Emotion Verification for Emotion Detection and Unknown Emotion Rejection

Hoon-Young Cho, Kaisheng Yao, and Te-Won Lee

Institute for Neural Computation
University of California, San Diego, USA
{hcho,kyao,tewon}@ucsd.edu

Abstract

This paper focuses on detection of a single emotion and verification of a specific emotion type in a test utterance. To utilize a probabilistic output of a classifier as well as to exploit various long term acoustic features, we built a probabilistic output SVM and applied several approximated log likelihood ratio tests for emotion verification. Experimental results on SUSAS and AIBO emotion database show that anger and sadness are easier emotions to be detected than boredom and happiness. Results also verify the efficacy of applying log likelihood ratio with respect to neutral emotion as a measure for emotion verification.

1. Introduction

Among several different information conveyed in a speech signal, affective state of a speaker is being studied actively. Several major issues in this area include definition of emotions, searching an effective feature set for emotion classification, design of emotion classifiers, and construction of more authentic databases [1].

Most of the recent studies on automatic emotion recognition deal with four to seven prominent emotions, e.g., anger, sadness, boredom, joy, and neutral emotion. Emotional features are derived from pitch and energy trajectories, duration, and spectrum features. Many studies have revealed that pitch and energy are the most useful features for emotion recognition. However, it is still an open question to find conclusively a set of effective features. In the recognition end, several classifiers such as hidden Markov models, support vector machines, neural networks, fuzzy inferences have been considered. The accuracies of state-of-the-art emotion recognition systems are around 60% for four emotion classes. The performance is dropped dramatically as the number of emotion classes is increased.

Instead of considering several emotions whose current performance is far below from being applicable, some studies have focused on distinguishing anger versus neutral emotion and has shown over 90% accuracy [3, 4, 2]. This can be directly applied to an automatic call routing of angry caller to human agents for interactive voice response (IVR) systems or to search an angry emotional speech utterance from a large audio archive [3].

These studies motivated our research on emotion verification. In the emotion verification task, we are interested in detection of a single emotion. All other suspicious emotion are to be rejected and those utterances with high confidence scores are accepted. We calculate the log likelihood ratio (LLR) of a specific emotion \( e \) and its complementary emotion \( \bar{e} \). By comparing this LLR value of a test utterance with a predetermined threshold, a decision of whether to accept or to reject an utterance is made. Since it may not be feasible to obtain all data for the complementary emotion \( \bar{e} \), we propose three different methods to obtain an approximated model for \( \bar{e} \).

Also, it seems obvious that a set of effective acoustic features varies according to emotion categories. For example, in verification of angry emotion, pitch and energy information might be effective. However, for verification of other emotions, the role of durational features might increase. Therefore, selection of relevant features are necessary for emotion verification as well as for emotion classification.

2. Feature extraction

This section describes emotional features used in this study. Three feature selection methods are considered.

2.1. Basic emotional features

We extract 19 basic emotional features from energy, pitch, duration, and spectrum. For energy based features, we calculated logarithmic energy from each frame and determined a speech duration. From the energy trajectory and its delta trajectory, we obtained minimum, maximum, average, and range of energy of the entire speech interval.

We used a Viterbi search based pitch tracking algorithm. From the estimated pitch trajectory and its delta trajectory, we calculated minimum, maximum, average, and range of pitches and delta pitches.

After the pitch estimation, we segmented the whole speech interval as voiced and unvoiced segments. Let us denote the pitch period of \( i \) th frame by \( p(i) \) and the autocorrelation of the same frame by \( r(i, \tau) \), where \( \tau \) is a lag value. The fundamentalness \( fn(i) \) is calculated as follows.

\[
fn(i) = \frac{r(i, p(i))}{r(0)} \quad (1)
\]

As \( i \) th frame is more voiced, the value of numerator may

\[
\text{The fundamentalness} \quad fn(i) \quad \text{is calculated as follows.}
\]

\[
fn(i) = \frac{r(i, p(i))}{r(0)} \quad (1)
\]

As \( i \) th frame is more voiced, the value of numerator may
be high. Therefore, \( fn(i) \) may become larger. We compare
\( fn(i) \) to a threshold and mark the frame as voiced or un-
voiced to obtain durational features. The length of voiced
segment is represented by dividing the number of voiced
frames by the number of whole speech frames and the length
of unvoiced part is calculated in the same way.

For spectral slope feature, the energy of a sub-band range
2000 ~ 4000 Hz was divided by that of a sub-band 0 ~ 1000
Hz. Total dimension of basic features is 19.

2.2. Additional features

Other than the basic features, acceleration features are ex-
tracted from the energy and the pitch trajectory. Also, two
standard deviations corresponding to pitch and energy se-
quences are added and 13 dimensional averaged MFCCs are
calculated from the voiced segment. The number of addi-
tional features are 23.

2.3. Feature selection

From a combination of 42 basic and additional features, we
try to reduce the dimensionality to 19 and compare the per-
formance with the basic features. Three feature selection
methods are considered: 1) The Pearson correlation coeffi-
cient (PCC) is defined as

\[
R(i) = \frac{\text{cov}(X_i, Y)}{\sqrt{\text{var}(X_i) \cdot \text{var}(Y)}},
\]

(2)

where \( X_i \) and \( Y \) are \( i \)-th dimensional feature variable and its
corresponding class label, respectively. \( R(i)^2 \) is used as a
variable ranking criterion [5]. 2) The single variable classi-
fiers (SVC) builds a classifier for each dimension and mea-
sures their individual predictive power [5]. 3) The SVM-RFE
ranking criterion for a given variable \( i \) is

\[
R(i) = \left| \| w \|^2 - \| w^{(i)} \|^2 \right|, \quad w = \sum_{k=1}^{m} \alpha_k \Phi(x_k),
\]

(3)

where \( \alpha_k \) and \( x_k \) denote Lagrangian multipliers and support
vectors, respectively. \( \Phi(x) \) is a nonlinear mapping function.
Calculation of \( \| w^{(i)} \|^2 \) is the same as that of \( \| w \|^2 \) except
that \( i \)-th feature is not considered. Higher \( R(i) \) value means
higher ranking [6].

3. Probabilistic Output SVM

We used support vector machine (SVM) for emotion recog-
nition, since it can easily incorporate long term acoustic fea-
tures. Long term features can be easily added and evaluated
in the framework of SVM. This may be an advantage of us-
ing SVM over hidden Markov Model (HMM) which requires
a sequence of short term features. Let us describe the output
of SVM by

\[
f(x) = \text{sgn}(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b),
\]

(4)

where \( x \) is an input feature, \( l \) is a number of support vec-
tors, and \( K \) is a kernel function. The output of conventional
SVM is a label of classes and it does not provide any likely-
hood information about the output label. With an additional
sigmoid function, \( f(x) \) are mapped to probabilities. If we
denote \( f(x) \) in equation 4 by \( f_i \), the probability is approximated by

\[
p_i = \frac{1}{1 + \exp(A f_i + B)},
\]

(5)

where values of \( A \) and \( B \) are fit by maximum likelihood es-
timation from training data [8].

The modified SVM explained so far have been on a bi-
ary class case. We further extend the method to a mul-
tiple class case. Assume that pairwise class probabilities
\( r_{ij} = p(y = i | y = i \lor j, x) \) are obtained by using sig-
moid approximations on SVM outputs. The goal is to esti-
mate \( p_i = p(y = i | x), i = 1, \cdots, k \), where \( k \) is the number of
classes. In [9], from the following equation,

\[
\sum_{j:j \neq i} p(y = i \lor j | x) - (k - 1) p(y = i | x) = \sum_{j=1}^{k} p(y = j | x) = 1,
\]

(6)

together with

\[
r_{ij} \approx \frac{p(y = i | x)}{p(y = i \lor j | x)},
\]

(7)

they obtained

\[
p_i \approx \frac{1}{\sum_{j:j \neq i} 1/r_{ij} - (k - 2)}
\]

(8)

We have applied the sigmoid training to SVM and extended it
to a multi-class case using the above mentioned method [9].

4. Emotion Verification

It may be useful to tell whether a given speech is accepted
as a specific emotion of our interest. As in speaker verifica-
tion, many different kinds of verification functions can be
considered. We propose a hypothesis test based method. For
a feature vector \( x \), the log likelihood ratio of a specific emo-
tion class \( C_r \), and a set of all the other emotions \( C_r \) can be
defined as

\[
V(x) = \log \frac{Pr(x|C_r)}{Pr(x|C_{\bar{r}})} | \bar{r} = 1,
\]

(9)

where the probability is calculated by a certain type of clas-
sifier such as hidden Markov model or probabilistic output
SVM. Since it is impossible to collect all different kinds of
emotion data for \( C_r \), the denominator term should be approx-
imated. One possible way of approximation is using neutral
emotion instead of \( C_r \) as,

\[
V(x) = \log \frac{Pr(x|C_r)}{Pr(x|C_{\bar{r}})}
\]

(10)
In figure 1, the solid line shows the distribution of $V$ values when the label of $x$ is anger and the dotted line is that of $V$ when labels of $x$ are from one of the four different speaking styles, i.e., neutral, fast, question, and slow speaking styles. The values of $V$ are generally higher for angry utterances than other emotions. Therefore, we can compare values of $V(x)$ with a threshold $\theta$ and decide whether we accept or reject the emotion recognition result. A high threshold makes it difficult for other emotional speech to be accepted, but at the price of falsely rejecting valid emotional speech.

The second way of obtaining a model for $C_e$ is to collect all the other emotion data as much as possible and to train one probabilistic model for $C_e$.

The third method is to collect data for $N - 1$ different emotions which belong to the complimentary emotion class $\bar{C}$. Then, learn probabilistic model for each class and approximate the calculation of $Pr(C_e|x)$ by

$$V(x) \approx \log Pr(x|C_e) - \log \left( \frac{1}{N-1} \sum_{i \neq e} \exp(\gamma \cdot \log Pr(x|C_i)) \right)^{1/\gamma}$$

(11)

where $\gamma \in R^+$. This method has been used as a confidence measure in the area of utterance verification [7].

So far, we have dealt with the verification of a single emotion. One can further extend this to cases of multiple emotions. Let us assume that an emotion classifier which discriminates $N$ different emotions gives us an emotion label for an arbitrary utterance. Among test utterances, some might have been generated by other emotion states than the $N$ emotion category and should be rejected. For a feature vector $x$, the log probability ratio between the registered emotion classes and unknown emotion classes is defined as

$$V(x) = \log \frac{Pr(x|C_e)}{Pr(x|C_\bar{e})} \quad e = 1, 2, \ldots, N$$

(12)

where the denominator, $Pr(x|C_\bar{e})$, can be approximated similarly as equation 11.

5. Experimental Results and Discussion

We conducted several experiments on emotion verification, feature selection, and emotion recognition. Description on emotion databases is followed by discussions on experimental results.

5.1. Databases

We used SUSAS and AIBO database for the experiments. The SUSAS database was originally designed for robust speech recognition under stressed conditions. The database contains 35 words uttered by 9 speakers for 11 different speaking style including angry, loud, Lombard, fast, slow, and neutral, etc. The AIBO emotional database by Sony entertainment robot consist of German speech and is sampled at 16 kHz [4, 10]. The set of emotions included in the database are angry, bored, happy, neutral, and sad. The emotional expressions are not exaggerated and therefore, the classification task is more difficult than SUSAS database [4].

Table 1: Comparison of feature selection methods on five speaking styles of SUSAS database

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCC(19)</td>
<td>68.2</td>
</tr>
<tr>
<td>SVC(19)</td>
<td>63.5</td>
</tr>
<tr>
<td>SVM-RFE(19)</td>
<td>66.7</td>
</tr>
</tbody>
</table>

5.2. Feature selection and emotion recognition

We used 540 utterances of five different speaking styles such as anger, question, fast, neutral, and slow to train a SVM. For test set, 2430 utterances were used. Firstly, we used 19 basic emotional features described in section 2.1. The overall accuracy of baseline emotion classifier was 65.0%.

After combining 23 additional features to the basic features, the accuracy was 65.6%. Then, the three feature selection method in section 2.3 were applied to this 42 dimensional features. Table 1 shows classification accuracies of the three methods.

5.3. Emotion verification

For verification of anger emotion, we used the same training and test set in the previous section. Among the test set, 486 utterances are angry utterance and all others are from four different speaking styles.

In our first experiment, we compared the performance of the three approximation methods mentioned in section 4. For the first approximation, equation 10, we trained a binary SVM classifier for anger and neutral emotion. For the second approximation method, we trained an anti-model using all training utterances except for the anger utterances. For the third approximation, we trained SVM models for the five emotions and then applied them to equation 11.

Figure 2 shows the ROC curve for anger emotion verification using the three approximation methods. As the thresh-
olds decreases, the false acceptance rate increases and the false rejection of anger increases. It appears that the first approximation method, LLR of angry and neutral emotion, performs better than other methods.

In our second experiment, we used the AIBO database to find out what emotion is most feasible for emotion verification. We used 3534 utterances to train four binary SVMs for angry-neutral, bored-neutral, happy-neutral, and sad-neutral emotion pairs. The whole test data consists of 1681 utterances, which includes 413, 314, 349, 304, 301 utterances for angry, bored, happy, neutral, and sad emotion, respectively. In a test set of anger verification, 413 utterances were labeled as anger and the other 1268 utterances were labeled as nonanger. In a similar way, four verification test sets corresponding to each emotion were built. The result is summarized in figure 3. From the result, it is seen that anger and sadness are comparatively the easiest emotions for verification.

6. Conclusions

In this paper, we focused on verification of a single emotion. We proposed using a probabilistic output SVM for emotion verification and compared three different approximation methods for log likelihood ratio tests. Extension of emotion verification from a single emotion to multiple emotions is underway.

7. Acknowledgements

This work was in part supported by the DIMI-Sony grant and the Post-doctoral Fellowship Program of Korea Science and Engineering Foundation (KOSEF).

8. References


