# MOOD DETECTION FROM DAILY CONVERSATIONAL SPEECH USING DENOISING AUTOENCODER AND LSTM

*Kun-Yi Huang, Chung-Hsien Wu, Ming-Hsiang Su and Hsiang-Chi Fu*

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan

## ABSTRACT

In current studies, an extended subjective self-report method is generally used for measuring emotions. Even though it is commonly accepted that speech emotion perceived by the listener is close to the intended emotion conveyed by the speaker, research has indicated that there still remains a mismatch between them. In addition, the individuals with different personalities generally have different emotion expressions. Based on the investigation, in this study, a support vector machine (SVM)-based emotion model is first developed to detect perceived emotion from daily conversational speech. Then, a denoising autoencoder (DAE) is used to construct an emotion conversion model to characterize the relationship between the perceived emotion and the expressed emotion of the subject for a specific personality. Finally, a long short-term memory (LSTM)-based mood model is constructed to model the temporal fluctuation of speech emotions for mood detection. Experimental results show that the proposed method achieved a detection accuracy of 64.5%, improving by 5.0% compared to the HMM-based method.

***Index Terms***—Long-term emotion tracking, mood detection, denoising autoencoder, long short-term memory

## 1. INTRODUCTION

According to the World Health Organization (WHO) [1], depression will be one of the leading causes of death and disability by 2020 and will be the biggest health burden on society economically and sociologically. Recognition of emotions in text, speech and facial video plays an important role in affective computing [2-4]. For many people who struggle with negative emotion, it is a great challenge to discuss their situation with other people. Sometimes we do not even know what is happening until our negative emotions have completely rolled over us. In fact, we can relieve stress and manage our emotions with assistances such as counseling, learning emotion management and meditation. The traditional subjective self-report method used for measuring emotions generally lacks objective judgment for diagnosis. Thus, an objective emotion tracking system can complement subjective self-report measurement and assist users in managing their emotions.

In the research of affection, Robbins [5] divided the affection into two parts: emotion and mood. The emotions have more clear and specific types, such as anger, fear, sadness and happiness. Also, it is usually accompanied by distinct expressions. Different from emotion, the duration of mood is longer. The mood dimension consists of positive and negative emotions. Additionally, the cause in mood is unclear in general, and thus it is not indicated by distinct expressions. Moreover, it is commonly accepted that the speech emotion perceived by the listener is an approximation of the intended emotion conveyed by the speaker. However, [6-7] investigated the validity of the mismatch between the expression and perception of emotion. [6] used the IEMOCAP database [8] to analyze the difference between perceived emotion and expressed emotion. In addition, personality also affects a person's emotions and mood swings directly [9-10]. In affective speech computing, several databases [11-16] were employed for emotion recognition, but the databases discussed above were collected from short-term utterances. There is no emotional speech database considering personality for long-term emotion tracking and mood detection.

To summarize, this work aims to collect a long-term mood database and reduce the difference between the perceived emotion and the expressed emotion. Also, we considers the personality from different subjects, using the autoencoder-based conversion function to convert the perceived emotion into expressed emotion. Finally, the long-short term memory (LSTM) is employed to detect the mood of the subject based on long-term tracking of the converted expressed emotion.

## 2. DATABASE DESIGN AND COLLECTION

There are three databases for this work: **MHMC** emotion database, Emotion with Personality Database (**EP-DB**) and Long-Term Mood Database (**LM-DB**). The MHMC emotion database collected 896 utterances consisting of 236 angry, 199 happy, 214 sad, and 319 neutral utterances from 53 university students. The other two databases are detailed in the following sections.

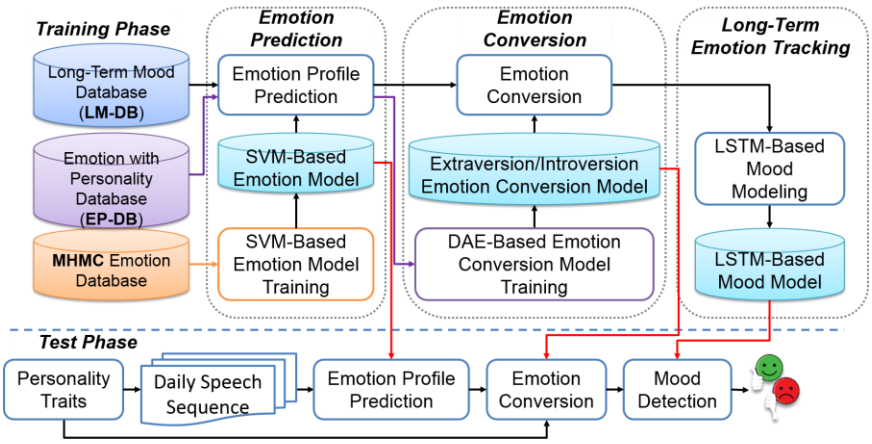


Figure 1. System framework of the proposed method.

### 2.1 Emotion with Personality Database (EP-DB)

In order to collect the database with labeled emotion and personality (extraversion and introversion), the participants were asked to complete the Big Five Inventory-10 (BFI-10) questionnaire [17] before the collection process for personality determination.

In data collection, the emotions of the participants were elicited by watching four videos consisting of the emotions of happiness, sadness, anger and neutrality. For eliciting emotional video selection, this work preselected six high-rating emotional videos [1][17] as the candidate eliciting videos. 25 students of National Cheng Kung University were invited to evaluate these six emotional videos by using questionnaires and Chi-squared test. Finally, three emotional videos passing the Chi-squared test were selected as the eliciting videos. After watching each video, two questions are played to ask the participant for response sequentially. The two questions are:

*1. What do you think about the above video?*

*2. Which scene in the movie is impressive? Why?*

The speech signals of the participants responding to the above questions were recorded to construct the Emotion with Personality Database (**EP-DB**). For emotion labeling of the **EP-DB**, the participants were asked to label their emotions as well as intensity before and after watching each video. The intensity of the emotion lies in the range [0, 5]. Zero represents no emotion expression, while five represents strong intensity of the emotion. In total, the emotion database collected the data from 21 participants, 11 introverts and 10 extroverts.

### 2.2 Long-Term Mood Database (LM-DB)

In previous studies, to the best of our knowledge, there is no speech database for mood evaluation. Since mood needs a longer time to trace and detect than emotion, this work collected the user's speech data in a day. The **LM-DB** database in this work recorded the daily conversations of a graduate student in the laboratory. Thus, the collected data are spontaneous dialogues. The participant has to complete a personality test before he/she records the speech data. After finishing the recording, he/she needs to label his/her mood as positive and negative at the end of the daily recording for constructing a long-term emotion tracking model. The time period for emotional speech recording is about two weeks. There are ten participants, made up of 6 males, 4 females, 7 introverts, 3 extraverts, with an average age of 27. In Table 1, "E" and "I" represent the personality of the participants, as extraverted and introverted, respectively.

Table 1. Basic information of the participants in the LM-DB

| Participant | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sex | M | F | F | M | M | F | M | M | M | F |
| Age | 23 | 23 | 23 | 23 | 38 | 24 | 42 | 24 | 24 | 24 |
| Personality | E | I | E | I | I | I | I | E | I | I |
| No. of days for recording | 26 | 32 | 32 | 37 | 23 | 8 | 10 | 10 | 10 | 12 |

## 3. PROPOSED METHOD

The system framework of the proposed method is shown in Fig. 1. The training phase in the proposed mechanism can be divided into three parts: emotion prediction, emotion conversion, and long-term tracking. The MHMC emotion database is used to construct the emotion prediction model by using the Support Vector Machine (SVM). In addition, the difference between the perceived emotion by the listener and the intended emotion by the speaker can be converted based on a denoising autoencoder [18] (DAE). Finally, long-term emotion tracking for mood detection can be done using the Long-Short Term Memory [19] (LSTM)-based mood model, which models the relationship between emotion trajectory and mood. In the test phase, the daily speech segment sequences are used to predict the emotion profile by the SVM. To obtain the expressed emotion from the subject, the predicted emotion will be converted by the DAE for the subject according to his/her personality. Finally, whether the subject's mood is positive or negative is

determined from the converted expressed emotion trajectory using the LSTM-based emotion tracking model.

### 3.1 DAE for Emotion Conversion with Personality

As described in previous sections, there is a difference between expressed and perceived emotions. This work attempts to obtain the expressed emotion based on an emotion conversion model and track the converted emotion sequence for mood detection. As shown in Fig. 2, $[PE_1,...,PE_i]$ is the predicted emotion profile and it can be regarded as perceived emotion. $[EE_1,...,EE_i]$ is the self-tagged emotion profile and it can be regarded as the expressed emotion profile. The emotion profile represents the posterior probabilities providing a quantitative measure for expressing the degree of the presence or absence of a set of basic emotions within an expression. This figure shows the difference between perceived emotion and expressed emotion for happiness. With the difference, the perceived emotion can be regarded as a noisy version of the expressed emotion. However, it is difficult to collect more labeled emotion data by self-annotation for different personalities. The distribution of the difference between expressed and perceived emotions is characterized by a Gaussian distribution and thus could be used to generate more noise to represent the differences for model training. Noise expressed emotion intensity is thus regarded as the new perceived emotion intensity for DAE training.

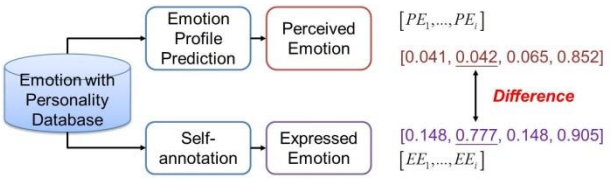


Figure 2. Difference between expressed and perceived emotions

Since there is difference between the predicted emotion profile and the self-tagged emotion profile, this work normalizes the self-tagged emotion profile in order to have the same scale with the detected emotion profile. First, the intensity of each emotion for each segment in the EP-DB is evaluated by the participants using the scale of [0, 5]. For each segment, the intensity for each emotion was normalized by a scaling function, defined as follows.

$$EE_i^p = Sigmod\left(EE_i^p - \frac{1}{I}\sum_{i=1}^{I} EE_i^p\right) \quad (1)$$

where $EE_i^p$ is the intensity of emotion $i$ for the *p-th* participant from the self-tagged emotion after scaling. The sigmoid function $1/1+e^{-s}$ was used. Second, the value *Diff* is the difference between the scaled self-tagged emotion intensity and the predicted emotion intensity defined as

$$Diff_i^p = PE_i^p - EE_i^p \quad (2)$$

The value of *Diff* is assumed the Gaussian distribution $N\left(\mu,\sigma^2\right)$ with mean $\mu = mean(Diff_i^p)$ and standard deviation $\sigma$ and is used to generate the Gaussian white noise as follows.

$$noise_i^p \sim N\left(\mu,\sigma^2\right) \quad (3)$$

where $noise_i^p$ is independent and identically distributed and drawn from a normal distribution with mean $\mu$ and standard deviation $\sigma$. The noisy expressed emotion intensity data are generated as follows to increase the number of training data.

$$\begin{cases} EE_i^p = EE_i^p + \omega \times noise_i^p & , if\ \mu-\sigma \le Diff_i^p \le \mu+\sigma \\ EE_i^p = EE_i^p & , otherwise \end{cases} \quad (4)$$

where $\omega$ is a weighting factor for the generated additive noise. Thereafter, the DAE-based emotion conversion model is constructed based on the self-annotated (expressed) as well as the noisy expressed emotion data (generated) and the annotated emotion data from other subjects (perceived). The trained DAE-based emotion conversion model is then used to convert the perceived emotion profile into the corresponding expressed emotion profile.

### 3.2 Long-Term Emotion Tracking for Mood Detection

This work considers mood swing as a long-term accumulation of emotions. As shown in Fig. 3, the daily conversational speech is used to extract a sequence of 384-dimensional feature vectors.

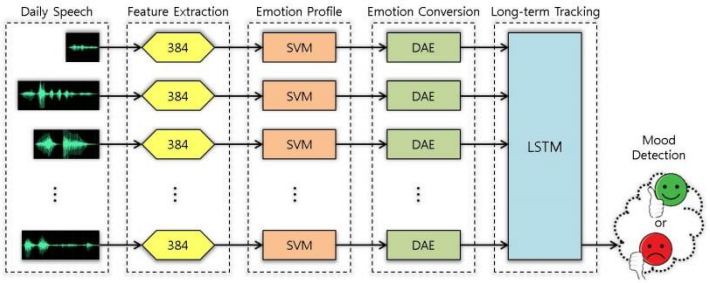


Figure 3. Flowchart of long-term emotion tracking for mood detection

Then, each feature vector is fed to the SVM-based emotion prediction model to obtain the emotion profile. The predicted emotion profile is then converted into the

expressed emotion profile using the personality-based conversion function. Finally, the expressed emotion profile sequence is used for mood detection using an LSTM-based long-term emotion tracking model.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

The **MHMC** emotion database was used to evaluate the emotion prediction performance based on 5-fold cross validation by using SVM. The accuracy of overall emotion recognition was 81.30%. As shown in Table 2, 1212 emotion utterances were collected to construct the **EP-DB**. The self-tagged samples after passing the consistency test in **EP-DB** were kept for experiments. The utterances were used to evaluate the emotion conversion performance for each emotion in **EP-DB**.

Table 2. Statistical information of self-tag for each personality in EP-DB

| Personality | Angry | Happy | Sad | Neutral |
|---|---|---|---|---|
| Extraversion | 128 | 128 | 132 | 133 |
| Introversion | 176 | 170 | 174 | 171 |

The weighting factor $\omega$ for the generated additive noise was fixed at 0.1. Table 3 shows the performance evaluation in emotion conversion model with/without personality. It can be observed that the R-square in the emotion conversion with personality is higher than the emotion conversion without personality for different hidden node numbers in DAE. Among different hidden node numbers in DAE, as the hidden node number was set to 64, the highest R-square value was achieved and this node number was used in the following experiments.

Table 3. Comparison of the emotion conversion models without/with personality

| | Number of hidden nodes in DAE | | | | |
|---|---|---|---|---|---|
| | 8 | 16 | 32 | 64 | 128 |
| Without personality | 0.91 | 0.91 | 0.91 | **0.91** | 0.91 |
| With personality | 0.91 | 0.92 | 0.92 | **0.92** | 0.91 |

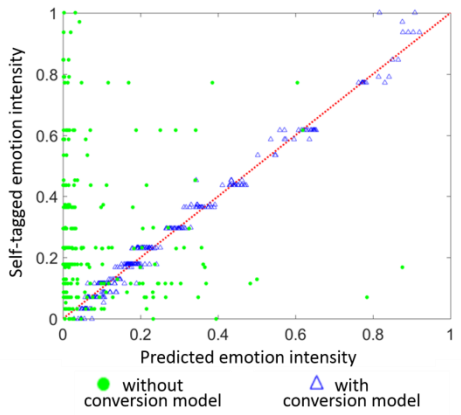


Figure 4. Correlation between the self-tagged and predicted emotion without/with emotion conversion model for angry emotion

Fig. 4 shows the correlation between the self-tagged and predicted emotions with/without emotion conversion model for angry emotion. This result shows the predicted emotion using emotion conversion model is closer to the self-tagged data than that without using emotion conversion model. Therefore, the proposed model can reduce the difference between the perceived emotion and the expressed emotion.

For mood detection, leave-one-speaker-out cross validation was employed for evaluation. As the LSTM hidden node number was set to 8, the mood detection performance achieved the best as shown in Fig. 5.

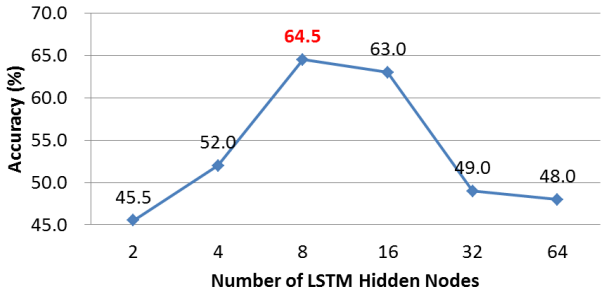


Figure 5. Performance of the proposed method

This work compared the proposed method with the traditional models. As shown in Table 4, the accuracy of mood detection using LSTM is better than that using HMM under different conditions.

Table 4. Performance comparison with traditional model

| | HMM | | LSTM | |
|---|---|---|---|---|
| | Without personality | With personality | Without personality | With personality (Proposed method) |
| Accuracy | 55% | 59.5% | **63%** | **64.5%** |

## 5. CONCLUSION

This work proposed an approach to mood detection based on long-term expressed emotion tracking. First, a support vector machine (SVM)-based emotion model is developed to generate a perceived emotion profile. Then, the DAE was used to construct an emotion conversion model to characterize the relationship between the perceived emotion and the expressed emotion of the subject for a specific personality. Finally, a long short-term memory (LSTM)-based mood model is constructed to model the temporal fluctuation of emotions for mood detection. Experimental results show that the proposed DAE+LSTM method achieved a detection accuracy of 64.5%, a 5.0% improved accuracy compared to the HMM-based method.

## REFERENCES


[1] M. Reddy, "Depression: the disorder and the burden," *Indian Journal of Psychological Medicine*, vol. 32, no. 1, pp. 1, 2010.

[2] C. H. Wu, J. C. Lin and W. L. Wei, "Two-Level Hierarchical Alignment for Semi-Coupled HMM-Based Audiovisual Emotion Recognition with Temporal Course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880-1895, December 2013.

[3] J. C. Lin, C. H. Wu and W. L. Wei, "Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp.142-156, 2012.

[4] C. H. Wu, and W. B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10-21, 2011.

[5] S. P. Robbins, Organizational behavior, 14/E: Pearson Education India, 2001.

[6] C. Busso, and S. S. Narayanan, "The expression and perception of emotions: comparing assessments of self versus others," in Proc. Interspeech, pp. 257-260, 2008.

[7] K. P. Truong, M. A. Neerincx, and D. A. Van Leeuwen, "Assessing agreement of observer-and self-annotations in spontaneous multimodal emotion data," in Proc. Interspeech, pp. 318-321, 2008.

[8] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[9] S. Kshirsagar, "A multilayer personality model," in Proc. International symposium on Smart graphics, pp. 107-115, 2002.

[10] Personality-Central, "Extroversion-Introversion preferences."

[11] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in Proc. Artificial Neural Networks in Engineering, vol. 710, 1999.

[12] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in Proc. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, pp. 39-44, 2000.

[13] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," in Proc. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, pp. 29-33, 2000.

[14] F. Yu, E. Chang, Y. Q. Xu, and H. Y. Shum, "Emotion detection from speech to enrich multimedia content," in Proc. Pacific-Rim Conference on Multimedia, pp. 550-557, 2001.

[15] F. Schiel, S. Steininger, and U. Türk, "The SmartKom Multimodal Corpus at BAS," in Proc. LREC, 2002.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in Proc. Interspeech, vol. 5, pp. 1517-1520, 2005.

[17] R. Wiseman, 59 Seconds: Motivation: Think A Little, Change A Lot: Pan Macmillan, 2011.

[18] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, Jan. 2012.

[19] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.