AFFECTIVE STRUCTURE MODELING OF SPEECH USING PROBABILISTIC CONTEXT FREE GRAMMAR FOR EMOTION RECOGNITION

Kun-Yi Huang[1], Jia-Kuan Lin[1], Yu-Hsien Chiu[2], and Chung-Hsien Wu[1]

[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
[2]Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University, Kaohsiung, Taiwan
chunghsienwu@gmail.com

ABSTRACT

A complete emotional expression typically contains a complex temporal course in a natural conversation. Related research on utterance-level and segment-level processing lacks understanding of the underlying structure of emotional speech. In this study, a hierarchical affective structure of an emotional utterance characterized by the probabilistic context free grammars (PCFGs) is proposed for emotion modeling. SVM-based emotion profiles are obtained and employed to segment the utterance into emotionally consistent segments. Vector quantization is applied to convert the emotion profile of each segment into codewords. A binary tree in which each node represents a codeword is constructed to characterize the affective structure of the utterance modeled by PCFG. Given an input utterance, the output emotion is determined according to the PCFG-based emotion model with the highest likelihood of the speech segments along with the score of the affective structure. For evaluation, the EMO-DB database and its expansion in utterance length were conducted. Experimental results show that the proposed method achieved emotion recognition accuracy of 87.22% for long utterances and outperformed the SVM-based method.

Index Terms— Speech emotion recognition, probabilistic context free grammar, affective structure model

1. INTRODUCTION

As the rapid development of human-computer interface, providing emotion detection capability to a computer is desirable for versatile affective applications. Recognition of emotions in speech and facial video plays an important role in affective computing [1][2][3][4]. Recently, several research focused on utterance-level or frame/segment-level speech emotion recognition [5]. However, in conversational speech, the emotional expression of an utterance can be described by the temporal courses [3] considering its static and dynamic properties typically. There still lacks understanding of the underlying structure of emotional speech to completely characterize an emotion expression. This study aimed to develop an automatic recognition mechanism by detecting and modeling the structural fluctuation of emotions embedded in a speech utterance.

Most of traditional emotion recognition methods focused on utterance-level recognition. Even though additional acoustic information has been integrated into the recognition process [6], it is still difficult to describe the complete dynamic characteristics of human emotion expression due to high degree of abstraction. Some recent work used segment-level methods. More detailed information on segment units in an utterance can provide structural information. Segmentation with a fixed-length window, which is called technical units such as frame, time slices, etc. [7][8][9], could provide fast processing on emotion recognition. However, important prosodic information in human emotional expression may be lost and emotionally consistent units also affect the classification performance. In order to keep the variation information, varying-length segmentation approach has been proposed to extract meaningful units, such as phrase, word, syllable, etc. [10][11][12][13]. The segmented units could provide more useful information for emotion expression modeling [14]. The hierarchical structure not only considers the relation of neighboring prosodic states but also the higher or lower level prosodic relations. Modeling different temporal courses in an expression is helpful to emotion recognition.

Motivated by the findings, we proposed a probabilistic context free grammar (PCFG)-based approach [15] to model emotion fluctuation as an affective structure by more precise segmentation and characterization of the relation between segments. Canny edge detection algorithm [16] is employed to detect the hypothesized segment boundaries of speech signal according to spectral similarity. A multiclass SVM-based emotion classifier is constructed to output the emotion profiles [14] which are then used to estimate the emotional difference between two consecutive segments. A hierarchical binary tree is constructed until all the hypothesized boundaries have been determined. Followed by vector quantization, the entire speech segments within the layered structure are clustered and constructed with a codebook by using the k-means algorithm. Multi-layered segments are mapped to a codeword-based structure for eliminating the data sparseness problem and further modeled by the probabilistic context free grammar. The hierarchical structure can provide useful structural

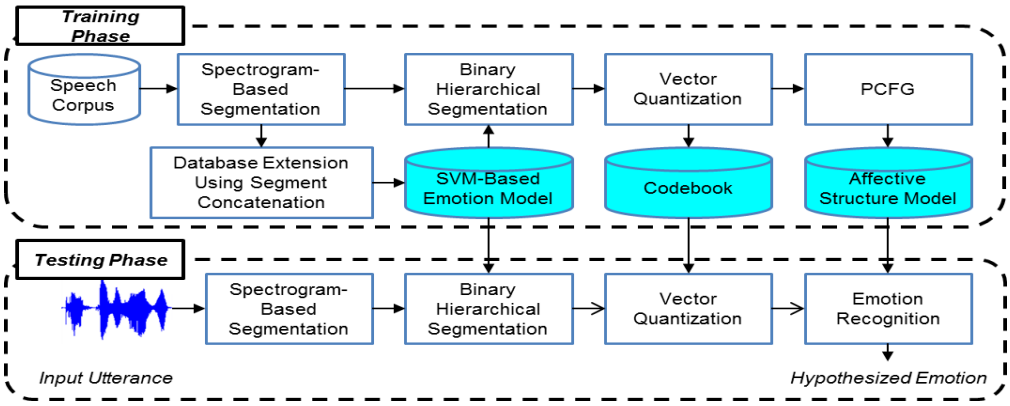

Figure 1. The proposed system framework.

information as the side information for emotion recognition of a speech utterance.

2. PROPOSED METHOD

Figure 1 shows the proposed system framework. The training phase performs the binary segmentation for extracting more precise variable-length speech segments according to the emotogram (a sequence of emotion profiles) of the input utterance [14]. An emotion profile codebook is generated for constructing the proposed PCFG-based affective structure. In the test phase, the emotion is determined by jointly estimating the binary hierarchy of speech segments and its corresponding affective structure of the speech utterance. Using the Canny algorithm, the input utterance is pre-segmented into a sequence of acoustically related speech segments with a length of N. Feature vectors in the speech segment sequence $S_{1,N} = s_1, s_2, ..., s_N$ are extracted to form the feature vector sequence $X_{S_{1,N}} = x_{s_1}, x_{s_2}, ..., x_{s_N}$. $\Omega_E = \{E_e \mid 1 \le e \le 7\}$ represents the emotion space with seven emotion categories. The emotionally-consistent segments are further classified using the SVM to obtain the associated emotion profiles, which could provide multiple probabilistic class information for emotion likelihood estimation. The feature vector sequence is further converted into a codeword sequence $C_{S_{1,N}} = c_{s_1}, c_{s_2}, ..., c_{s_N}$. The emotion category with the highest probability for a given sequence $S_{1,N}$ from the speech level and the corresponding codeword-level structural sequence is obtained.

$$\begin{aligned}\hat{E} &= \arg\max_{E_e \in \Omega_E} \log P(E_e \mid S_{1,N}) = \arg\max_{E_e \in \Omega_E} \log \frac{P(S_{1,N} \mid E_e) P(E_e)}{P(S_{1,N})} \\ &\cong \arg\max_{E_e \in \Omega_E} \log P(X_{S_{1,N}}, C_{S_{1,N}} \mid E_e) \\ &\cong \arg\max_{E_e \in \Omega_E} \left(\log P(X_{S_{1,N}} \mid E_e) + \log P(C_{S_{1,N}} \mid E_e) \right) \\ &\cong \arg\max_{E_e \in \Omega_E} \left(\alpha \log P(X_{S_{1,N}} \mid E_e) + (1-\alpha) \log P(C_{S_{1,N}} \mid E_e) \right)\end{aligned} \quad (1)$$

where $P(X_{S_{1,N}} \mid E_e)$ represents the speech segment-based emotion likelihood obtained from SVM and $P(C_{S_{1,N}} \mid E_e)$ is the affective structure likelihood based on PCFG.

2.1. Emotion Profile Likelihood

Consider the Q hypothesized boundaries $B = \{b_1, b_2, .., b_Q\}$ in a speech segment sequence $S_{1,N}$. $X_{S_{1,N}}$ is feature vector sequence extracted from the sequence $S_{1,N}$. Figure 2 shows the structural representation of the proposed binary hierarchical segmentation. The basic idea is based on the binary tree, which is constructed according to the emotion profile difference in the speech signal. The Euclidean distance between the emotion profiles of two segments for the segmentation point b_q is estimated as follows.

$$Dis(b_q) = \sqrt{\sum_{e=1}^{7} \left(\log P(X_{S_{1,b_q}} \mid E_e) - \log P(X_{S_{b_q,N}} \mid E_e) \right)^2} \quad (2)$$

The point with the highest distance above a threshold λ is determined as the most probable boundary b^*. The binary hierarchical algorithm repeats until all the boundaries at the hierarchical levels have been detected.

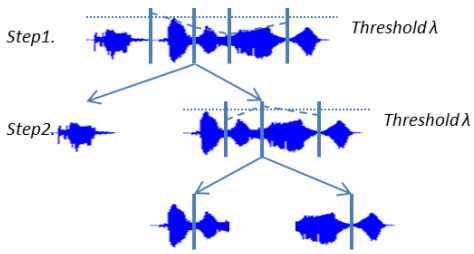

Figure 2. Illustration of binary hierarchical segmentation.

The emotion likelihood of the binary tree is estimated as follows:

$$\log P\left(X_{S_{1,N}} \mid E_e\right) = \sum_{l=1}^{L} \frac{1}{2^{l-1}} \sum_{k=1}^{N_l} \log P\left(X^l_{S^k} \mid E_e\right) \quad (3)$$

where $P\left(X^l_{S^k} \mid E_e\right)$ represents the emotion likelihood of the k-th segment S_k and N^l is number of speech segments at the *l-th* level. The factor $1/2^{l-1}$ is used for weighting the emotion likelihoods at the *l-th* level of the affective structure.

2.2. Affective Structure Likelihood

Based on the above process, an utterance can be represented as a hierarchical binary tree. Each node in the tree represents an emotion profile vector which corresponds to a speech segment in the input utterance. In order to alleviate the data sparseness problem, this study further transforms the emotion profiles generated from the speech segments into codewords. All the segments in the structures are vector quantized to construct a codebook by using the k-means algorithm. Figure 3 shows the transformation process. After the process, the relation between codewords in the hierarchical structure is further modeled by the Chomsky Normal Form (CNF) [15] generally used in PCFG. In CNF, each node only has two child nodes and each node has only two term types – nonterminal term and terminal term. The right side in Fig. 3 shows a CNF-based structure. NC_0 is the root node; NC_i is a nonterminal node and c_l is a terminal node. The probabilities of nonterminal and terminal terms should follow the following constraints:

$$\sum_{e}\sum_{j,k} P\left(NC_i \rightarrow NC_j NC_k \mid G_e\right) + \sum_{e}\sum_{i=l} P\left(NC_i \rightarrow c_l \mid G_e\right) = 1 \quad (4)$$

where G_e is the PCFG for emotion E_e. Given a codeword sequence $C_{S_{1,N}} = c_{s_1}, c_{s_2}, ..., c_{s_N}$, the emotion score of $C_{S_{1,N}}$ with respect to PCFG G_e for emotion E_e is estimated as

$$P\left(C_{S_{1,N}} \mid E_e\right) \equiv P\left(NC_0 \Rightarrow c_{s_1}, c_{s_2}, ..., c_{s_N} \mid G_e\right) = \sum_i \begin{pmatrix} P\left(NC_i \Rightarrow c_{s_a, s_b} \mid G_e\right) \\ \times P\left(NC_0 \Rightarrow c_{s_1, s_{a-1}} NC_i c_{s_{b+1}, s_N} \mid G_e\right) \end{pmatrix} \quad (5)$$

where $P\left(NC_i \Rightarrow c_{s_a, s_b} \mid G_e\right)$ is the inside probability which represents the probability of codeword sequence c_{s_a, s_b} derived from the nonterminal term NC_i. $P\left(NC_0 \Rightarrow c_{s_1, s_{a-1}} NC_i c_{s_{b+1}, s_N} \mid G_e\right)$ is the outside probability for deriving the root node NC_0 into an inner node NC_i with left codeword sequence $c_{s_1, s_{a-1}}$ and right codeword sequence c_{s_{b+1}, s_N}.

3. RESULTS AND DISCUSSION

In order to evaluate the proposed method, the Berlin emotional speech database (EMO-DB) with 535 utterances and 7 emotional states was used. Ten actors performed 7 emotion categories, such as angry, fear, happy, bored, sad, neutral and disgust. For investigating the effect of utterance length, this study further expanded a new data set with a more number of long-duration utterances from EMO-DB. The concatenation criterion is to concatenate two adjacent utterances from the same speaker and emotion, randomly selected from EMO-DB. The concatenated corpus consists of 1495 utterances. Table 1 shows the statistics of the original EMO-DB and the extended database with concatenated utterances and their average length of the utterances (MLU) is shown in Figure 4. OpenSMILE [17] was employed for feature extraction. LLDs and their functionals of the speech segements were extracted. An SVM classifier was constructed using LibSVM [18]. The leave-one-speaker-out cross validation scheme was employed for the following experiments.

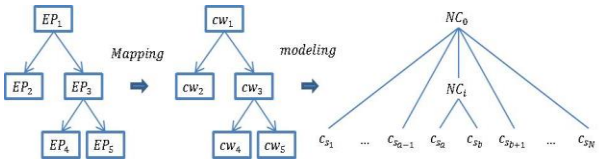

Figure 3. The mapping from binary hierarchy to codeword-based affective structure.

Table 1. Statistics of the original EMO-DB and the concatenated database.

	Number of utterances	
	EMO-DB	Concatenated
Anger	127	362
Fear	69	174
Boredom	81	194
Disgust	46	116
Happiness	71	234
Neutrality	79	224
Sadness	62	191

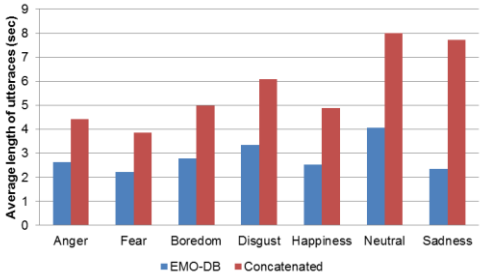

Figure 4. Descriptive statistics in average length of the utterances in the two databases.

Considering the computational efficiency, a codeword number of 7 and $\lambda = 0.4$ were selected for the experiments. Table 2 shows some of the derived rules of anger. The relations between codewords are modeled with the probabilities trained from the concatenated database.

In this study, utterance-level recognition by SVM was regarded as a baseline system. Table 3 shows the recognition results and the comparisons of the SVM-based baseline and the proposed methods. For the two corpora, the average accuracies for the baseline system are similar. This study considered the joint estimation of the speech-based emotion profile and the PCFG-based affective structure for emotion recognition.

Table 2. Some of the derived rules of anger.

Rule	Probability	Rule	Probability
$cw_1 \rightarrow cw_1 cw_2$	0.0095	$cw_2 \rightarrow cw_4 cw_2$	0.0139
$cw_2 \rightarrow cw_1 cw_2$	0.0767	$cw_2 \rightarrow cw_5 cw_2$	0.0331
$cw_2 \rightarrow cw_2 cw_1$	0.0575	$cw_5 \rightarrow cw_4 cw_2$	0.0063

Table 3. Comparisons of the SVM-based baseline and the PCFG-based methods.

	Baseline	Proposed method
EMO-DB	75.88%	76.64%
Concatenated	77.19%	87.22%

Figure 5 shows the experiment on the effect of the weighting factor α. The condition $\alpha = 0$ represents that the utterance is recognized only based on the affective structure model. For the expanded corpus, the proposed method achieved 87.22% when $\alpha = 0.2$. The experimental results also show that the proposed method achieved good performance in long-duration utterances and shows effectiveness to model emotional fluctuation in speech. Our proposed affective structure plays a crucial role on the recognition. Every emotion has its corresponding affective structures useful for emotional speech recognition as shown in Fig. 6. More detailed evaluations on different utterance lengths is shown in Fig. 7. The results show that the proposed method outperformed the conventional SVM-based method when MLU > 2 seconds. Short utterances (< 2 sec) with less emotional fluctuation has low recognition performance.

4. CONCLUSIONS

This work proposed a PCFG-based method to model affective structure of speech for emotion recognition. Canny edge detection algorithm was used to detect hypothesized boundaries from speech. SVM-based emotion model outputs the emotion profiles for detecting the hypothesized boundaries. The proposed binary segmentation approach provide an alternative and efficient modeling of affective structure of emotional speech. The mapping of emotion profile-based binary hierarchical structure is helpful in modeling the underlying structure in speech and every emotion has its corresponding affective structure modeled by PCFG. The experiment results reveal that the proposed method outperformed the conventional SVM-based method for the utterances with long duration. It also shows the potential in modeling emotion fluctuation in long utterance, or even for conversation. Future study on collecting more real corpus is needed for the analysis and recognition of emotions in spontaneous speech.

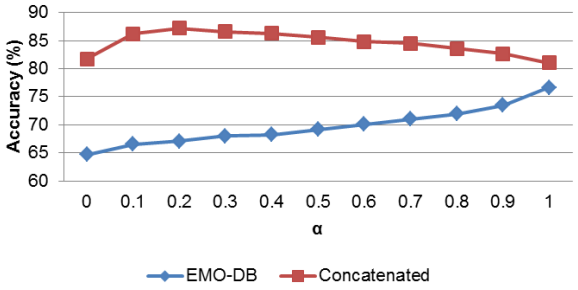

Figure 5. Experimental results on the weighting factor α.

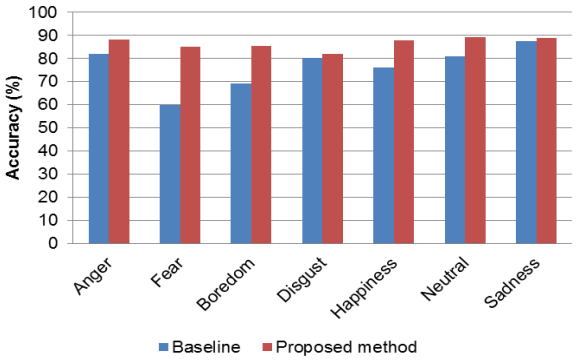

Figure 6. Experimental results for different emotions.

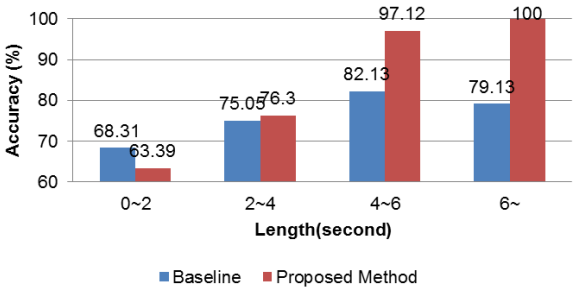

Figure 7. Experimental results for different durations.

5. ACKNOWLEDGEMENT

This paper was supported by the Ministry of Science and Technology of Taiwan under Contract NSC 102-2221-E-006-094-MY3 and the Headquarters of University Advancement at the National Cheng Kung University, which is sponsored by the Ministry of Education, Taiwan.

6. REFERENCES

[1] R. W. Picard, “Affective Computing,” MIT Press, Cambridge, 1997.

[2] C.-H. Wu, J.-C. Lin, W.-L. Wei, "A Survey on Audiovisual Emotion Recognition: Databases, Features, and Data Fusion Strategies,” APSIPA Transactions on Signal and Information Processing, Vol. 3, e12, DOI: 10.1017/ATSIP.2014.11, Published online: 11 November 2014.

[3] 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 Transactions on Multimedia, vol. 15, no. 8, pp. 1880-1895, December 2013.

[4] 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, January 2012, pp.142~156.

[5] C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective Speech based on Multiple Classifiers using Acoustic-Prosodic Information and Semantic Labels,” IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10–21, 2011.

[6] B. Schuller and G. Rigoll, “Timing Levels in Segment-Based Speech Emotion Recognition,” in Proc. INTERSPEECH, Pittsburgh, Pennsylvania, pp. 1818-1821, 17-21 September 2006.

[7] B. Schuller and L. Devillers, “Incremental Acoustic Valence Recognition: An Inter-Corpus Perspective on Features, Matching, and Performance in A Gating Paradigm,” in Proc. INTERSPEECH, Makuhari, Chiba, Japan, pp. 801-804, 26-30 September 2010.

[8] E. Mower and S. Narayanan, “A Hierarchical Static-Dynamic Framework for Emotion Classification,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2372-2375, 22-27 May 2011.

[9] J. H. Jeon, R. Xia, and Y. Liu, “Sentence Level Emotion Recognition Based on Decisions from Subsentence Segments,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4940-4943, 2011.

[10] D.-N. Jiang and L.-H. Cai, “Speech Emotion Classification with The Combination of Statistic Features and Temporal Features,” in Proc. IEEE International Conference on Multimedia and Expo (ICME), vol. 3, pp. 1967-1970, 27-30 June 2004.

[11] A. Batliner, S. Steidl, D. Seppi and B. Schuller, “Segmenting into Adequate Units for Automatic Recognition of Emotion-Related Episodes: A Speech-Based Approach,” Advances in Human-Computer Interaction, vol. 2010, January 2010.

[12] D. Bitouk, R. Verma and A. Nenkova, “Class-Level Spectral Features for Emotion Recognition,” Speech Communication, vol. 52, issues 7-8, pp. 613-625, July-August 2010.

[13] C.-Y. Tseng, S.-H. Pina, Y.-L. Lee, H.-M. Wang and Y.-C. Chen, “Fluent Speech Prosody: Framework and Modeling,” Speech Communication, vol. 46, issues 3-4, pp. 284-309, July 2005.

[14] E. Mower Provost and S. Narayanan, “Simplifying Emotion Classification Through Emotion Distillation”, in Proc. Asia-Pacific Signal and Information Processing Association (APSIPA), Los Angeles, CA, 2012.

[15] C. D. Manning, H. Schutze, “Foundations of Statistical Natural Language Processing,” The MIT Press Cambridge, Massachusetts London, England, June 1999.

[16] J. Canny, “A Computational Approach to Edge Detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679-698, November 1986.

[17] F. Eyben, M. Wöllmer and B. Schuller, “openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in Proc. International Conference on Multimedia, pp. 1459-1462, 2010.

[18] C. C. Chang and C. J. Lin, “LIBSVM: A Library for Support Vector Machines,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, issue. 3, April 2011.