Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels (Extended Abstract)

Chung-Hsien Wu and Wei-Bin Liang

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
chunghsienwu@gmail.com; liangnet@gmail.com

***Abstract*—This work presents an approach to emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information (AP) and semantic labels (SLs). For AP-based recognition, acoustic and prosodic features are extracted from the detected emotional salient segments of the input speech. Three types of models GMMs, SVMs, and MLPs are adopted as the base-level classifiers. A Meta Decision Tree (MDT) is then employed for classifier fusion to obtain the AP-based emotion recognition confidence. For SL-based recognition, semantic labels are used to automatically extract Emotion Association Rules (EARs) from the recognized word sequence of the affective speech. The maximum entropy model (MaxEnt) is thereafter utilized to characterize the relationship between emotional states and EARs for emotion recognition. Finally, a weighted product fusion method is used to integrate the AP-based and SL-based recognition results for final emotion decision. For evaluation, 2,033 utterances for four emotional states were collected. The experimental results reveal that the emotion recognition performance for AP-based recognition using MDT achieved 80.00%. On the other hand, an average recognition accuracy of 80.92% was obtained for SL-based recognition. Finally, combining AP information and SLs achieved 83.55% accuracy for emotion recognition.**

Keywords—Emotion recognition; acoustic-prosodic features; semantic labels; meta decision tree

I. INTRODUCTION

With the exponential growth in available computing power and significant progress in speech technologies, spoken dialogue systems have been successfully applied to several domains, such as navigation systems, air travel information systems, etc. [1][2]. More complex applications (e.g. home nursing [3] educational/tutoring, and chatting [4]) based on new capabilities, like affective interaction, have attracted intensive research due to the overwhelming needs for practical applications. However, to achieve the goal of affective interaction via speech, several problems in speech technologies, including low accuracy in recognition of highly affective speech and lack of affect-related common sense and basic knowledge, still exist.

In the past years, different kinds of affective information, such as emotional keywords [5], speech signals, facial expressions [6][7], linguistic information, dialogue acts [8], etc., have been widely investigated for emotion recognition. Of the affective information previously used, speech is one of the most popular and easily accessible information for emotion recognition. In speech-based emotion recognition, many studies considered acoustic or/and prosodic features, such as pitch, intensity, voice quality features, spectrum and cepstrum [8][9]. On the other hand, several approaches have been proposed to recognize emotional states from purely textual data [5][10][11]. Traditionally, research on the recognition of emotion from text focused on the discovery and utilization of emotional keywords. However, all keyword-based systems have the following problems: (1) ambiguity in emotional keyword definition, (2) sentences with no emotional keywords, and most importantly (3) lack of affect-related semantic and syntactic knowledge base. With further analysis, some researchers proved that textual data is rich with emotion at the semantic level; that is, the emotion is also embedded in the semantic structure of a sentence [5]. A semantic network-based emotion recognition mechanism [5] was proposed using emotional keywords, semantic/syntactic information, and emotional history. However, the link between the parts-of-speech of two words lacks meticulous propagation criteria in the semantic network. Furthermore, for emotion classifier modeling, a variety of pattern recognition methods are utilized to construct a classifier, such as Gaussian mixture model (GMM), support vector machine (SVM), multilayer perceptron (MLP) and decision trees [12][13]. However, a base-level classifier may not perform well on all emotional states. For example, a GMM-based classifier may fail to correctly recognize the neutral emotion, while the MLP-based classifier shows its superiority on neutral emotion recognition. Some studies [7][12] have proved that hybrid/fusion based approaches can achieve higher recognition performance than individual classifiers.

This work presents an emotion recognition approach based on multiple base-level classifiers using acoustic-prosodic information and semantic labels. Three classifiers consisting of GMMs, SVMs, and MLPs are used for emotion detection based on acoustic-prosodic features. A Meta Decision Tree (MDT) [13] is then employed for the fusion of the three classifiers to obtain the emotion recognition confidence. For emotion recognition using semantic labels extracted from the recognized word sequence, the maximum entropy model

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

(MaxEnt) is utilized for emotion recognition. Finally, a weighted product fusion model is used to integrate the results from AP-based and SL-Based approaches to output the recognized emotional state.

II. Framework of Emotion Recognition Using Multiple Classifiers

Fig. 1 illustrates the block diagram of the training and testing procedures for emotion recognition. For AP-based approach, emotional salient segments (ESS) are firstly detected from the input speech. Acoustic and prosodic features including spectrum-, formant-, and pitch-related features are extracted from the detected emotional salient segments and used to construct the GMM-based, SVM-based, and MLP-based base-level classifiers. The MDT is then employed to combine the three classifiers by selecting the most promising classifier for AP-based emotion recognition. On the other hand, the word sequence recognized by a speech recognizer is used in SL-based emotion recognition. The semantic labels of the word sequence derived from an existing Chinese knowledge base called the HowNet [14] are extracted and then a text-based mining approach is employed to mine the Emotion Association Rules (EARs) of the word sequence. Next, the MaxEnt model [15] is employed to characterize the relation between emotional states and EARs and output the emotion recognition result. Finally, the outputs from the above two recognizers are integrated using a weighted product fusion method to determine the final emotional state.

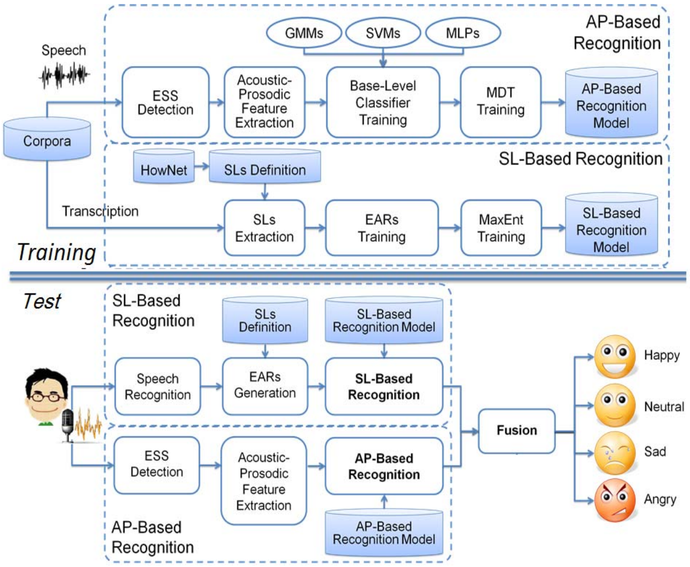

Fig. 1. System block diagram of the training and testing procedures for emotion recognition

III. Multiple Classifiers for Emotion Recognition

A. Acoustic-Prosodic Information-based Classifiers and MDT for Classifier Fusion

Generally, an entire utterance comprises pause/breath segments and salient speech segments. A salient speech segment is defined as the segment in an utterance between two pause/breath segments. In this work, the pitch contour is firstly used to detect the emotional salient segment (ESS) [16] for further acoustic-prosodic information extraction. As shown in Fig. 2, according to the pitch accent tones [17], three types of smoothed pitch-contour patterns are defined as the ESS. The Legendre polynomial-based curve fitting approach used in [18] is adopted for contour smoothing. Of the three pitch-contour types, Type-1 ESS is defined as a complete pitch-contour segment that starts from the point of a pitch rise to the point of the next pitch rise. Type-2 ESS is a monotonically decreasing pitch-contour segment and Type-3 ESS is a monotonically increasing pitch-contour segment. Therefore, the speech features such as pitch, intensity, formants 1 to 4 and formant bandwidths 1 to 4, four types of jitter-related features, six types of shimmer-related features, 3 types of harmonicity-related features and Mel-frequency cepstrum coefficients (MFCCs) are extracted as the acoustic-prosodic information in each ESS for emotion recognition. Moreover, statistics (e.g., mean, standard deviation, maximum, and minimum) and the slope of the above-mentioned features are also used to characterize the ESSs.

For the base-level classifier modeling, given a training data set $\mathbf{F}_j=\{\mathbf{f}_1^j,...,\mathbf{f}_N^j\}$ extracted from N ESSs belonging to the j-th emotional state e_j, where $\mathbf{f}_n^j$ indicates the n-th feature vector, in this work, GMM, SVM, and MLP are employed to model the acoustic-prosodic information. In GMM-based classifier modeling, given the test feature vector $\mathbf{f}$, the emotion output probability can be obtained by formulating the GMM for each emotional state as:

$$P_{GMM}^{j}(\mathbf{f})=\sum\nolimits_k \omega_{j,k}P(\mathbf{f}\mid\mathbf{u}_{j,k},\boldsymbol{\Sigma}_{j,k}) \tag{1}$$

where index j denotes the j-th emotional state, $\omega_{j,k}$ is the mixture weight of the k-th mixture in GMM for the j-th emotional state, and $\boldsymbol{\mu}_{j,k}$ and $\boldsymbol{\Sigma}_{j,k}$ are the mean and covariance of the k-th mixture in the j-th emotion GMM, respectively.

In SVM framework [19], given the feature space $\mathbf{F}$ for all emotional states, a hyperplane is searched to optimally separate the feature space into two subspace, by maximizing the margin $\gamma(\mathbf{f})$ in which the states are represented by +1 and -1, respectively. Since the recognition results of SVMs typically include a number of positives and negatives, the Platt's conversion method [20] is performed to obtain a likelihood output. The conversion is defined as:

$$P_{SVM}^{j}(\mathbf{f})=\frac{1}{1+\exp\{\alpha\cdot\gamma_j(\mathbf{f})+\beta\}} \tag{2}$$

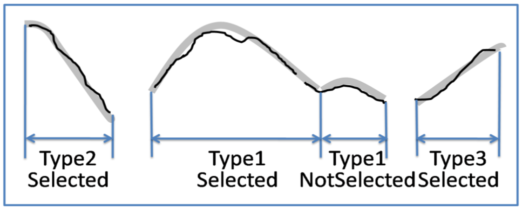

Fig. 2. An illustration of the definition and extraction of emotionally salient segments (ESSs)

where $\gamma_j(\mathbf{f})$ denotes the maximum margin for the j-th emotional state, the parameters α and β are maximum likelihood estimates based on the training features.

Since MLP trained using the back-propagation algorithm can achieve a good emotion recognition performance [21], an MLP with four outputs are adopted and defined as follows.

$$P_{MLP}^{j}(\mathbf{f}) = \frac{1}{1+\exp\{-\alpha\sum_{i}\omega_{i,j}h_i(\mathbf{f})\}} \tag{3}$$

$$h_i(\mathbf{f}) = \frac{1}{1+\exp\{-\alpha\sum_{m}\omega_{m,n}\mathbf{f}_m\}} \tag{4}$$

where ω is the weight in MLP and h_i is the output of the i-th hidden unit in the hidden layer.

Finally, the MDT is employed for classifier selection to output the emotion recognition confidence. By observing the feature vector **f,** the base-level classifier L (e.q. GMM) returns the probability distribution $P_L(\mathbf{f})$ for each emotional state:

$$P_L(\mathbf{f}) = \{P_L^1(\mathbf{f}), P_L^2(\mathbf{f}), ..., P_L^J(\mathbf{f})\} \tag{5}$$

where J is the number of emotional states. The following three probability properties of the emotional state probability distributions obtained from classifier L are used as the attributes in MDTs. First, the maximum probability over all emotional states is denoted as *L_MaxProb*:

$$L\_MaxProb = \max_{j=1}^{J} P_L^j(\mathbf{f}) \tag{6}$$

Next, the entropy of emotional state probability distribution is denoted as *L_Entropy*:

$$L\_Entropy = -\sum_{j} P_L^j(\mathbf{f})\log_2 P_L^j(\mathbf{f}) \tag{7}$$

where j is the index of emotional state. Finally, *L_Weight* represents the fraction of the training data used by the base-level classifier L to estimate the distribution of the emotional state for the test data. Both the *Entropy* and *MaxProb* of a probability distribution can be interpreted as the estimates of the recognition confidence of the model. Moreover, the weight quantifies how reliable the model's estimate of its own confidence is. In the training phase for inducing MDT, the **MLC4.5** algorithm [22] focuses on the accuracy of each base-level classifier L from the classifier set $L^{\#}$ based on the acoustic-prosodic feature set S. Therefore, the measure used in **MLC4.5** is defined as

$$info(S) = 1 - \max_{L\in L^{\#}} \mathbf{accuracy}(L,S) \tag{8}$$

where $info(S)$ is the information gain of S, and $\mathbf{accuracy}(L,S)$ denotes the relative frequency of features in S that are correctly classified by the base-level classifier L.

B. Semantic Label-based Classifier Using MaxEnt

For text-based emotion recognition, an affect-robust speech recognition system based on maximum likelihood linear regression-based adaptation method is constructed to output the recognized text.

1) Emotion Generation Rules

Textual data analysis for emotion recognition shows that not only emotional keywords but also some general terms convey the emotion information. For example, a speaker may say "I finally finished the annoying job." instead of "*I am so glad that* I finally finished the annoying job." The two sentences express the same emotional state, but the critical emotional keyword "*glad*" is not uttered in the former sentence. To solve this problem, the mechanism for generating the emotional states from the viewpoint of psychology should be investigated first. The conditions for generating emotions are summarized according to the previous research on psychology of emotion [23]. These kinds of conditions or environmental situations are based on emotion psychology and are manually derived as emotion generation rules (EGRs). Although EGRs are able to describe situations producing specific emotional states, there still exist some ambiguities inside EGRs. For example, it is clear that "*One may be Happy if someone obtains something beneficial*," but the emotional state of "*some*one *lost something beneficial*" may be "ANGRY" or "SAD". Accordingly, to eliminate the ambiguities in EGRs, the EGRs deduced in [5] with only two opposite emotional states are adopted; POSITIVE contains the emotional states of "HAPPY" and NEGATIVE consists of the emotional states of "ANGRY" and "SAD". Table I illustrates some examples of EGRs and the corresponding emotional states.

Table I. Some examples of EGRs

Emotion State	EGRs
Positive (Happy)	*One may be* HAPPY *if someone reaches his goal*
	One may be HAPPY *if someone have someone's support*
	One may be HAPPY *if someone loses something harmful*
Negative (Unhappy)	*One may be* SAD *if someone failed his goal*
	One may be SAD *if someone lost someone's support*
	One may be ANGRY *if someone disputed with someone*

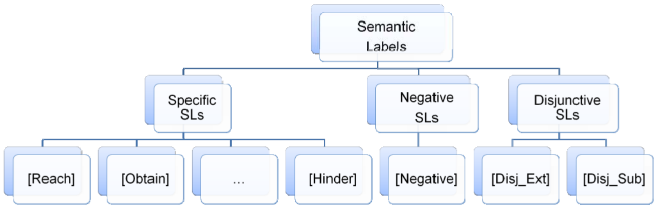

Fig. 3. Tree structure of the Semantic Labels [5]

For EGRs, the definition and extraction of semantic labels (SLs) using HowNet are critical to the entire process. HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts connoted in the lexicon containing Chinese and their English equivalents. In this knowledge base, concept of a word or phrase and its description constitute an entry. For the literal meaning, SL is defined as a word or a phrase that indicates some specific semantic information according to concept definition (DEF). In SL tree structure of Fig.3, three kinds of SLs are defined [5]: *Specific SLs* (SSLs), *Negative SLs* (NSLs) connoting negation words, and *Disjunctive SLs* (DSLs) connoting disjunction words.

With the help of hypernyms defined in HowNet, we first define the SSLs which form the main EGR intention, such as [REACH], [OBTAIN], or [LOSE]. Because most of the concepts with the same intention share the same hypernyms in HowNet, these concepts can be defined by simply defining their hypernyms. Finally, 147 hypernyms used in the evaluation corpora are selected manually from 803 hypernyms in HowNet for the definitions of 15 SSLs in HowNet. Using these semantic labels, emotion recognition does not depend only on the emotional keywords but some general terms.

2) Emotion Association Rules and MaxEnt Modeling

Given a recognized word sequence, the converted SSLs, NSLs and DSLs are obtained by simply comparing the recognized words with the predefined words. Then the hierarchical hypernym structure of the remaining words in the sentence is checked to find appropriate SLs. When a word matches more than one SL, all of the matched SLs will be retained for training the EARs. The *a priori* algorithm [24] is then employed to mine the association rules from the training data.

To model the emotion based on semantic labels, MaxEnt is employed to model the abovementioned EARs as follows.

$$P_{MaxEnt}(e_j \mid EAR) = \frac{1}{Z(EAR)} \exp\{\sum_k \lambda_k O_k(e_j, EAR_k)\} \qquad (9)$$

where $O_k(e_j,EAR_k)$ is the k-th observation function with the weight λ_k. For item-sets used in the previous example, $O_k(e_j,EAR_k)$ can be defined as:

$$O_k(e_j, EAR_k) = \begin{cases} 1, & \begin{array}{l} if\ e_j = Happy, \\ EAR_k = \left\{ \begin{array}{l} < OBTAIN|\text{得到} >, \\ < MONEY|\text{錢財} > \end{array} \right\} \end{array} \\ 0, & Otherwise \end{cases} \qquad (10)$$

$Z(EAR)$ is a normalization term for all training data and defined as:

$$Z(EAR) = \sum_j \exp\{\sum_k \lambda_k O_k(e_j, EAR_k)\} \qquad (11)$$

3) Integration of AP- and SL-based Approaches

Since emotion is generally embedded in speech and textual data which convey different information, acoustic-prosodic information and textual information are expected to complement each other. Therefore, the final step of this work is to fuse the results obtained from AP-based MDT and SL-based MaxEnt. Given the speech signal **X**, the recognized emotion result is determined from Eq. (12) by finding the optimal emotional state e^* over all possible emotional states. In this work, the acoustic-prosodic features **f** are extracted and the EARs are obtained from speech recognition output with SLs. Herein, the EAR and **f** are assumed statistically independent. Lastly, Bayes theory is utilized to obtain Eq. (13).

$$e^* = \arg\max_{e_j \in \mathbf{E}} P(e_j \mid \mathbf{X}) \qquad (12)$$

$$\cong \arg\max_{e_j \in \mathbf{E}} P_{MDT}(e_j \mid \mathbf{f}) P_{MaxEnt}(e_j \mid EAR) \qquad (13)$$

where $P_{MDT}(e_j|\mathbf{f})$ is the emotion recognition confidence from MDT using acoustic-prosodic information. $P_{MaxEnt}(e_j|EAR)$ represents the SL-based emotion recognition confidence using MaxEnt. Moreover, typically in classifier combination, components are weighted differently to obtain an optimal performance based on the contribution of each component. Hence, Eq. (13) is modified as a weighted product fusion mechanism with a weighting factor λ_{AP}.

$$e^* = \arg\max_{e_j \in \mathbf{E}} P_{MDT}(e_j \mid \mathbf{f})^{\lambda_{AP}} P_{MaxEnt}(e_j \mid EAR)^{1-\lambda_{AP}} \qquad (14)$$

where λ_{AP} represents the weight for AP-based recognition ranging from 0 to 1.

IV. Experiment

A. Corpora

Two dialogue corpora were collected for the following experiments. Corpora A and B consist of the utterances from six and two volunteers, respectively. To counterbalance the day-to-day variations, Corpus A was continuously collected for about one month. The data collection system for Corpus A was basically a dialogue system that could guide the speakers to talk about their daily lives by asking questions and recording the answers with acted emotional states. Corpus A can be regarded as an acted emotional speech corpus and is an almost balanced corpus because the distributions of four emotional states of the subjects' daily lives are similar. For Corpus B, the subjects expressed their emotions via the interaction with a computer game. The game often stimulates the invited subjects to generate angry or happy emotional states and speak out the guided comments on the scenario of the game generated by the system while they win or lose the

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

game, respectively. Totally, 2,033 sentences were collected in a lab environment. In both corpora, each sentence was manually annotated as neutral, happy, angry and sad emotional states by the subject who provided the utterance. If different annotation occurred for the same utterance, they will discuss and determine the final annotation. Mixed corpus contains the two above-mentioned corpora.

B. Experimental Setup

For speech recognition, an HTK-based speech recognition system was constructed [25] using a read Mandarin speech corpus TCC-300. The average word accuracy of the speech recognizer is 84.6%. In AP-based recognition, the Praat software [26] was utilized to extract the acoustic-prosodic features. Four GMM-based classifiers, each for one emotional state, were constructed with sixteen mixtures. For the SVM-based classifiers, each SVM for the *j*-th emotional state is trained to discriminate a specific emotion from the others using the open source-LIBSVM [27]. An MLP with one hidden layer, twenty hidden nodes and four output nodes, each representing one emotional state, was constructed. In EAR training, according to the SL-based recognition results, the thresholds of *support* and *confidence* were chosen as 0.3 and 0.5 respectively. The MaxEnt model training for SL-based emotion recognition was realized using an open-source software [28]. In the following evaluations, *K*-fold (*K*=5) cross validation [29] was employed to evaluate the proposed approach.

C. Evaluation Results

The first evaluation is the recognition performance of the three base-level classifiers with/without the ESS in AP-based recognition. For the evaluations without ESS, each segment in an utterance is employed for emotion recognition using the product of probabilities estimated from all segments. For comparison, only the ESS with the largest duration will be selected for evaluation. The comparisons between the approach with the ESS and without the ESS are demonstrated in Table II. The results of GMM-based approach reveal that GMM can model the emotional states well except for the neutral emotion. In other words, GMM-based classifier is not robust enough to recognize all emotional states. SVM-based classifier achieved 78.16% accuracy of emotion recognition with the ESS and outperformed other base-level classifiers in happy, sad, and angry emotions. Compared to GMM-based approach using all training data, SVM-based approach only uses the support vectors to decide the separation hyperplane. Hence, confusing speech features will not be included in the training phase. For MLP-based approach, the classifier obtained the best recognition performance on neutral emotion without/with ESS and the accuracy is 72.40% and 74.40%, respectively. MLP-based approach is useful to process such features because each input node (i.e. each speech feature) and each hidden node contribute differently to each hidden node and each output node (i.e. each emotional state). Briefly, the evaluation results of the three base-level classifiers with ESS are better than those results without ESS. In these evaluation results, there is confusion between neutral emotion and sad emotion because speech data of sad emotion were sometimes uttered normally. Additionally, there is confusion between happy emotion and angry emotion because the speech data of these two emotional states are often uttered loudly.

Then, the effect of the AP-based recognition using MDT for combining the above three base-level classifiers was evaluated with ESS. Because the MDT training procedure needs the meta data (e.g., the probability distributions of the GMM-based emotional recognition), the meta data were obtained from the outputs of the three base-level classifiers using the training data set. Table III is designed to evaluate the effect of ESS. Two corpora were evaluated independently and the results of the mixed corpora show the average performance of the proposed approach. As shown in Table III based on the strategy of selecting a proper classifier, the recognition performance of each emotional state was improved, especially the neutral emotion. The evaluation results of SVM-based recognition are close to the results of MDT-based classifier combination because the MDT is a classifier selection approach instead of combining all classifiers directly.

Table IV shows the evaluation results of SL-based recognition with/without SLs. The average accuracy of SL-based emotion recognition with SLs is 80.92% compared to 67.04% for the evaluation without SLs. Since there are no salient words used for neutral emotional state, the recognition performance is lower than other emotional states. Conversely, the sentences with angry emotion often comprise “intense” words. Therefore, it can achieve the best performance. For the comparison between two corpora, the evaluation results of Corpus A is lower than the results of Corpus B because the text of Corpus A transcribed from the utterances of the subjects’ daily lives is more widespread than the text in Corpus B in which most of the textual contents in Corpus B are speech commands. According to the evaluation on the experiments with/without semantic labels, 86.1% speech recognition accuracy for emotional words is still not good enough. Generally speaking, text-based approach such as semantic labels is more reliable than signal-based approach such as acoustic-prosodic based classifier. For example, in normal utterances, the word “angry” means somebody is in unhappy emotion explicitly. However, high intensity value can be classified into happy (e.g. laughing) and angry emotions (e.g. roaring).

Finally, in Eq. (14), the AP-based and the SL-based recognition results are combined based on the weighting factor λ_{AP}. Fig. 4 shows the evaluation results for the four emotional states. The evaluation result for λ_{AP} =0.4 achieved the best performance of 83.55% for emotion recognition. Generally speaking, speech data harvested by interaction with a game should be able to express the emotion obviously. However, the speech data were affected by the speakers while they were waiting for the system responses.

V. CONCLUSION

This work presents a fusion-based approach to emotion recognition of affective speech using multiple classifiers with acoustic-prosodic information (AP) and semantic labels (SLs). The acoustic-prosodic information was adopted for emotion recognition using multiple classifiers and the MDT was used to select an appropriate classifier to output the recognition confidence. In SL-based approach, the MaxEnt was utilized to model the relationship between emotional states and EARs for emotion recognition. Finally, the integrated results from AS-based and SL-based approaches are used to determine the emotion recognition output. The experimental results show that emotion recognition performance based on MDT outperformed each individual classifier. On the other hand, SL-based recognition obtained an average recognition accuracy of 80.92%. Finally, combining acoustic-prosodic information and semantic labels achieved 83.55% accuracy, which is superior to the classifiers using either acoustic-prosodic information or semantic labels only.

Table II. Evaluation results of the base-level classifiers with/without the ESS

	GMM without ESS			
	Neutral	Happy	Sad	Angry
Neutral	**58.33%**	11.67%	20.00%	11.80%
Happy	11.67%	**72.57%**	3.33%	12.53%
Sad	21.67%	3.33%	**71.67%**	3.33%
Angry	8.33%	12.43%	5.00%	**72.33%**
Average Accuracy	68.73%			

(a)

	GMM with ESS			
	Neutral	Happy	Sad	Angry
Neutral	**63.28%**	10.75%	17.31%	10.45%
Happy	9.66%	**76.08%**	3.26%	11.03%
Sad	18.94%	2.98%	**75.57%**	3.03%
Angry	8.12%	10.19%	3.86%	**75.49%**
Average Accuracy	72.61%			

(b)

	SVM without ESS			
	Neutral	Happy	Sad	Angry
Neutral	**70.00%**	10.33%	11.67%	5.00%
Happy	8.00%	**76.33%**	4.33%	15.00%
Sad	15.67%	4.00%	**78.33%**	3.33%
Angry	6.33%	9.34%	5.67%	**76.67%**
Average Accuracy	75.33%			

(c)

	SVM with ESS			
	Neutral	Happy	Sad	Angry
Neutral	**72.98%**	9.14%	10.36%	4.40%
Happy	7.72%	**78.81%**	2.77%	13.93%
Sad	14.15%	3.48%	**81.01%**	1.84%
Angry	5.15%	8.57%	5.86%	**79.83%**
Average Accuracy	78.16%			

(d)

	MLP without ESS			
	Neutral	Happy	Sad	Angry
Neutral	**72.40%**	11.39%	18.33%	9.41%
Happy	12.37%	**66.98%**	0.00%	11.67%
Sad	9.28%	7.42%	**75.00%**	13.88%
Angry	5.95%	14.21%	6.67%	**65.04%**
Average Accuracy	69.86%			

(e)

	MLP with ESS			
	Neutral	Happy	Sad	Angry
Neutral	**74.40%**	10.26%	17.19%	9.21%
Happy	11.16%	**69.15%**	0.00%	10.41%
Sad	8.99%	7.54%	**76.88%**	13.33%
Angry	5.45%	13.05%	5.93%	**67.05%**
Average Accuracy	71.87%			

(f)

Table III. Evaluation results of MDT-based classifier combination

	MDT				
	Neutral	Happy	Sad	Angry	Average
A	78.13%	80.52%	83.18%	82.97%	81.20%
B	74.09%	79.74%	81.64%	79.75%	78.81%
Mixed	76.11%	80.13%	82.41%	81.36%	80.00%

Table IV. Evaluation results of SL-based recognition

	MaxEnt with Semantic Labels (Accuracy %)				
	Neutral	Happy	Sad	Angry	Average
Corpus A	73.41%	80.55%	82.74%	83.64%	80.09%
Corpus B	77.23%	82.31%	81.92%	85.58%	81.76%
Mixed	75.32%	81.43%	82.33%	84.61%	80.92%
	MaxEnt Without Semantic Labels (Accuracy %)				
Mixed	60.52%	70.14%	67.28%	70.23%	67.04%

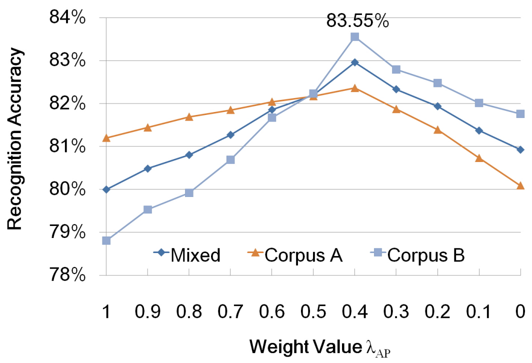

Fig. 4. Evaluation results using weighted product fusion as a function of the weight value

REFERENCES

[1] J. Liu, Y. Xu, S. Senef, and V. Zue, "CityBrowser II: A Multimodal Restaurant Guide in Mandarin," in Proc. *International Symposium Chinese Spoken Language Processing* (*ISCSLP*), pp. 1-4, 2008.

[2] C.-H. Wu and G.-L. Yan, “Speech Act Modeling and Verification of Spontaneous Speech with Disfluency in a Spoken Dialogue System,” *IEEE Trans. on Speech and Audio Processing*, Vol.13, pp.330-344, May 2005.

[3] N. Roy, J. Pineau, and S. Thrun, "Spoken Dialogue Management Using Probabilistic Reasoning," in Proc. *Annual Meeting on Association for Computational* Linguistics (AM-ACL), pp. 93-100, 2000.

[4] D. Jurafsky, R. Ranganath, D. McFarland, “Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation,” in Proc. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for* Computational *Linguistics* (NAACL HLT), pp. 638-646, 2009.

[5] 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, June 2006.

[6] C.-H. Wu, W.-L. Wei, J.-C. Lin, and W.-Y. Lee, “Speaking Effect Removal on Emotion Recognition from Facial Expressions Based on Eigenface Conversion,” *IEEE Trans. Multimedia*, VOL. 15, NO. 8, December 2013, pp. 1732-1744.

[7] 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. Multimedia*, VOL. 15, NO. 8, December 2013, pp.1880-1895.

[8] B. Schuller, G. Rigoll and M. Lang, “Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture,” In Proc. the *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), pp.17-21, 2004.

[9] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” in Proc. *INTERSPEECH*, pp. 312-315, 2009.

[10] C.-M. Lee, and S. S. Narayanan, “Toward Detecting Emotions in Spoken Dialogs,” *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 2, pp. 293-303, MARCH 2005.

[11] L. Devillers, L. Lamel and I. Vasilescu, “Emotion Detection in Task-Oriented Spoken Dialogues,” In Proc. the *IEEE International Conference on Multimedia and Expo* (*ICME*), pp. 549-552, 2003.

[12] I. Luengo, and E. Navas, and I. Hernáez, “Combining Spectral and Prosodic Information for Emotion Recognition in The Interspeech 2009 Emotion Challenge,” in Proc. *INTERSPEECH*, pp. 332-335, 2009.

[13] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion Recognition Using Hierarchical Binary Decision Tree Approach,” in Proc. *INTERSPEECH*, pp. 320-323, 2009.

[14] Z. Dong, and Q. Dong, *HowNet* [Online] Available: http://www.keenage.com/

[15] A. Berger, S. Della Pietra, and V. Della Pietra, “A Maximum Entropy Approach to Natural Language Processing,” *Computational Linguistics*, Vol.22, No. 1, pp. 39-71, 1996.

[16] C.-H. Wu, and Z.-J. Chuang, “Emotion Recognition from Speech Using IG-based Feature Compensation,” *International Journal of Computational Linguistics and Chinese Language Processing*, Vol.12, No. 1, pp.65-78, 2007.

[17] X. Huang, A. Acero, and H.-W. Hon, “Prosody” in Spoken *Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st edition, Prentice Hall PTR, 2005, ch. 15, sec. 15.4.4, pp. 753-755.

[18] C.-H. Wu and J.-H. Chen, “Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis,” *Speech Communication*, Vol.35, pp. 219-237, 2001.

[19] V. Vapnik, *The Natural of Statistical Learning Theory*, Springer-Verlag, New York, 2005.

[20] J. C. Platt, “Probabilities for SV machines,” *Advances in Large Margin Classifiers*, pp. 61-74, MIT Press, 2000.

[21] V. Petrushin, “Emotion Recognition in Speech Signal: Experimental Study, Development, and Application,” in *Proc. International Conference on Spoken Language Processing* (*ICSLP*), pp. 222-225, 2000.

[22] J. R. Quinlan, C4*.5:Programs for Machine Learning*, Morgan Kaufmann

[23] H. Soltau, and A. Waibel, “Acoustic Models for Hyperarticulated Speech,” in Proc. the *International Conference on Spoken Language Processing* (*ICSLP*), 2000.

[24] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proc. the ACM Special Interest Group on Management of Data (SIGMOD)*, pp. 207-216, 1993.

[25] S. J. Young, G. Evermann*, M. J. F. Gales*, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, The *HTK Book*, version 3.4 [Online] Available: http://htk.eng.cam.ac.uk/

[26] P. Boersma, and D. Weenink, *Praat: doing phonetics by computer (version 5.1.05)*, [Computer program] Retrieved May 1, 2010, from http://www. praat.org/

[27] C.C. Chang and C.J. Lin, LIBSVM – A Library for Support Vector Machines,

[Online] Available: http*://www.csie.ntu.edu.tw/~cjlin/libsvm/*

[28] Z. Le, *Maximum Entropy Modeling Toolkit for Python and C++*, [Computer program] from http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[29] P. A. Devijver, and J. Kittler, *Pattern Recognition: A* Statistical *Approach*, Prentice-Hall, London, 1982.