EMOTION RECOGNITION VIA ACOUSTIC FEATURES AND SEMANTIC CONTENTS IN SPEECH

Ze-Jing CHUANG, Chung-Hsien WU

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan
bala@csie.ncku.edu.tw, chwu@csie.ncku.edu.tw

ABSTRACT
Recent researches into human-machine communication make more emphasis on the recognition of nonverbal information, especially on the topic of emotional reaction. Many kinds of physiological characteristics are used to extract emotions, such as voice, facial expression, hand gesture, body movement, even heartbeat and blood pressure. In this paper, based on the idea that humans are capable of detecting human emotions through speech input without other visual or physiological information, an emotion recognition system that can detect the emotion from acoustic features and semantic contents in speech is proposed. In this approach, the acoustic features are extracted for feature-based emotion extraction. On the other hand, the speech signal is also fed to a speech recognizer and the recognized contents are then used for content-based emotion extraction. Finally, the integration of the results from acoustic features and semantic contents is used to determine the final emotion.

1. INTRODUCTION
In the past years, there are several research works in this topic. Some focused on a bimodal emotion extraction approach to increase the recognition accuracy. For example, Cohn and Katz [1] developed a semi-automated method on emotion extraction from face and voice. Silva [2] used HMM structure to recognize emotion from both video and audio. Yoshitomi, et al [3] combined HMM and NN structures to extract emotion from speech and facial expression. Other researchers focused on extracting emotion from speech data only. Fukuda and Kostov [4] applied a Wavelet/Cepstrum based software tools for emotion extraction from speech. Yu, et al [5] developed an SVM-based emotion extraction system. However, few papers focused on the emotion extraction from textual input only. In [6], we have developed a semantic network for emotion extraction from textual content.

In this paper, an emotion recognition system is proposed to classify six basic emotions including anger, happiness, sadness, surprise, antipathy, and fear. The proposed emotion recognition system can detect the emotion from acoustic features and semantic contents in speech. There are three steps in recognition via acoustic features: corpus collection, feature extraction and selection, and model training. In corpus collection, we adopt a broadcast drama with speech and textual contents as our corpus instead of artificial emotional speech. In feature selection, an initial feature set that contains totally 33 features was firstly analyzed and adopted. These speech features contain several possible aspects, such as intonation, timbre, acoustics, tempo, and rhythm. We also extract some features to represent the special intonations, such as trembly speech, unvoiced speech, and crying speech. Among these diverse features, the most significant features are selected by the principle component analysis (PCA).

Finally, the support vector machine model (SVM) [7] is adopted to construct the emotion recognition system.

In the recognition via semantic contents, due to the diversity of natural language, it is more difficult to recognize the emotions from the textual information than from other kinds of information. The main problem is that the emotion representation of a single sentence is ambiguous and context-dependent. In order to handle the complexity of natural language, we propose an emotional semantic network in this system. The semantic network is composed of two sub-networks: a static semantic network and a dynamic semantic network. The static semantic network is established from an existing Chinese knowledge base called HowNet and used to estimate the emotion trigger value of each word. The dynamic semantic network accepts the textual input and dynamically constructs the nodes and links, which represent the emotion carrier and the emotion propagator respectively. Initiated by the emotion trigger value, the emotion in the dynamic semantic network will propagate and finally converge to the final emotion output. Finally, the output of these two systems is integrated to give the final emotion output.

2. FEATURE-BASED EMOTION EXTRACTION
The feature-based emotion extraction system consists of three components: feature extraction, PCA, and subspace SVM. The diagram of this system is shown in Figure 1:

![Diagram of feature-based emotion extraction](image)

2.1. Feature extraction
In our approach, four basic acoustic features: pitch, energy, formant 1 (F1), and zero crossing rate (ZCR) are firstly estimated. A short-time processing technique is applied to obtain the contours of four features. For these features, what we really concern is the tendency of their contours. The tendency of a
contour can be represented by its mean, slope, or difference of slope. A Legendre polynomial technique is adopted to represent the contours of these four features.

According to previous research, emotion expression is strongly related to the pitch and energy of speech signal. For this result, we adopt several parameters that are based on the pitch and energy. Finally, we extract the following 13 kinds of acoustic features:

1. 4th-order Legendre parameters for pitch contour
2. 4th-order Legendre parameters for energy contour
3. 4th-order Legendre parameters for F1 contour
4. 4th-order Legendre parameters for ZCR contour
5. Maximum of energy
6. Maximum of smoothed energy
7. Minimum, median, and standard deviation of pitch contour
8. Minimum, median, and standard deviation of energy contour
9. Minimum, median, and standard deviation of smoothed pitch contour
10. Minimum, median, and standard deviation of smoothed energy contour
11. Ratio of up-slope sample number to down-slope sample number for pitch contour
12. Ratio of up-slope sample number to down-slope sample number for energy contour
13. Pitch vibration:
   The pitch vibration is defined and calculated in Equation (1):

   \[ P_v = \frac{\sum_{i=0}^{N-1} \delta[(P(i) - \bar{P}) \times (P(i+1) - \bar{P})]}{N}, \quad \delta[x] = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \]  

where \( \bar{P} \) is the mean value of pitch contour.

2.2. Principle component analysis

Principal component analysis (PCA) is a statistical approach that can extract the main components from a lot of variables. In the previous section, an initial feature set that contains 33 features was firstly extracted. The dimension of parameters can be reduced by selecting an appropriate percentage of principle components. In the further analysis, the influence of an original parameter \( Z_i \) to a new principle component \( Y_j \) is the \( j \)-th element in the transformation matrix \( \mathbf{q}_{ij} \). According to this value we can select the most important original features in a new principle component.

In order to magnify the importance of the features, we classify the features into several feature subspaces by PCA. Firstly, the important principle components are selected according to a threshold of 90%. Secondly, for each principle component, the significant features of this component form a feature subspace. We use a threshold of 0.2 to decide the significance of features. Finally, we have a feature set of 14 sub-spaces.

2.3. Support vector machine

Support vector machine (SVM) has been widely applied in many research topics such as data mining, pattern recognition, linear regression, data clustering, and so on.

Given a set of data that belong to two classes, the basic idea of SVM is to find out a hyperplane that can completely separate the different classes. The hyperplane is decided by the maximal margin of two classes, and the samples that lie on the margin are what we called “support vector.” The equation of hyperplane is described in Equation (2).

\[ D(x) = \sum_{i=1}^{N} \alpha_i y_i (x \cdot x_i) + w_0 \]  

In this paper, we apply the modified SVM that classifies the input data in a space and produce a continuous probability for emotion extraction. Given the test sample \( x' \), the probability that \( x' \) belong to class \( c \) is \( P(\text{class},x') \). The value should relate to the following factors:

- The distance between testing input and the hyperplane:
  \[ R = \frac{D(x')}{\|\mathbf{x}\|} = D(x') \]  

- The distance from class centroid to the hyperplane:
  \[ R' = \frac{R}{D(\mathbf{x})} = D(x') \]  

where \( \mathbf{x} \) is the centroid of the training data in a class.

- The classification confidence of the class.
  The classification accuracy calculated from the training data is used to define the classification confidence of class \( c \):
  \[ P_c = \frac{\text{Number of correctly recognized classes}}{\text{Total data number}} \]  

Finally, the output probability is defined according to the above factors:

\[ P(\text{class},x') = \frac{P_c}{1 + \exp(1 - R')} = \frac{P_c}{1 + \exp \left( 1 - \frac{D(x')}{D(\mathbf{x})} \right)} \]  

As the previous description, the original feature set is divided into 14 feature sub-spaces. For each sub-space, an SVM model is applied to decide the best class of the input test sample. And the final output is the combination of these 14 SVM outputs and shown in Equation (7):

\[ P(\text{class},x') = \prod_{i=1}^{N} P(\text{class},x')^{\frac{1}{S}} \]  

\[ = \prod_{i=1}^{N} \left( 1 + \exp \left( 1 - \frac{D(x')}{D(\mathbf{x})} \right) \right)^{\frac{1}{S}} \]  

where the probability \( P(\text{class},x') \) is the output of SVM in the \( i \)-th feature subspace, and \( S \) is the number of sub-spaces. In this case, \( S \) is 14.

3. CONTENT-BASED EMOTION EXTRACTION

Figure 2 shows the diagram of content-based emotion extraction system. In this system, the textual content of input utterance is processed as a semantic network structure. The emotion contained in the sentence will be propagated in this network. The whole process includes three components: preprocessing, static semantic network, and dynamic semantic network.

3.1. Speech recognizer

The hidden Markov models (HMM) with continuous density function are adopted for speech recognition. Mel-frequency cepstrum coefficients (MFCC) are selected as the speech features. Since the semantic network is constructed according to the
emotional keywords and other significant POSs, a keyword spotting system is sufficient for our purpose.

![Diagram of content-based emotion extraction](image)

**3.2. Preprocessing**

Since we are only interested in the emotion representation of a sentence, a parser is designed to focus on the special syntactic structures useful for emotion recognition. There are three steps in the parsing process: (a) word segmentation, (b) POS to syntactic constituent mapping, and (c) grammatical rule mapping.

In order to disambiguate the syntactic constituent of each word, we induce the mapping from POSs to syntactic constituents, which is shown in Figure 3.

![Mapping from POSs to syntactic constituents](image)

**3.3. Static semantic network**

**3.3.1. Emotion trigger value**

In order to quantify the emotional effect of different kinds of verbs, we define an emotion trigger value \( T_D \). In general, the positive verb increases the positive emotion and decreases the negative emotion, while the negative verb decreases the positive emotion and increases the negative emotion. The classification result is listed in Table 1.

<table>
<thead>
<tr>
<th>Group</th>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Happiness, Anger, Surprise</td>
</tr>
<tr>
<td>Negative</td>
<td>Sadness, Fear, Antipathy</td>
</tr>
</tbody>
</table>

**3.3.2. Emotion trigger value generation using HowNet**

We tagged the emotion trigger value of each definition manually and calculated the emotion trigger values of all concepts according to the relationships defined in HowNet. The emotion trigger value is estimated as follows:

If a concept \( C \) is defined by \( k \) definitions \( (D_1, D_2, \ldots, D_k) \) with \( k \) relationships \( (r_i) \), shown in Equation (8):

\[
DEF = r_1D_1, r_2D_2, \ldots, r_kD_k
\]

The emotion trigger value \( T_C \) for concept \( C \) is calculated as:

\[
T_C = \frac{1}{k} \sum_{i=1}^{k} F(r_i, T_D)
\]

where the function \( F() \) in Equation (9) is used to transform the emotion trigger value from definitions to concepts. For each relationship, the transformation function is defined and listed in Table 2.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Functions</th>
<th>Symbols</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONE</td>
<td>0.5 ( T_D )</td>
<td>$S$</td>
<td>$T_D$</td>
</tr>
<tr>
<td>%</td>
<td>( s \times (2^{l_1} - 1) )</td>
<td>@</td>
<td>0</td>
</tr>
<tr>
<td>&amp;</td>
<td>( s \times (2^{l_1} - 1) )</td>
<td>#</td>
<td>0.2 ( T_D )</td>
</tr>
<tr>
<td>?</td>
<td>( s \times (2^{l_1} - 1) )</td>
<td>^</td>
<td>(- T_D )</td>
</tr>
<tr>
<td>*</td>
<td>( s \times \log_{10}(9 \times</td>
<td>T_D</td>
<td>+ 1) )</td>
</tr>
</tbody>
</table>

**3.4. Dynamic semantic network**

**3.4.1. Network construction**

The primary components in the dynamic semantic network are nodes and links. The nodes (emotion carrier) can carry the emotional information while the links (emotion propagator) can propagate the emotional information. There are five components in a node: word name, trigger value acceptor \( (T_C) \), initial emotion vector \( (E_0) \), emotion propagation value \( (E_p) \), and emotion vector \( (E_l) \). There are also four types of links defined and shown in Figure 4.

![Link types in dynamic semantic network](image)

**3.4.2. Emotion propagation**

The propagation of emotions is triggered by a new input sentence and ends when the emotion propagation value is below a threshold. In the emotion propagation process, the sources of the propagation are firstly decided according to three criteria as follows:

(a) The pre-tagged emotional keywords
(b) The pre-tagged predicates in the broadcast drama
(c) The words with emotion trigger value.

According to the above criterion, we define the decay function as follows:

\[
E_p = D(E_{s-1}) = \begin{cases} E_{s-1} & \delta(l) \times W(E_{s-1} + 0.5g \times T_C) \text{, equal link} \\ \delta(l) \times W(E_{s-1}) & \text{, other wise} \end{cases}
\]

\[
\delta(l) = \exp(-0.05 \times l^2)
\]

\[
W(x) = \begin{cases} 1, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}
\]
where \( E_{t-1} \) denotes the emotion vector calculated in the last propagation step. \( l \) is the number of propagation steps from the propagation source, and \( g \) is the sign of emotion vector \( E_{t-1} \). Functions \( \delta() \) and \( W() \) limit the emotion propagation ranging from -1 to +1 and decaying to zero after ten links.

### 3.4.3. Emotion history

The current emotion vector was calculated recursively from \( E_p \) and emotion history \( E_{t-1} \). The emotion history calculated in Equation (11) is used to remain the continuity of the emotion.

\[
E_0 = E_i \\
E_t = E_{t-1} + 0.5\{E_p - E_{t-1}\} \tag{11}
\]

where \( E_t \) indicates the emotion reaction for the \( t \)-th input sentence. \( E_i \) and \( E_p \) indicate the initial emotion vector and the emotion propagation value respectively. The final emotion reaction is the average of the values of \( E_t \) over all nodes in the dynamic semantic network.

### 4. SYSTEM INTEGRATION

So far the emotion output can be received from both speech and semantic module. The final output shown in Equation (12) is a linear combination of these two results.

\[
E_{\text{final}} = (1 - \eta)E_{\text{semantic}} + \eta E_{\text{speech}} \tag{12}
\]

where the variable \( \eta \) is a weighting factor. In this system, we use the classification accuracy of SVM as the weighting factor; it is defined in Equation (5).

### 5. EXPERIMENTAL RESULTS

In order to obtain the natural emotion reaction, we collected the training corpus from a broadcast drama. There are 558 sentences contained in 137 dialogues from the leading man and 453 sentences contained in 136 dialogues from the leading woman. We tagged the emotion reaction of each sentence manually.

In the broadcast drama, the numbers of the sentences for the six emotions are not equal. We selected 80\% of the sentences for inside test and the other 20\% for outside test. The following table shows the recognition results.

The recognition results for inside test achieved an accuracy of 87\%. Table 3 shows the result for outside test. The result is calculated over all input sentences. This result shows that the accuracy is proportional to the number of sentences. The results for content-based and feature-based systems achieved the accuracy of about 55.92\% and 63.35\%, respectively.

Table 4 shows the results for the integrated system from content-based and feature-based systems. About 2\% improvement compared to the feature-based system was obtained for outside test.

### 6. CONCLUSION

In this paper, an emotion recognition system with speech input is proposed. The emotion is extracted from both acoustic features and semantic content. The acoustic features are reduced and segmented into several feature sub-spaces by PCA. The SVM model is applied to construct the emotion classifier based on the speech features. The semantic content is analyzed by an emotional semantic network, which consists of a static semantic network and a dynamic semantic network. The static semantic network is constructed from HowNet to generate the emotion trigger value. The dynamic semantic network is constructed from the analysis of Chinese grammatical rules and input sentences. The integration of these two systems achieved a promising improvement compared to individual systems. In the future, this approach is hopefully to be integrated with other physiological characteristics to achieve more robust emotion recognition.

### 7. REFERENCE


