion from speech utterance by considering only the global features extracted from the whole utterance lacks the dynamic or temporal information and results in unsatisfactory performance in distinguishing the arousal-similar emotions, e.g. Anger and Joy [4], [7]. Opposite to utterance level recognition (i.e., static modeling approaches), the frame-level recognition (i.e., dynamic modeling approaches) was proposed to capture the temporal characteristics of affective speech [3], [8], [9]. In such methods, local features [3], such as the values of pitch, intensity, and formant of low-level descriptors (LLDs) in each speech frame, instead of the global features, are commonly used. The motivation for using dynamic modeling is that the fluctuation of the frame-based emotional characteristic is inevitable, and it is very difficult to represent the temporal variation by using the global features. However, as each speech frame only covered a very short period of speech utterance, the frame-level information may not be sufficient to represent the emotional intent. For better modeling the emotional content, exploring more appropriate speech units is thus critical.

Recent studies began to investigate the intermediate alternatives, which used appropriate analysis unit in addition to utterance-level or frame-level units. A few studies have explored and investigated the effectiveness of different units such as words, syllables, and fixed-length segments [9]–[11]. The results show that combining the information from the intermediate units can effectively improve system performance. However, the mentioned intermediate units either lack accurate boundary detection results (i.e., speech
recognition error resulting in incorrect detection of word and syllable will dramatically degrade the emotion recognition performance) or lack the emotion-oriented meaning (i.e., fixed-length segment). Accordingly, different from the mentioned intermediate units, in this study, the emotion-oriented multilevel unit chunking approach is firstly proposed to chunk the input speech utterance into LLDs-based sub-utterance level, emotion profile (EP)-based [12]–[14] sub-utterance level and utterance level units to better represent the temporal emotional content. A hierarchical correlation model (HCM) is then proposed to model the correlations among the corresponding vector-quantized codewords of the extracted EPs. Finally, the emotion recognition result is determined by using the proposed hierarchical correlation model in the test phase.

II. SYSTEM FRAMEWORK

Figure 2 illustrates the system framework of the proposed approach. The training phase is composed of multilevel unit chunking, emotion profile construction and hierarchical correlation modeling. First, the chunking module is proposed to chunk each input speech utterance into LLDs-based sub-utterance level, emotion profile (EP)-based [12]–[14] sub-utterance level and utterance level units to better represent the temporal emotional content. A hierarchical correlation model (HCM) is then proposed to fuse the three level outputs from their corresponding recognizers and further model the correlation among them to determine the final emotion of the utterance.

III. MULTILEVEL UNIT CHUNKING

Before unit chunking, the spectrogram of the speech utterance is used to extract the speech segments with similar spectral properties using the Canny edge detection algorithm [16]. At the first level of the chunking process based on the obtained speech segments, the LLDs-related features including the functionals of LLDs and ΔLLDs in each speech segment are first extracted [15]. The Canny edge detection algorithm is then employed to chunk the speech segments into a sequence of LLDs-based sub-utterance units, based on the two-dimensional LLDs-related features (i.e., LLD-related features in Y-axis and segment indices in X-axis). Then, the LLD-related features in the LLDs-based sub-utterance units are re-estimated and fed to the SVM-based EP model to obtain the EPs. In the second level, the EPs are further chunked to obtain a sequence of EP-based sub-utterance units. Finally, the LLDs-based sub-utterance units, EP-based sub-utterance units and utterance unit (the third-level unit) are obtained as the temporal units to construct the hierarchical correlation model for emotion recognition.

IV. HIERARCHICAL CORRELATION MODEL

Given the observations $O$ of the input utterance, the most likely emotional state $e^*$ can be obtained by

$$e^* = \arg \max_{e_i \in E} P(e_i | O)$$  \hspace{1cm} (1)

where $e_i$ represents the $i$-th emotional state in the emotional state set $E$. Using Bayes' rule, (1) is rewritten as

$$e^* = \arg \max_{e_i \in E} P(O | e_i)P(e_i)$$  \hspace{1cm} (2)

where $P(O | e_i)$ is the probability of $O$ obtained from the recognizer $e_i$ based on the multilevel temporal units, and $P(e_i)$ is the prior probability of each emotional state and regarded as an uniform distribution in this study.

For temporal units estimation, the observation $O$ extracted from the multilevel temporal units is regarded as an observation sequence from $L$ temporal unit levels and represented as $O = O^{1}, O^{2}, \ldots, O^{L}$. The term $P(O | e_i)$ in (2) is thus re-written as

$$P(O | e_i) = P(O^{1}|e_i)P(O^{2}|e_i)\ldots P(O^{L}|e_i)$$  \hspace{1cm} (3)
maximum entropy, the joint probability is estimated from two different signal sources. Invoking the principle of maximum entropy, the joint probability is estimated as

\[ P(X, Y) = P(X)P(Y) \frac{p(w, v)}{p(w)p(v)} \]  
(4)

where \( w \) and \( v \) are the output of mapping functions \( f_X(X) \) and \( f_Y(Y) \) for dimensionality reduction, respectively. The roles of the two mapping functions are two-fold: to make \( p(w,v) \) easier to estimate than \( p(X,Y) \) and to capture the dependency between \( X \) and \( Y \). In this study, we employ the \( k \)-means clustering algorithm to construct the codebooks of clustered EPs for the temporal units at each level. The EP of each speech segment is mapped to a codeword \( d_i^l \) which represents the \( n \)-th codeword in the codebook at the \( l \)-th level.

To simplify the derivation, we consider a top-down hierarchical structure for the temporal evolution shown in Fig. 4. Suppose that the temporal unit at each level is dependent on its parent. So, (3) is further derived as (5) and simplified as follows using the example shown in Fig. 4.

\[ P(O_l | e_l) = \prod_{i=1}^{N_l} P(O_{l-1} | e_{l-1}) \frac{P(d_l^i | d_{l-1}^i, e_{l-1})P(d_{l-1}^i | d_{l-2}^i, e_{l-2})}{P(d_{l-2}^i | e_{l-2})P(d_{l-1}^i | e_{l-1})} \]  
(5)

where

\[ P(d_l^i | d_{l-1}^i, e_{l-1}) = P(d_l^i | d_{l-1}^i, e_{l-1})P(d_{l-1}^i | d_{l-2}^i, e_{l-2}) \approx P(d_l^i | d_{l-1}^i, e_{l-1})P(d_{l-1}^i | d_{l-1}^i, e_{l-1}) \]

and

\[ P(d_{l-1}^i | d_{l-2}^i, e_{l-2}) = P(d_{l-1}^i | d_{l-1}^i, e_{l-1}) \approx P(d_l^i | d_{l-1}^i, e_{l-1})P(d_{l-1}^i | d_{l-1}^i, e_{l-1}) \]

Based on the proposed hierarchical structure, each term in (5) presents the relation between the parent and its children. According to the analysis results, longer duration of a chunked segment may contain relatively sufficient information to represent emotion content; therefore, the probability at the \( l \)-th level \( P(O_l | e_l) \) is further weighted by the duration defined as

\[ P(O^l | e_l) = \left( \frac{\text{dur}(o^l)}{\text{dur}(O)} \right)^{1/N_l} \]  
(6)

where \( O^l = o_1^l, o_2^l, \ldots, o_N^l \), \( N_l \) is the number of the chunked segments at the \( l \)-th level and \( \text{dur}(.) \) represents the number of frames in the temporal unit.

V. EXPERIMENTAL RESULT

The proposed system was evaluated by the Berlin Speech Emotional Database (EMO-DB) [18], which covers seven categories of emotional states (anger, boredom, disgust, fear, joy, sadness and neutrality) with 535 utterances in total. For feature extraction used in the openSMILE script of INTERSPEECH 2009 Emotion Challenge, a total of 384 types of features were extracted. Also, speaker normalization by feature normalization [19] with respect to the whole individual speaker contexts was considered to improve the system performance. Herein, speaker normalization is realized by mapping the overall distribution of the features for each speaker to a normal distribution. This preserves the differences between distributions for each emotion while normalizing the values across speakers. For the construction of the emotion recognizers, the LibSVM [20] was employed to construct the multi-class SVMs with RBF kernel.

Evaluation was performed using Leave-One-Speaker-Out cross validation in the speaker independent mode. We employed the weighted average (WA) and unweighted average (UA) reflecting unbalance among classes) of class-wise recognition rates. In this study, evaluation was focused on the effects of the proposed hierarchical correlation model for each emotion class and the sizes of the codebooks at the three levels were chosen as 10, 15, and 20 for utterance unit, EP-based sub-utterance unit, and LLDs-based sub-utterance unit levels, respectively, that is, achieved better performance in our preliminary experiments.

Figure 5 shows the WA and UA recognition accuracies conditioned on the combinations of different temporal units using the proposed HCM. The UA performance improvement of the HCM among different temporal unit combinations for each emotion class is further shown in Fig. 6. For performance comparisons, even though the combination of LLDs-based sub-utterance level (LLDs-SL) and utterance level (UL) units slightly degraded the performance than considering only the UL, it improved the performance by combining the three levels (i.e., UL, EP-SL (EP-based sub-utterance level) and LLDs-SL in the “proposed” HCM) and reached the highest performance as shown in Fig. 5. The results imply that different temporal unit levels are complementary to each other for speech emotion recognition. Compared to the recognition result of
UL, the result further shows in Fig. 6 that different temporal unit combinations in HCM have different effects on different emotion classes. The results provide indications that different temporal units needed for the systems to recognize emotion varies by emotion type.

To further demonstrate the effectiveness of the proposed HCM, we also compared the HCM to the score-level fusion approach as shown in Table I. For the score fusion-based method, the \( P(O|e) \) obtained for each temporal level unit were concatenated as a feature vector and then used for emotion recognition using an SVM-based classifier. The results demonstrated that the proposed HCM achieved a better performance. The reason may be attributed to that the correlations between each temporal unit are further modeled in the proposed HCM. Based on the mentioned analyses, compared to the utterance level recognition and score-level fusion approaches, the proposed HCM achieved the best recognition performance shown in Fig. 5 and Table I.

The average number of speech segments and chunked LLDs-SL and EP-SL segments are further shown in Fig. 7. From the results, the average number of speech segments and chunked segments for sad utterances were larger than the other emotions. More segments obtained for sadness may result from the acoustic variation in speaking rate, pitch and energy when speaking with sad emotion. People in sad emotion are unable to properly control their pronunciation, such as steady speaking rate, less-varied pitch and moderate energy. In addition, we also found that the average number of the chunked segments was greatly reduced compared to that of the speech segments. Even though the average number of chunked LLDs-SL and EP-SL segments are relatively few, the extracted chunked LLDs-SL and EP-SL segments are still promising in speech emotion recognition as shown in Fig. 5 and 6. The result demonstrates that the proposed multilevel unit chunking approach provides the potential for selecting appropriate segments in analyzing the affective content of a speech utterance.

**VI. CONCLUSION**

This paper presented a new approach to speech emotion recognition using multilevel unit chunking and a hierarchical correlation model. Variable-length units are used to interpret the temporal characteristics of affective utterance, and provide the contextual information at different levels. Experimental results show that modeling the relationships among temporal units at different levels in the proposed HCM provides the potential to improve the speech emotion recognition accuracy. There are several issues that need to be further explored in the future. First, conversation-based speech corpus should be further considered for evaluating the proposed approach. Second, an emotion language model which characterizes temporal evolution between emotional states should be considered. Finally, exploring the expression styles from different users is a viable direction for effective emotion recognition, which is not only related to the expression intensity, but also related to the expression manner and may be significantly associated to personality trait [5].

**REFERENCES**


