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

Chung-Hsien Wu, *Senior Member*, *IEEE*, and Wei-Bin Liang

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 including spectrum, formant, and pitch-related 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 derived from an existing Chinese knowledge base called HowNet 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 the final emotion decision. For evaluation, 2,033 utterances for four emotional states (Neutral, Happy, Angry, and Sad) are collected. The speaker-independent experimental results reveal that the emotion recognition performance based on MDT can achieve 80.00 percent, which is better than each individual classifier. On the other hand, an average recognition accuracy of 80.92 percent can be obtained for SL-based recognition. Finally, combining acoustic-prosodic information and semantic labels can achieve 83.55 percent, which is superior to either AP-based or SL-Based approaches. Moreover, considering the individual personality trait for personalized application, the recognition accuracy of the proposed approach can be further improved to 85.79 percent.

Index Terms—Emotion recognition, acoustic-prosodic features, semantic labels, meta decision trees, personality trait.

1 Introduction

Speech is one of the most fundamental and natural communication means of human beings. With the exponential growth in available computing power and significant progress in speech technologies, spoken dialogue systems (SDS) have been successfully applied to several domains. However, the applications of SDSs are still limited to simple informational dialog systems, such as navigation systems and air travel information systems [1], [2]. To enable more complex applications (e.g., home nursing [3], educational/tutoring, and chatting [4]), new capabilities such as affective interaction are needed. 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, different kinds of affective information, such as emotional keywords [5], speech signals, facial expressions [6], linguistic information, and dialogue acts [7], 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], [10], [11], [12]. Since communication systems can identify the users' emotional states from different communication modalities, several approaches have been proposed to recognize emotional states from purely textual data [5], [13], [14]. Traditionally, research on the recognition of emotion from text focused on the discovery and utilization of emotional keywords. Using emotional keywords is undoubtedly the most direct way to recognize a speaker's emotions from textual input [13], [14]. However, all keyword-based systems have the following problems: 1) ambiguity in emotional keyword definition, 2) difficulty in emotion recognition of the 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 are 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. In [5], semantic labels (SLs) and attributes were manually defined and adopted for emotion recognition from textual data only. However, the automatically extracted semantic labels and attributes from the imperfectly recognized word sequence are still

- *The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, No. 1, Ta-Hsueh Road, Tainan 701, Taiwan. E-mail: {chunghsienwu, liangnet}@gmail.com.*

Manuscript received 6 May 2010; revised 31 Aug. 2010; accepted 7 Dec. 2010; published online 16 Dec. 2010.
Recommended for acceptance by S. Narayanan.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2010-05-0031.
Digital Object Identifier no. 10.1109/T-AFFC.2010.16.

unreliable for emotion recognition. For emotional state modeling, a variety of pattern recognition methods are utilized to construct a classifier, such as a Gaussian mixture model (GMM), support vector machine (SVM), multilayer perceptron (MLP), and decision trees [15], [16], [17]. 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], [15] have proven 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 (AP) 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) [17] 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 (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.

According to previous study on the role of personality traits in emotion recognition, investigation showed evidence for the role of individual and cultural factors in emotion recognition on a purely verbal task [18]. Besides, emotion recognition generally fails to a certain extent when confronted with different speakers with diverse personality traits. Accordingly, this study also investigates the effect of individual personality characteristics on emotion recognition. In this evaluation, a personality trait of a specific speaker obtained from the Eysenck personality questionnaire (EPQ) [19] is transformed into a weighting factor and integrated into the classifier for personalized emotion recognition.

The rest of this work is organized as follows: Section 2 introduces the framework of an emotion recognition system based on multiple base-level classifiers. Next, MDT for acoustic-prosodic information-based classifier selection, MaxEnt for semantic label-based recognition, and the fusion of the outputs from MDT and MaxEnt are separately described in Section 3. Section 4 presents the experiments for the evaluation of the proposed approach. Finally, concluding remarks are given in Section 5.

2 Framework of Emotion Recognition Using Multiple Classifiers

Fig. 1 illustrates the block diagram of the training and testing procedures for AP and SL-based emotion recognition. For the AP-based approach, emotional salient segments (ESSs) are first 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 [20] 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 [21] 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. Furthermore, in order to investigate the effect of individual personality characteristic, the personality trait obtained from EPQ for a specific speaker is considered for personalized emotion recognition.

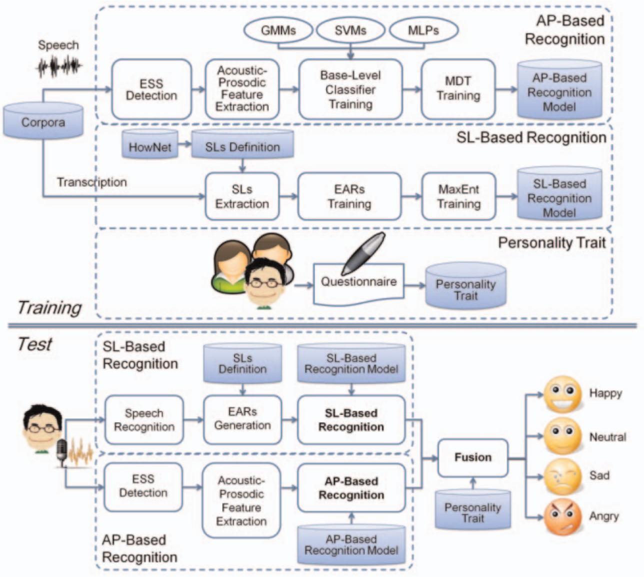


Fig. 1. An overview of the training and testing flowchart of the acoustic-prosodic information-based recognition, the semantic label-based recognition, and the personality trait.

3 Multiple Classifiers for Emotion Recognition

3.1 Acoustic-Prosodic Information-Based Classifiers and MDT for Classifier Fusion

Previous work showed that acoustic and prosodic features reveal different performance due to different data set and task design [9], [15], [22], [23], [24]. This work therefore adopts a broad variety of features from speech signal. Generally, an entire utterance is comprised of 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 first used to detect the ESS [25] for further acoustic-prosodic information extraction. As shown in Fig. 2, according to the pitch accent tones [26], three types of smoothed pitch-contour patterns are defined as the ESS. The Legendre polynomial-based curve fitting approach used in [27] 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. Moreover, only the ESS with the largest duration will be selected for further prosodic feature extraction

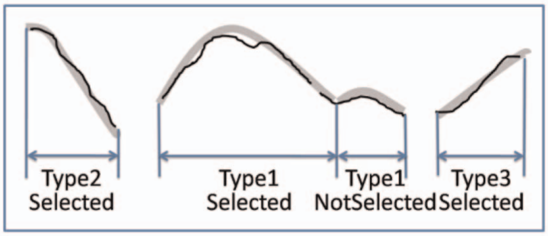


Fig. 2. An illustration of the definition and extraction of emotionally salient segments.

because of two major reasons: 1) The smoothed pitch contour can connect two separate speech fragments to form a longer ESS (e.g., Type1-Selected ESS in Fig. 2) and 2) prosodic cues play an important role in various information extraction tasks [28]; a speech segment with longer pitch contour is likely to include more useful prosodic information for further processing. Therefore, the speech features, such as pitch, intensity, formants 1-4 and formant bandwidths 1-4, four types of jitter-related features, six types of shimmer-related features, three types of harmonicity-related features, and Mel-frequency cepstrum coefficients (MFCCs), are adopted 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 $\mathbf{F}_j = \{\mathbf{f}_1^j, \ldots, \mathbf{f}_N^j\}$ extracted from N ESSs belonging to the jth emotional state e_j, where $\mathbf{f}_n^j$ indicates the nth 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

$$P_{GMM}^j(\mathbf{f}) = \sum_k \omega_{j,k} P(\mathbf{f}|\mathbf{u}_{j,k}, \Sigma_{j,k}), \tag{1}$$

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

In the SVM framework [29], given the feature space $\mathbf{F}$ for all emotional states, a hyperplane is searched to optimally separate the feature space into two subspaces, one representing the target emotional state and the other representing nontarget emotional state, by maximizing the margin $\gamma(\mathrm{f})$ 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 [30] is performed to obtain a likelihood output. The conversion is defined as

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

where $\gamma_j(\mathbf{f})$ denotes the maximum margin for the jth 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 [31], an MLP with four outputs is adopted and defined as follows:

$$\begin{aligned} P_{MLP}^j(\mathbf{f}) &= \frac{1}{1 + \exp\{-\alpha \sum_i \omega_{i,j} h_i(\mathbf{f})\}}, \\ h_i(\mathbf{f}) &= \frac{1}{1 + \exp\{-\alpha \sum_m \omega_{m,n} \mathbf{f}_m\}}, \end{aligned} \tag{3}$$

where ω is the weight in MLP and h_i is the output of the ith hidden unit in the hidden layer.

Finally, the MDT, a novel method for combining base-level classifiers, is employed for classifier selection to output the emotion recognition confidence. Compared to an ordinary decision tree (ODT), the structures are identical on both types of decision trees. The difference between an MDT and an ODT is that the leaves in MDT specify which base-level classifier should be used instead of predicting the fusion probability of emotional state directly. The attributes are derived from the emotional state probability distributions predicted by the base-level classifiers for a given feature vector. By observing the feature vector $\mathbf{f}$, the base-level classifier L (e.g., GMM) returns the probability distribution $P_L(\mathbf{f})$ for each emotional state

$$P_L(\mathbf{f}) = \{P_{{}_L}^1(\mathbf{f}), P_{{}_L}^2(\mathbf{f}), \ldots, P_{{}_L}^J(\mathbf{f})\}, \tag{4}$$

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*:

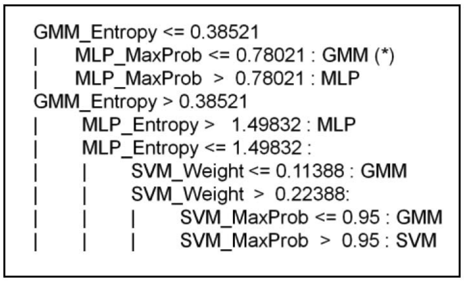
$$L\_MaxProb = \max_{j=1}^{J} P_{{}_L}^j(\mathbf{f}). \tag{5}$$

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

$$L\_Entropy = -\sum_j P_L^j(\mathbf{f}) \log_2 P_L^j(\mathbf{f}), \tag{6}$$

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. Fig. 3 shows an example of an MDT on emotion recognition using above properties as attributes. The leaf denoted by (*) specifies that the GMM-based classifier is to be used to recognize the input data if the Entropy of the probability distribution returned is smaller than 0.385 and the MaxProb in the probability distribution returned by the MLP-based classifier is smaller than 0.78. In the training phase, the algorithm **MLC4.5** algorithm [32] for inducing MDT 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

$$\mathit{info}(S) = 1 - \max_{L \in L^{\#}} \mathbf{accuracy}(L, S), \tag{7}$$



```
GMM_Entropy <= 0.38521
|    MLP_MaxProb <= 0.78021 : GMM (*)
|    MLP_MaxProb  >  0.78021 : MLP
GMM_Entropy > 0.38521
|    MLP_Entropy >   1.49832 : MLP
|    MLP_Entropy <= 1.49832 :
|    |    SVM_Weight <= 0.11388 : GMM
|    |    SVM_Weight  >  0.22388:
|    |    |    SVM_MaxProb <= 0.95 : GMM
|    |    |    SVM_MaxProb  >  0.95 : SVM
```

Fig. 3. Example of a meta decision tree.

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.

3.2 Semantic Label-Based Classifier Using MaxEnt

It is a known fact that affective speech in general degrades recognition performance of a speech recognition task. Therefore, in this work, acoustic models for hyperarticulated speech [33] and the maximum likelihood linear regression (MLLR)-based adaptation method are adopted for affect-robust speech recognition. In this study, emotion-related salient words are defined and a verification process is adopted to provide a word sequence with reliable emotional salient words.

3.2.1 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 previous research on the psychology of emotion [34]. The conditions are assumed to be generic even though different individual backgrounds and environments are involved. For example, one may feel happy when he/she is promoted to a higher position in his/her job, obtains a benefit, or has a good reputation. Then, "*One may be Happy if someone obtains something beneficial*" will be one of the general rules for a "Happy" emotion. 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," which is highly related to the personality trait of the speaker. Accordingly, to eliminate the ambiguities in EGRs, the EGRs deduced in [5] with only two opposite emotional states are adopted; POSITIVE includes the emotional states of "HAPPY" and NEGATIVE includes the emotional states of "ANGRY" and "SAD". Table 1 illustrates some examples of EGRs and the corresponding emotional states.

TABLE 1
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*

For EGRs, the definition and extraction of SLs using HowNet are critical to the entire process. HowNet is an online common-sense knowledge base unveiling interconceptual relations and interattribute 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. Regardless of language types, an entry will be comprised of four items. These items are arranged according to the following sequence:

$$\begin{aligned} \text{W\_C} &= \text{word/phrase form of Chinese,} \\ \text{G\_C} &= \text{word/phrase syntactic class,} \\ \text{E\_C} &= \text{example of usage,} \\ \text{DEF} &= \text{concept definition.} \end{aligned}$$

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 the SL tree structure of Fig. 4, 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]. Table 2 shows 15 predefined SSLs. 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. For example, the concept "FULFILL" is the hypernym of concepts "ACHIEVE," "END," "FINISH," and "SUCCEED." With the definition "*concepts that have a hypernym* 'FULFILL' *can be the* SL [ACHIEVE]," we can convert the four words "ACHIEVE," "END," "FINISH," and "SUCCEED" in the word sequence to SL [ACHIEVE]. Finally, 147 hypernyms used in the evaluation corpora are selected manually from 803 hypernyms in HowNet for the definitions of 15 SSLs in HowNet.

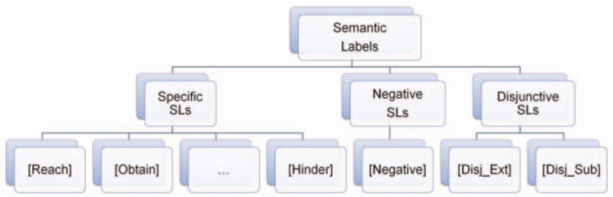
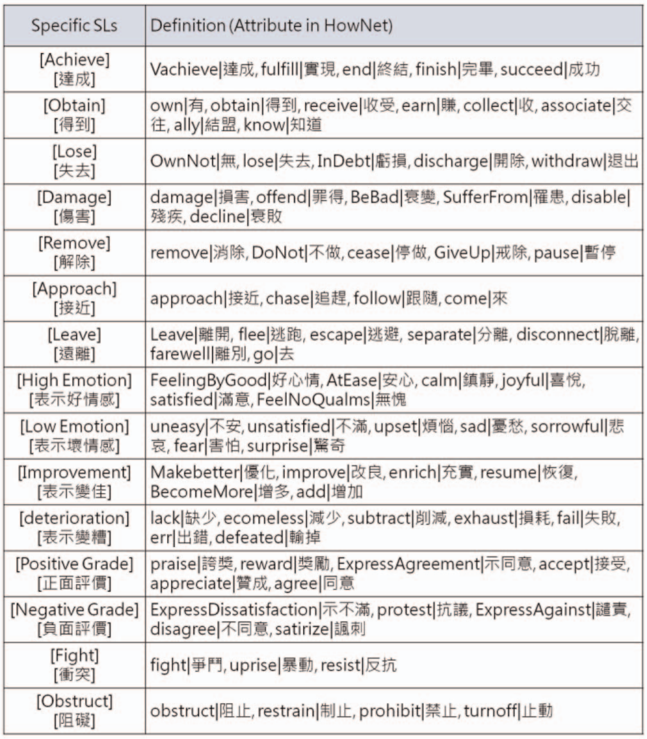


Fig. 4. Tree structure of the Semantic Labels [5].

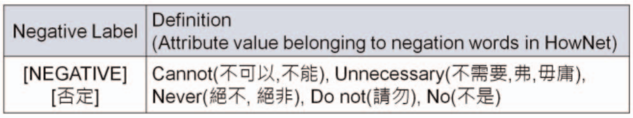
TABLE 2
Some Example Words Belonging to Specific SLs [5]



<table>
<tr><th>Specific SLs</th><th>Definition (Attribute in HowNet)</th></tr>
<tr><td>[Achieve]
[達成]</td><td>Vachieve|達成, fulfill|實現, end|終結, finish|完畢, succeed|成功</td></tr>
<tr><td>[Obtain]
[得到]</td><td>own|有, obtain|得到, receive|收受, earn|賺, collect|收, associate|交往, ally|結盟, know|知道</td></tr>
<tr><td>[Lose]
[失去]</td><td>OwnNot|無, lose|失去, InDebt|虧損, discharge|開除, withdraw|退出</td></tr>
<tr><td>[Damage]
[傷害]</td><td>damage|損害, offend|罪得, BeBad|衰變, SufferFrom|罹患, disable|殘疾, decline|衰敗</td></tr>
<tr><td>[Remove]
[解除]</td><td>remove|消除, DoNot|不做, cease|停做, GiveUp|戒除, pause|暫停</td></tr>
<tr><td>[Approach]
[接近]</td><td>approach|接近, chase|追趕, follow|跟隨, come|來</td></tr>
<tr><td>[Leave]
[遠離]</td><td>Leave|離開, flee|逃跑, escape|逃避, separate|分離, disconnect|脫離, farewell|離別, go|去</td></tr>
<tr><td>[High Emotion]
[表示好情感]</td><td>FeelingByGood|好心情, AtEase|安心, calm|鎮靜, joyful|喜悅, satisfied|滿意, FeelNoQualms|無愧</td></tr>
<tr><td>[Low Emotion]
[表示壞情感]</td><td>uneasy|不安, unsatisfied|不滿, upset|煩惱, sad|憂愁, sorrowful|悲哀, fear|害怕, surprise|驚奇</td></tr>
<tr><td>[Improvement]
[表示變佳]</td><td>Makebetter|優化, improve|改良, enrich|充實, resume|恢復, BecomeMore|增多, add|增加</td></tr>
<tr><td>[deterioration]
[表示變糟]</td><td>lack|缺少, ecomeless|減少, subtract|削減, exhaust|損耗, fail|失敗, err|出錯, defeated|輸掉</td></tr>
<tr><td>[Positive Grade]
[正面評價]</td><td>praise|誇獎, reward|獎勵, ExpressAgreement|示同意, accept|接受, appreciate|贊成, agree|同意</td></tr>
<tr><td>[Negative Grade]
[負面評價]</td><td>ExpressDissatisfaction|示不滿, protest|抗議, ExpressAgainst|譴責, disagree|不同意, satirize|諷刺</td></tr>
<tr><td>[Fight]
[衝突]</td><td>fight|爭鬥, uprise|暴動, resist|反抗</td></tr>
<tr><td>[Obstruct]
[阻礙]</td><td>obstruct|阻止, restrain|制止, prohibit|禁止, turnoff|止動</td></tr>
</table>

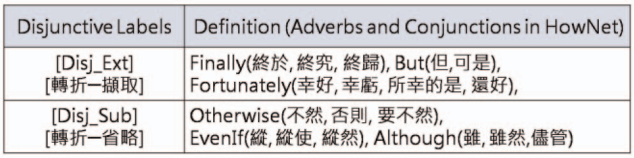
The definitions of two other types of SLs are simpler than for SSLs. Negative SLs indicate the words with negative semantic meanings in the item "DEF" of an HowNet entry. Only the SL [NEGATIVE] is considered in this SL type. For this type of SL, 47 Chinese words are directly defined, such as "CANNOT," "UNNECESSARY," "NO," "NEVER," and so on. Table 3 demonstrates some example words of SL [NEGATIVE]. The disjunctive SLs, comprised of two SLs, [DISJ_EXT] and [DISJ_SUB], indicate the disjunctive semantic meaning. The disjunctive-extractive SL [DISJ_ EXT] implies that the succeeding sentence is more important than the preceding sentence, so the SLs before [DISJ_EXT] will be ignored. Conversely, the SLs following the SL [DISJ_SUB] will be ignored. Because the concepts belonging to these two SLs are enumerable, it is unnecessary to define hypernyms hierarchically in HowNet. Finally, 18 and 11 words can be directly defined as [DISJ_EXT] and [DISJ_SUB], respectively. Table 4 shows some examples of these two types of SLs. The corresponding words of SL [DISJ_EXT] include "FINALLY," "BUT," "FORTUNATELY," and so on. The corresponding words of SL [DISJ_SUB] include "OR," "EVEN THOUGH," "ALTHOUGH," etc.

TABLE 3
Some Example Words Belonging to Negative SLs



Negative Label	Definition (Attribute value belonging to negation words in HowNet)
[NEGATIVE] [否定]	Cannot(不可以,不能), Unnecessary(不需要,弗,毋庸), Never(絕不, 絕非), Do not(請勿), No(不是)

TABLE 4
Some Example Words Belonging to Disjunctive SLs



Disjunctive Labels	Definition (Adverbs and Conjunctions in HowNet)
[Disj_Ext] [轉折一擷取]	Finally(終於, 終究, 終歸), But(但,可是), Fortunately(幸好, 幸虧, 所幸的是, 還好),
[Disj_Sub] [轉折一省略]	Otherwise(不然, 否則, 要不然), EvenIf(縱, 縱使, 縱然), Although(雖, 雖然,儘管)

Using these semantic labels, emotion recognition does not depend only on the emotional keywords but on some general terms. Take the previous sentence "I finally finished the annoying job." as an example. The possible emotional state can be deduced from the general terms, <FINISH> and <JOB>, without the emotional keyword "*glad*."

3.2.2 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. For example, the salient word sequence **W** of the sentence "我(I)賺(earn)很多(a lot of)錢財(money)" is {賺(earn),錢財(money)} and the corresponding SLs are <OBTAIN | 得到> and <MONEY | 錢財>, where "得到(obtain)" is the hypernym of "賺(earn)." The a priori algorithm [35] is then employed to mine the association rules from the training data through two important parameters: support and confidence. These two parameters are calculated as follows:

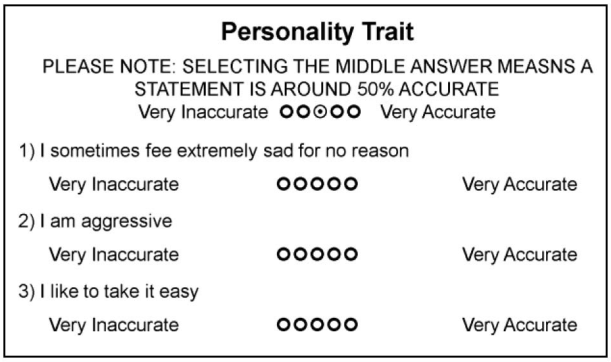
$$Support(A, B) = \frac{Count(A, B)}{Count_{total}}, \tag{8}$$

$$Confidence(A \rightarrow B) = \frac{Count(A, B)}{Count(A)}, \tag{9}$$

where A and B are two item-sets, such as <OBTAIN | 得到> and <MONEY | 錢財>. $Support(A, B)$ and $Confidence(A \rightarrow B)$ are the support and confidence, respectively. $Count(A)$ is the occurrence count of A and $Count(A, B)$ is the count of co-occurrences of A and B over all training data. The value is the total number of training data in which other items co-occur with item A. The threshold values of these two parameters are determined by the SL-based recognition result. If these two parameters are greater than the threshold value, the union of A and B will be retained for MaxEnt modeling. To model the emotion based on semantic labels, MaxEnt is employed to model the abovementioned EARs as follows:

$$P_{MaxEnt}(e_j|EAR) = \frac{1}{Z(EAR)} \exp\left\{ \sum_k \lambda_k O_k(e_j, EAR_k) \right\}, \tag{10}$$

where $O_k(e_j, EAR_k)$ is the kth 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



Personality Trait

PLEASE NOTE: SELECTING THE MIDDLE ANSWER MEASNS A STATEMENT IS AROUND 50% ACCURATE

Very Inaccurate ○○◉○○ Very Accurate

1) I sometimes fee extremely sad for no reason

Very Inaccurate ○○○○○ Very Accurate

2) I am aggressive

Very Inaccurate ○○○○○ Very Accurate

3) I like to take it easy

Very Inaccurate ○○○○○ Very Accurate

Fig. 5. Some example questions in EPQ [36].

$$O_k(e_j, EAR_k) = \begin{cases} 1 & , \text{ if } e_j = Happy, \\ & EAR_k = \left\{ \begin{matrix} <OBTAIN|\text{得到}>, \\ <MONEY|\text{錢財}> \end{matrix} \right\} \\ 0 & , \text{ Otherwise} \end{cases} \tag{11}$$

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

$$Z(EAR) = \sum_j \exp\left\{ \sum_k \lambda_k O_k(e_j, EAR_k) \right\}. \tag{12}$$

3.3 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 acoustic-prosodic information-based MDT and semantic label-based MaxEnt. Given the speech signal $\mathbf{X}$, the recognized emotion result is determined from (13) by finding the optimal emotional state e^* over all possible emotional states. Using Bayes theory, (13) can be rewritten into (14). In this work, the acoustic-prosodic features $\mathbf{f}$ are extracted and the EARs are obtained from speech recognition output with SLs. Hence, (14) is further rewritten as (15). Herein, the EAR and $\mathbf{f}$ are assumed statistically independent, and therefore, (16) is obtained. Last, Bayes theory is utilized to obtain (18).

$$e^* = \arg\max_{e_j \in \mathbf{E}} P(e_j|\mathbf{X}) \tag{13}$$

$$= \arg\max_{e_j \in \mathbf{E}} P(\mathbf{X}|e_j)P(e_j) \tag{14}$$

$$\cong \arg\max_{e_j \in \mathbf{E}} P(\mathbf{f}, EAR|e_j)P(e_j) \tag{15}$$

$$\cong \arg\max_{e_j \in \mathbf{E}} P(\mathbf{f}|e_j)P(e_j)P(EAR|e_j) \tag{16}$$

$$\cong \arg\max_{e_j \in \mathbf{E}} P(e_j|\mathbf{f})P(e_j|EAR)P(EAR) \tag{17}$$

$$\cong \arg\max_{e_j \in \mathbf{E}} P_{MDT}(e_j|\mathbf{f})P_{MaxEnt}(e_j|EAR), \tag{18}$$

where $P_{MDT}(e_j|f)$ 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. $P(EAR)$ in (17) denotes the prior probabilities of the EAR corresponding to the recognized word sequence of the input speech and can be regarded as the same for all emotional states and can be ignored in (18). Moreover, typically, in classifier combination, components are weighted differently to obtain an optimal performance based on the contribution of each component. Hence, (18) is modified as a weighted product fusion mechanism with a weighting factor λ_{AP}.

$$e^* = \arg\max_{e_j \in \mathbf{E}} P_{MDT}(e_j|\mathbf{f})^{\lambda_{AP}} P_{MaxEnt}(e_j|EAR)^{1-\lambda_{AP}}, \tag{19}$$

where λ_{AP} represents the weight for AP-based recognition ranging from 1 to 0 and the weight for SL-based recognition is $1 - \lambda_{AP}$.

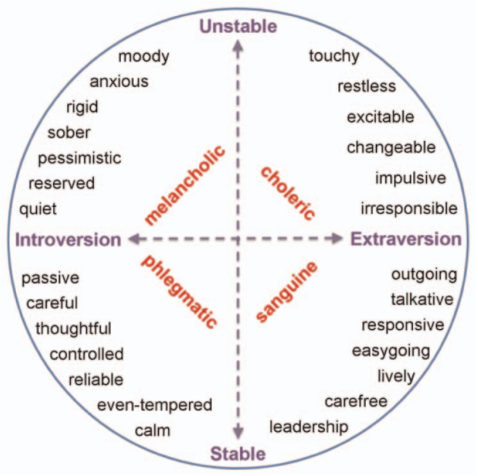


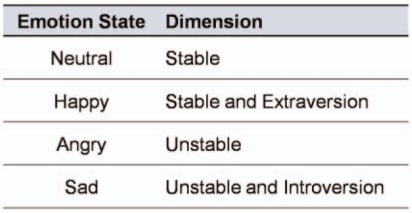
Fig. 6. Personality trait chart [37].

3.4 Consideration of Personality Trait

To investigate the effect of the individual personality trait on emotion recognition, in this work, the EPQ [19] is employed for personality trait evaluation of each individual. Because personality trait is not always reliable and may depend on the speaker's education and cultural background, the personality trait score for each emotion obtained from EPQ is only adopted to weight the system's decision. Fig. 5 illustrates an example of the EPQ. The EPQ is a questionnaire to assess the personality traits of a person. Eysneck initially conceptualizes the personality as two dimensions: Extraversion/Introversion and Neuroticism/Stability, which define four quadrants, as shown in Fig. 6. The scores of extraversion and stability are quantized based on two parameters: ε and δ, respectively.

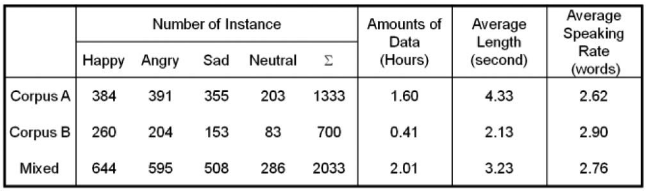
If there are three opinions in a question of the EPQ, the score can simply be zero, half, and one. According to the personality trait chart, the relationship between the emotional states and the two dimensions is shown in Table 5. The neutral emotional state has a tendency toward stable dimension, so the score of neutral emotion can be equivalent to the scale of stability, i.e., the value of parameter δ. Conversely, the angry emotion is opposite to the neutral emotion because people will lose control if they get angry. Therefore, the score of angry emotion is set to $1 - \delta$. Since sad

TABLE 5
The Relationship between
Emotion State and Two Dimensions Defined in EPQ



Emotion State	Dimension
Neutral	Stable
Happy	Stable and Extraversion
Angry	Unstable
Sad	Unstable and Introversion

TABLE 6
Description of the Collected Speech Data for Neutral, Happy, Sad, and Angry Emotions in the Evaluation Corpora



	Number of Instance					Amounts of Data (Hours)	Average Length (second)	Average Speaking Rate (words)
	Happy	Angry	Sad	Neutral	Σ			
Corpus A	384	391	355	203	1333	1.60	4.33	2.62
Corpus B	260	204	153	83	700	0.41	2.13	2.90
Mixed	644	595	508	286	2033	2.01	3.23	2.76

emotion locates in the melancholic quadrant, the score of sad emotion is set to half of the two opposite dimensions, i.e., the average of two parameters ε and δ. For the happy emotion, because extreme pleasure is followed by sorrow, the score of the happy emotion is defined as opposite to the sad emotion. So, we define the scores of the four emotional states as

$$q_N = \delta, \tag{20}$$

$$q_S = [(1-\varepsilon) + (1-\delta)]/2, \tag{21}$$

$$q_A = 1 - \delta, \tag{22}$$

$$q_H = (\varepsilon + \delta)/2, \tag{23}$$

where $0 < \varepsilon < 1$ and $0 < \delta < 1$; q_N, q_H, q_A, and q_S denote the scores of neutral, happy, angry, and sad emotions, respectively.

Considering the individual personality characteristics for personalized application, the emotion recognition output can be reestimated by incorporating the score of the individual personality trait into the probabilistic model as

$$e^* = \arg\max_{e_j \in \mathbf{E}} \{ q_j \times P_{MDT}(e_j|\mathbf{f})^{\lambda_{AP}} \times P_{MaxEnt}(e_j|EAR)^{1-\lambda_{AP}} \}, \tag{24}$$

where q_j is the score of the jth emotional state considering personality trait. Finally, the emotional state with the highest probability is determined as the recognized emotional state.

4 Experiment

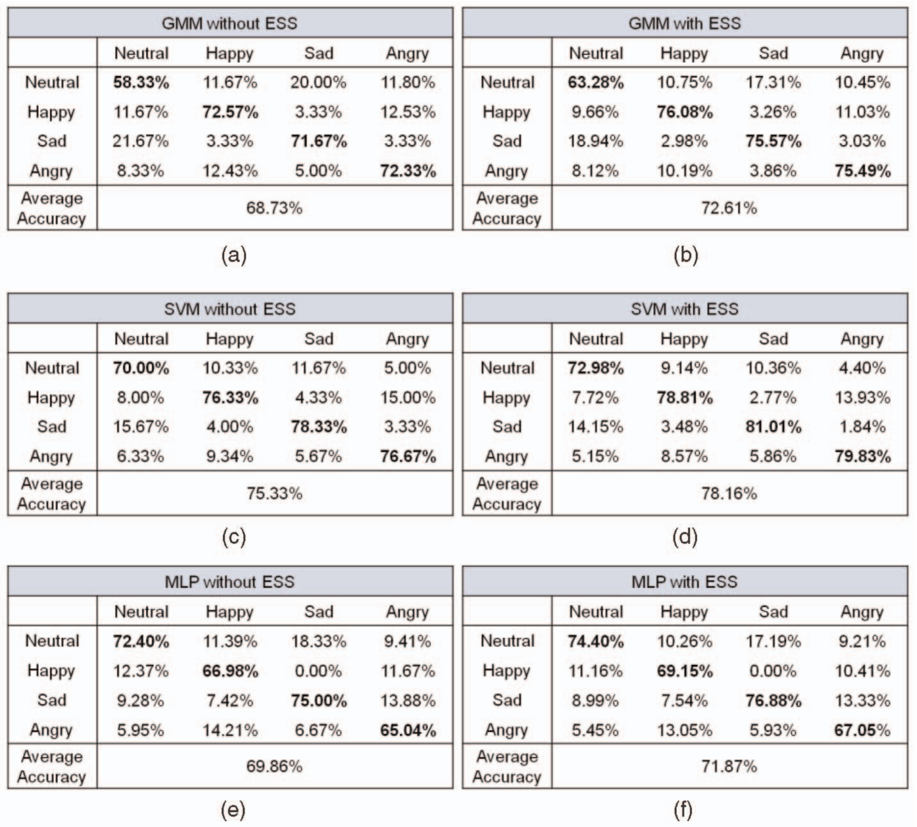
4.1 Corpora

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 the four emotional states of the subjects' daily lives are similar. For Corpus B, the subjects expressed their emotions via 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 game, respectively. In total, 2,033 sentences were collected in a lab environment using 16 KHz sampling rate and 16-bit PCM wave format. Table 6 illustrates the collected speech data for the neutral, happy, sad, and angry emotions in both corpora. The symbol Σ is used to denote the subtotal number of instances in each corpus. The row labeled as "Mixed" is used to represent a mix of two abovementioned corpora. The speaking rate of Corpus A is smaller than the speaking rate of Corpus B because some utterances were prolonged. For example, the duration of sentence "好無聊喔 (it is so boring)" can sustain more than 2 seconds. In both corpora, each sentence was manually annotated as neutral, happy, angry, and sad emotional states by the subject who provided the utterance. In order to consider the difference between perceived and intended emotions, one annotator other than the speaker who provided the utterance also annotated the evaluation data. If different annotations occurred for the same utterance, they discussed and determined the final annotation. After checking the annotation of the evaluation data, a few data of neural and sad utterances are ambiguous in speech signal but can be classified by textual transcription.

4.2 Experimental Setup

For speech recognition, an HTK-based speech recognition system was constructed [38] which contains 153 subsyllable-based acoustic models (115 right-context-dependent INITIALs and 38 context-independent FINALs), 37 particle models (e.g., "EN," "MA," and "OU"), 47 syllable-level models for hyperarticulated speech, and 13 filler models for Mandarin speech recognition. For robust acoustic model training, a read speech corpus TCC-300 was used for normal model construction, and then the collected emotional corpora were employed for MLLR-based acoustic model adaptation. The HMM topology is left-to-right without skips, with three states for each INITIAL subsyllable, five states for each FINAL subsyllable, and at most 32 mixtures in each state. The average word accuracy of the speech recognizer is 84.6 percent. In AP-based recognition, the Praat software [39] was utilized to extract the acoustic-prosodic features. The pitch features of affective speech are estimated over a frame of 80 ms at a rate of 10 ms and the pitch frequency range is from 75 to 500 HZ on the basis of an autocorrelation method described in [40], which is a part of the Praat software. The HTK was employed to extract the cepstral features—MFCCs. The cepstral features comprise 12 MFCCs plus energy, their delta, and acceleration for a frame size of 32 ms (512 samples) with a frame shift of

TABLE 7
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%			

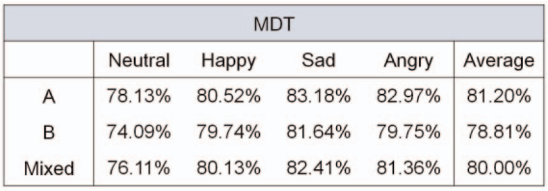
(f)

10.625 ms (170 samples). In total, 253 acoustic-prosodic features were adopted in this study. Four GMM-based classifiers, one for each emotional state, were constructed with 16 mixtures. For the SVM-based classifiers, each SVM for the jth emotional state is trained to discriminate a specific emotion from the others using the open source-LIBSVM [41]. An MLP with one hidden layer, 20 hidden nodes in the hidden layer, 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 [42]. In the following evaluations, K-fold ($K = 5$) cross validation [43] was employed to evaluate the proposed approach. In other words, the database was split into five disjoint subsets. Additionally, to decrease the effect of data bias problem, 80 percent of the instances (i.e., four subsets) of each emotional state for each speaker were randomly selected for classifier training and the remaining 20 percent of the instances were used for evaluation in each iteration.

4.3 Evaluation Results

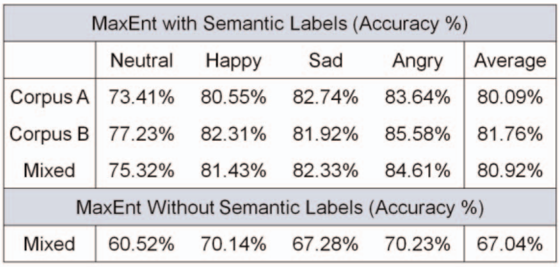
The first evaluation is the recognition performance of the three base-level classifiers with/without the ESS in AP-based recognition. For the evaluations without ESS, each segment in an utterance is employed for emotion recognition using the product of probabilities estimated from each segment. For comparison, 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 7, in which Tables 7a, 7c, and 7e were the evaluation results without ESS and Table 7b, 7d, and 7f were evaluated with ESS. The first row of each table is the base-level classifier and evaluation condition. The last row of each table is the average recognition accuracy of each base-level classifier. Moreover, the four emotional states listed in the second row are the target emotional states. Rows 3-6 are used to record the percentage amount of the test data being classified to the target emotion; for example, in Table 7a, the target emotion is “Neutral” but 21.67 percent test data was classified as “Sad.” The results of the GMM-based approach reveal that GMM can model the emotional states well except for the neutral emotion. In other words, the GMM-based classifier is not robust enough to recognize all emotional states. The SVM-based classifier can achieve 78.16 percent average accuracy of emotion recognition with the ESS and outperforms other base-level classifiers in the happy, sad, and angry emotions. Compared to the GMM-based approach using all training data, the 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 the MLP-based approach, the classifier obtained the best recognition performance on neutral emotion without/with ESS and the accuracy is 72.40 percent and 74.40 percent, respectively. The MLP-based approach is useful to process such features because each input node (i.e., each speech feature, such as intensity mean) and each hidden node contribute differently to each hidden node and each output

TABLE 8
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 9
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%

node (i.e., each emotional state). Briefly, the evaluation results of the three base-level classifiers with ESS are better than those results evaluated without ESS. The improvements of GMM-based and SVM-based are about 5 percent and 3 percent, respectively. The improvement of the MLP-based classifier is less than the others. In these evaluation results, there is confusion between neutral emotion and sad emotion because the 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 metadata (e.g., the probability distributions of the GMM-based emotional recognition), the metadata was obtained from the outputs of the three base-level classifiers using the training data set and then the test data set was used to evaluate the recognition performance of the MDT-based classifier combination. Table 8 is designed to evaluate the effect of ESS. Two corpora were evaluated independently and the results of mixed corpora show the average performance of the proposed approach. As shown in Table 8, based on the strategy of selecting a proper classifier, the recognition performance of each emotional state was improved, especially the performance on neutral emotion recognition. 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 9 shows the evaluation results of SL-based recognition with/without SLs. The average accuracy of SL-based emotion recognition with SLs is 80.92 percent against 67.04 percent for the evaluation without SLs. Since there are no salient words used for the neutral emotional state, the recognition performance is lower than other emotional states. Conversely, the sentences with angry emotion are often comprised of "intense" words. Therefore, it can achieve the best performance. For the comparison between two corpora, the evaluation results of Corpus A are 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, even though 86.1 percent speech recognition accuracy for emotional words is not good enough, the correctly recognized emotional words are still beneficial for text-based emotion recognition. Generally speaking, a text-based approach, such as semantic labels, is more reliable than a signal-based approach, such as the acoustic-prosodic-based classifier. For example, in normal utterances, the word "angry" explicitly means somebody is in an unhappy emotional state. However, high intensity value can be classified into happy (e.g., laughing) and angry emotions (e.g., roaring).

In (19), the AP-based and the SL-based recognition results are combined based on the weighting factor λ_{AP}. Fig. 7a shows the evaluation results of overall emotional state and Fig. 7b illustrates the result for each emotional state. For the evaluation of overall emotional states, the evaluation result for $\lambda_{AP} = 0.4$ can achieve the best performance of emotion recognition. So, the following evaluations will be performed based on this condition ($\lambda_{AP} = 0.4$). 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. For example, the speaker should be happy when he/she answers a question correctly; however, if he/she is unsure that his/her answer is correct or not, the emotion expression is therefore not explicit. This is one possible reason that the performance of AP-based recognition on Corpus B is lower than the evaluation result obtained from Corpus A. Conversely, the recognition performance on Corpus B is better than the result of Corpus A in SL-based recognition. Additionally, each emotional state is also evaluated with different weighted product fusion, shown in Fig. 7b. Similarly to the results shown in Fig. 7a, the weight values are highly affected by the sources with better recognition performance except angry emotion ($\lambda_{AP} = 0.5$). For angry and happy emotions, text information improved the ambiguity between speech features. For example, a high intensity value can be extracted from the utterances in happy emotion (e.g., laughing) and angry (e.g., roaring) emotion. For sad emotion, fewer keywords resulted in fewer EARs; hence, the sad emotion can outperform other emotions. Moreover, the speech features of sad emotion are different from those of happy and angry emotions so that the recognition performance is also the best one. For neutral emotion, both text information and speech features are scattered so that neutral emotion obtains the lowest performance. Table 10 is the EPQ scores of two speakers who are the two subjects in Corpus B. In this evaluation, these two subjects not only filled out the questionnaire but also identified what characteristic another speaker is. Fortunately, the results of the EPQ are similar to the identification results done by the subjects. The last evaluation is to evaluate the proposed approach as shown in Table 11. According to the result based on EPQ, speaker A is an extrovert and the recognition performance of the

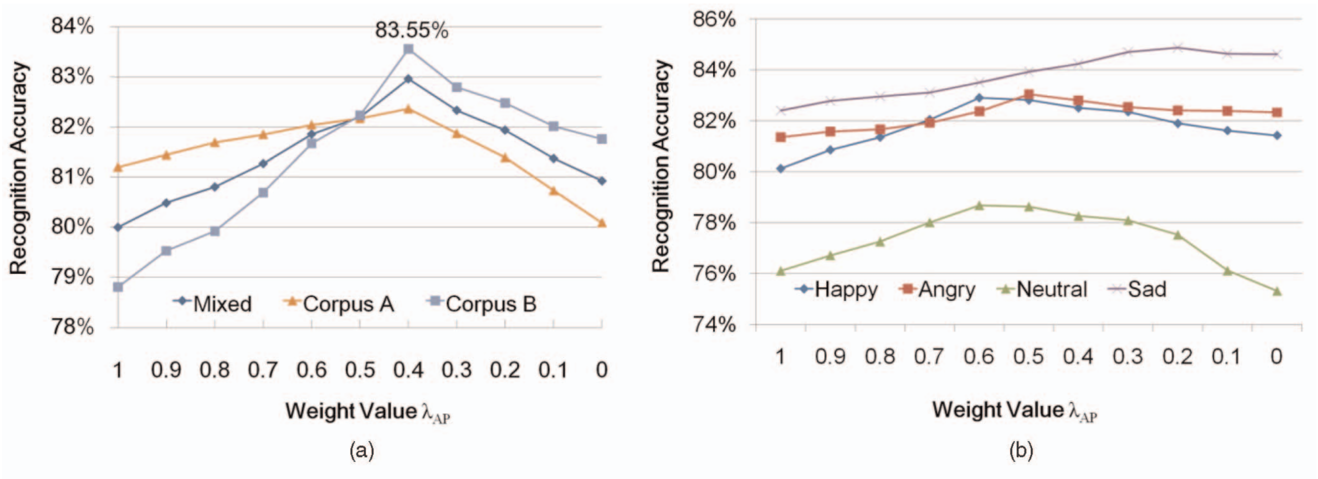


Fig. 7. Evaluation results using weighted product fusion as a function of the weight value: (a) results of overall data and separate corpora, and (b) results of each emotional state.

corresponding emotions—happy and angry emotions which have stronger expression—was improved. For speaker B, who is neither extrovert nor introvert, the difference in the evaluation results is small. Besides this evaluation, the subjects were satisfied with the fine-tuned system after they tested this system again. The evaluation of the proposed approach proved that the proposed approach can work well on the emotion recognition task.

In summary, the average recognition accuracy of this system can achieve 83.55 percent without considering personality trait, with a standard deviation of 4.61 percent. The results confirm the effectiveness of the proposed approach.

5 Conclusion

This work presents a fusion-based approach to emotion recognition of affective speech using multiple classifiers with AP and 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 the SL-based approach, the MaxEnt was utilized to model the relationship between emotional states and EARs for emotion recognition. Finally, the integrated results from the 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 can outperform each individual classifier. On the other hand, SL-based recognition obtains an average recognition accuracy of 80.92 percent. Finally, combining acoustic-prosodic information and semantic labels can achieve 83.55 percent accuracy, which is superior to the classifiers using either acoustic-prosodic information or semantic labels only. Furthermore, the proposed approach combining acoustic-prosodic information and semantic labels can achieve an average accuracy of 85.79 percent considering the personality trait of a specific speaker. Although acoustic-prosodic information and semantic labels of the neutral emotional state are ordinary (no strange energy, flat slope of the ESSs, and a few salient words), the proposed approach still obtains 78.10 percent and 81.39 percent accuracy for the neutral emotional state in corpus B without/with personality trait, respectively. Due to the imperfect recognition of affective speech, the acoustic models for hyperarticulated speech and MLLR-based model adaptation are employed to obtain more reliable recognized word sequence. Moreover, the extraction performance of the EARs may be restricted by the predefined SLs extracted from the recognized word sequence and the HowNet knowledge base. Thus, a mechanism utilized to flexibly mine the association rules is needed. On the other hand, since the acoustic-prosodic information is indistinguishable for all emotional states, the GMM-based classifier is competitive with other two complex classifiers. Finally, the experimental results show that an individual personality trait reveals a positive effect

TABLE 10
Evaluation Results of EPQ

Personality Trait		
	Speaker A	Speaker B
Stability (δ)	0.6	0.7
Extraversion (ε)	0.8	0.5

TABLE 11
Evaluation Results of AP-Based and SL-Based Emotion Recognition with Personality Trait

	MDT+MaxEnt (λ_{AP} = 0.4) (Accuracy %)				
	Neutral	Happy	Sad	Angry	Average
Speaker A	75.80%	85.97%	83.35%	87.81%	83.23%
Speaker B	80.40%	81.61%	88.49%	84.93%	83.86%
Coupus B	78.10%	83.79%	85.92%	86.37%	83.55%
	Proposed (MDT+MaxEnt+PT) (Accuracy %)				
	Neutral	Happy	Sad	Angry	Average
Speaker A	76.30%	88.79%	83.17%	89.91%	84.54%
Speaker B	86.48%	84.99%	91.49%	85.17%	87.03%
Coupus B	81.39%	86.89%	87.33%	87.54%	85.79%

on emotional recognition and can be applied to personalized applications like tutoring, entertainment, and chatting.

In this work, several types of speech features are employed for acoustic-level individual classifiers training. More speech features could be able to provide more detailed information of speech signal. However, different emotional states can be described using different types of speech features. Therefore, a proper feature selection approach is beneficial to emotion recognition [44] and can be employed in our future work.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for valuable comments that greatly improved the quality of this paper.

REFERENCES

[1] J. Liu, Y. Xu, S. Senef, and V. Zue, "CityBrowser II: A Multimodal Restaurant Guide in Mandarin," *Proc. Int'l Symp. Chinese Spoken Language Processing,* 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. Speech and Audio Processing,* vol. 13, no. 3, pp. 330-344, May 2005.

[3] N. Roy, J. Pineau, and S. Thrun, "Spoken Dialogue Management Using Probabilistic Reasoning," *Proc. Ann. Meeting Assoc. for Computational Linguistics,* pp. 93-100, 2000.

[4] D. Jurafsky, R. Ranganath, D. McFarland, "Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation," *Proc. Human Language Technologies: The 2009 Ann. Conf. North Am. Chapter of the Assoc. for Computational Linguistics,* 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. Asian Language Information Processing,* vol. 5, no. 2, pp. 165-182, June 2006.

[6] C. De Silva and P.C. NG, "Bimodal Emotion Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 332-335, 2000.

[7] R. López-Cózar, Z. Callejas, M. Kroul, J. Nouza, and J. Silovský, "Two-Level Fusion to Improve Emotion Classification in Spoken Dialogue System," *Lecture Notes in Artificial Intelligence,* pp. 617-624, Springer-Verlag, 2008.

[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," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 17-21, 2004.

[9] T. Vogt and E. André, "Exploring the Benefits of Discretization of Acoustic Features for Speech Emotion Recognition," *Proc. Int'l Speech Comm. Assoc.,* pp. 328-331, 2009.

[10] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," *Proc. Int'l Speech Comm. Assoc.,* pp. 312-315, 2009.

[11] F. Yu, E. Chang, Y.-Q. Xu, and H.-Y. Shum, "Emotion Detection from Speech to Enrich Multimedia Content," *Proc. IEEE Pacific-Rim Conf. Multimedia,* pp. 500-557, 2001.

[12] N. Amir, S. Ziv, and R. Cohen, "Characteristics of Authentic Anger in Hebrew Speech," *Proc. European Conf. Speech Comm. and Technology,* pp. 713-716, 2003.

[13] C.-M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Trans. Speech and Audio Processing,* vol. 13, no. 2, pp. 293-303, Mar. 2005.

[14] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion Detection in Task-Oriented Spoken Dialogues," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 549-552, 2003.

[15] I. Luengo and E. Navas, and I. Hernáez, "Combining Spectral and Prosodic Information for Emotion Recognition in the Interspeech 2009 Emotion Challenge," *Proc. Int'l Speech Comm. Assoc.,* pp. 332-335, 2009.

[16] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion Recognition Using Hierarchical Binary Decision Tree Approach," *Proc. Int'l Speech Comm. Assoc.,* pp. 320-323, 2009.

[17] L. Todorovski and S. Dzeroski, "Combining Classifiers with Meta Decision Trees," *Machine Learning,* vol. 50, no. 3, pp. 223-249, 2003.

[18] A. Terracciano, M.S. Merritt, A.B. Zonderman, and M.K. Evans, "Personality Traits and Sex Differences in Emotion Recognition among African Americans and Caucasians," *Ann. New York Academy of Sciences,* vol. 1000, pp. 309-312, Dec. 2003.

[19] H.J. Eysenck and S.B.G. Eysenck, *Manual of the Eysenck Personality Questionnaire.* Hodder and Stoughton, 1975.

[20] Z. Dong and Q. Dong, HowNet, http://www.keenage.com/, 2010.

[21] 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.

[22] M. Slaney and G. McRoberts, "A Recognition System for Affective Vocalization," *Speech Comm.,* vol. 39, pp. 367-384, 2003.

[23] A. Paeschke and W. Sendlmeier, "Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements," *Proc. Int'l Speech Comm. Assoc. Tutorial and Research Workshop Speech and Emotion,* pp. 75-80, 2000.

[24] T. Nwe, S. Foo, and L. De Silva, "Speech Emotion Recognition Using Hidden Markov Models," *Speech Comm.,* vol. 41, no. 4, pp. 603-623, 2003.

[25] C.-H. Wu and Z.-J. Chuang, "Emotion Recognition from Speech Using IG-Based Feature Compensation," *Int'l J. Computational Linguistics and Chinese Language Processing,* vol. 12, no. 1, pp. 65-78, 2007.

[26] X. Huang, A. Acero, and H.-W. Hon, "Prosody," *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development,* first ed., ch. 15, Section 15.4.4, pp. 753-755, Prentice Hall PTR, 2005.

[27] C.-H. Wu and J.-H. Chen, "Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis," *Speech Comm.,* vol. 35, pp. 219-237, 2001.

[28] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Comm.,* vol. 32, nos. 1/2, pp. 124-154, 2000.

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

[30] J.C. Platt, "Probabilities for SV Machines," *Advances in Large Margin Classifiers,* pp. 61-74, MIT Press, 2000.

[31] V. Petrushin, "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application," *Proc. Int'l Conf. Spoken Language Processing,* pp. 222-225, 2000.

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

[33] H. Soltau and A. Waibel, "Acoustic Models for Hyperarticulated Speech," *Proc. Int'l Conf. Spoken Language Processing,* 2000.

[34] R.S. Lazarus and B.N. Lazarus, *Passion and Reason: Making Sense of Our Emotions.* Oxford Univ. Press, 1996.

[35] R. Agrawal, T. Imielinski, and A.N. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proc. ACM SIGMOD,* pp. 207-216, 1993.

[36] Similarminds.com, *Personality Test,* http://similarminds.com/cgi-bin/eysenck.pl, 2010.

[37] Temperament: A Brief Survey, with Modern Applications, http://intraspec.ca/temper0.php, 2010.

[38] 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.* Cambridge Univ. Press, http://htk.eng.cam.ac.uk/, 2010.

[39] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer (Version 5.1.05), http://www. praat.org/, 2010.

[40] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound," *Proc. Inst. of Phonetic Sciences,* vol. 17, pp. 97-110, 1993.

[41] C.-C. Chang and C.-J. Lin, LIBSVM—A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2010.

[42] Z. Le, Maximum Entropy Modeling Toolkit for Python and C++, http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html, 2010.

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

[44] J. Tao and T. Tan, "Emotion Perception and Recognition from Speech" *Affective Information Processing,* ch. 6, Section 6.2, pp. 96-97, Springer, 2009.

Chung-Hsien Wu received the BS degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the MS and PhD degrees in electrical engineering from National Cheng Kung University, Tainan, Taiwan, Republic of China, in 1987 and 1991, respectively. Since August 1991, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University. He became a professor and distinguished professor in August 1997 and August 2004, respectively. From 1999 to 2002, he served as the chairman in the department. Currently, he is the deputy dean in the College of Electrical Engineering and Computer Science, National Cheng Kung University. He also worked at the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, Cambridge, in summer 2003 as a visiting scientist. He was the editor-in-chief for the *International Journal of Computational Linguistics and Chinese Language Processing* from 2004 to 2008. He serves as the guest editor of the *ACM Transactions on Asian Language Information Processing*, *IEEE Transactions on Audio, Speech and Language Processing* and *EURASIP Journal on Audio, Speech, and Music Processing* in 2008-2009. He is currently an associate editor of the *IEEE Transactions on Affective Computing*, *IEEE Transactions on Audio, Speech and Language Processing*, *ACM Transactions on Asian Language Information Processing*, and the subject editor on information engineering of the *Journal of the Chinese Institute of Engineers* (*JCIE*). His research interests include speech recognition, text-to-speech, and spoken language processing. He has been the president of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) since September 2009. He is a senior member of the IEEE and a member of the International Speech Communication Association (ISCA).

Wei-Bin Liang received the BS and MS degrees in computer science and information engineering from I-Shou University, Kaohsiung, Taiwan, in 2001 and 2003, respectively. Currently he is working toward the PhD degree in the Department of Computer Science and Information Engineering, National Cheng Kung University. His research interests include speech recognition, spoken language processing, and affective computing.