# Deep Learning-based Emotion Spatial Regression in Speech Recognition for Human-computer Interaction

Jui-Feng Yeh[1], Jian-Cheng Tsai[2], Bo-Wei Wu[3], Tai-You Kuang[4]

Department of Computer Science & Information Engineering,
National Chia-Yi University,
Chiayi city, Taiwan (R.O.C.)
{ralph, s1033027, s1032881, s1050470}@mail.ncyu.edu.tw

## Abstract

This paper propose a method which uses the CNN-LSTM speech recognition as a basis to convert the input of user's speech signal into text, and combine the CKIP system to break sentence or phrases into words and calculate Valence-Arousal values by the method of vector-based in the last. We calculate Valence-Arousal values with vector-based method and three lexicons which is emotional, degree and negative lexicon respectively for words and phrases. If the input is not exist in three lexicons, we will use the OOV(Out-of-Vocabulary) function to solve the problem. The OOV function not only solve the problem but also is helpful for Valence-Arousal values calculation. Because the sentence is consists of many modifiers and words, the system will have calculation deviation. In order to improve the accuracy of values, we use the way of detecting modifiers again and implement in OOV function to reach the goal. Finally, the system will output the Valence-Arousal values and present on the chart.

**Index Terms**: speech recognition, affective computing, textual sentiment analysis, emotional coordinates

## 1. Introduction

Sentiment analysis is a new trend in the development of computer science. It has invested a lot of research in the past two decades. As we know, the purpose of emotional computing is to make machines acquire and provide human emotions. Identifying, expressing, and generating emotions from the environment are also goals of emotional computing. Therefore, the data in real life is very important for the establishment of the system. The meaning of emotional computing is the technology that produces or intentionally affects emotions. These technical calculations are all related to emotions. Only with reasonable calculations, the machine can act like human beings. Since the Turing test, computer scientists have made many efforts in artificial intelligence, especially in detecting, perceiving, and expressing emotions. Therefore, there are many multimedia-type machines in the society. For example, computer vision technology can be used to detect emotional expressions, or combined with voice and dialogue, a multi-functional action agent with emotional interactions can be developed and applied to the real environment, so that emotional communication be emphasized in terms of human-computer interaction. "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions." said by Marvin Minsky. He tries to make computers have the ability to deal with emotional aspects in human-computer interaction. Classifying the type and degree of emotions displayed by users is the goal of emotion recognition.

First, we can classify emotion classification into two categories: discrete categories and dimensional models. The former means that the emotional categories are discrete and have different structures; the latter means that emotions should be described as continuous values regression based on dimensions. Li et al. [1] pointed out that the correct division of sentences has a great importance for influencing the meaning of emotions because it changes the usage of words. The method proposed in this paper is based on the enhanced mutual information (EMI) scores of new user words. The higher the EMI score of a word, the higher the likelihood that it is a new word. The same new emotion word detection is still one of the basic issues of emotional computing. Huang et al. [2] found that the use of emotional words and polarity to predict new emotional words is one of the main contributions of this paper. They used the likelihood ratio test (LRT) method to use the quantitative association degree to find new words and frames. The method of production is completely unsupervised. In addition, the prediction of the polarity of emotional words is conducive to the classification and sentiment analysis of emotional words. Tang et al. [3] and Yao and Li [4] each propose a related method of word embedding for sentiment classification that is also a mainstream method. Wang et al. [5] used the deep learning method CNN-LSTM to deal with sentiment analysis problems. Yao and Li [4] proposed a eigenvector approach to search for special words in the context of the word2vec tool. In addition, this article proposes a new idea that uses eigenvectors to represent the emotional polarity of words. However, this idea is still growing. The more studies need to be proven, because the data in a particular situation is limited. Yeh et al. [6] used a linear regression method to perform linear regression on Chinese characters' Valence-Arousal values. Saif et al. [7] use three different sentiment lexicons to derive word prior sentiments. Zheng et al. [8] presented a way for identifying aspect and sentiment words by the AEP-LDA model. Appraisal Expression Patterns (AEPs) represent the opinion of people express and services, which can regarded as a the syntactic relationship between aspect and sentiment words. Maynard et al. [9] consider the sarcasm is a important feeling in emotion, so they developed a hashtag tokeniser for GATE to detect the sarcasm. Lan et al. [10] combine the convolutional neural network and sentiment word vectors to sentence-level sentiment analysis. Hao et al. [11] use five different sentiments to build the complex network and research the sentiment diffusion mechanism. Huang et al. [12] propose a new way with applying Chinese opinion words for words extraction associated with attribute words. dos Santos et al.[13] use the small text and convolutional neural network from character- to sentence-level information to implement the sentiment analysis. Gilbert, CJ Hutto Eric. present VADER which is a simple rule-based model and build goldstandard list of lexical features for sentiment analysis

Speech emotion recognition plays an important role. Since people receive emotions expressed by the other party, they often determine the emotions contained in the sentences by experience or judgment on the person, or they may not be aware of them. Therefore, in order to find the key words or let the emotions can be noticed, this paper propose a method to let the emotion recognition system to distinguish words and phrases, and then perform different processes of emotional calculations based on words or phrases, and integrate speech recognition and emotion recognition technologies into an applicable system. In this paper, 375 Chinese balanced sentences were used to test the accuracy of the speech recognition system used for speech recognition. By combining the CKIP



ISBN:978-1-7281-0110-1

Chinese word segmentation system, the system can cut the speech recognition text and identify various parts of speech, so as to retrieve the information we need from it.

This paper also proposes a new method for calculating the Valence-Arousal value with vector-valued two-dimensional vector analysis. At the same time, the values of the degree thesaurus and the negative thesaurus in the system are also calibrated according to observation and training. All of these will have more detailed explanations in the research methods.

## 2. The model of CNN-LSTM Speech Emotion Spatial Regression

In this chapter, we will explain the method which can be divided into two part. One is the CNN-LSTM speech recognition and other one is emotion spatial regression.

### *2.1. CNN-LSTM speech recognition*

This paper mainly adopts the semantic part of speech sentiment analysis. In speech recognition, we will convert the speech signal to the text. According to the references, we consider that the CNN layers and LSTM layers are parallel processing. Taking $\tau$ as a time step, $x_{t-\tau+1}$ to $x_t$ is an MFCC vector as the start point of input and $x_t^{LSTM}$ is the sequence which is consist by MFCC feature. The previous feature sequence affects the subsequent output feature sequence. The $x_t^{LSTM}$ sequence finally outputs the $z_t^{LSTM}$ sequence through two layers of LSTM layers.

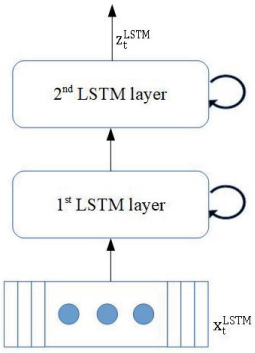


Figure 1: *LSTM layer processing.*

The spectrogram of convolution layers is a two-dimensional convolution between the predefined linear filter and the spectrum. By converting the short-time Fourier transform to a frequency spectrum, K different filters are applied to the spectrogram. K different filtered images are generated in the convolution layers, and SIF fragments composed of the characteristic F×τ matrix are extracted by the characteristics of the SIF. The time step of the SIF segment is the same length as the LSTM layer, followed by down sampling and denoising to form $x_t^{CNN}$:

$$F(f,t) = \left| \sum_{n=0}^{N-1} s(n)w(n)e^{\frac{-j2\pi nf}{N}} \right|, f = 0, \dots, N/2. \quad (1)$$

$F(f,t)$ in formula (1) represents the relationship between frequency and time. j is equal to $\sqrt{-1}$, N is the length of the cut-off frame, e is the natural logarithm constant, w(n) is the non-zero window function, and s(n) represents the audio frame. We enter into the pooling layer to cut the image into non-overlapping sub-regions, and then use the largest pooling method to gradually reduce the size of the presentation space. We can thereby extracting the main features of the sub-region. From the $x_t^{CNN}$ through two layers of CNN layers, the output is $z_t^{CNN}$ after the planarization output which adjust the output size in order to completely connect the structure of the upper layer. Finally, the full-connected layer stores the LSTM sequence information, the CNN frequency and time information. The system will output the result through two hidden layers consisting of 512 RELUs.

$$F_{down}(f,t) = \sum_{n=0}^{W-1} F(f+i,t)/W, f = 0, \dots, (K-1). \quad (2)$$

$$F_{dn}(f,t) = F_{down}(f,t) - \min_t\{F_{down}(f,t)\}, \quad f = 0, \dots, (K-1). \quad (3)$$

### *2.2. Emotion spatial regression and thesaurus building*

In the part of building the thesaurus, we will use the following four thesaurus as the knowledge base. We use the formula which is mentioned in chapter followed to calculate the VA value of the phrase. We will use the CVAP thesaurus to train the degree of influence of degree adverbs and negative adverbs on emotional words.

E-HowNet is extended via HowNet. We use E-HowNet to classify emotional words, which is beneficial to build our emotional thesaurus. ANTUSD contains 27,370 words. They collect frequently-used words for a long time to modify the lexicon. The purpose of SentiWordNet 3.0 was originally designed to be used for emotional analysis, and the words which was marked were classified as positive, negative, and neutral. The CVAW is a resource containing 2803 labeled Valence-Arousal values ??with a tagged Valence-Arousal value between 0 and 9. ANTUSD, SentiWordNet 3.0, and CVAP will be used by us to build the emotional thesaurus as a basis for our emotional judgment. There are 2251 data of phrases in CVAP which the way of tagged is as the same as CVAW thesaurus and we will use this as the training data for adverbs.

### *2.3. Calculation of phrase by Vector Representations*

We use the relationship between words in E-HowNet to group all the words into groups of synsets, and each synset has a corresponding hypernym. Finally, we completed our thesaurus by collecting words from CVAW to form hypernym. On the other hand, we can expand our thesaurus by gaining more and more hypernyms. When we will calculate the VA value of a word, we will find the VA value of the word based on the above thesaurus. For words that do not exist in the thesaurus, we use the OOV function to predict their Valence-Arousal values.

In the part of the calculation of phrase VA value, we divide the phrase into three parts: degree adverb, negative adverb, and emotional word. We obtain the VA value of emotional words through the above functions, and obtain their degree of influence on emotions from the degree adverbs that have been trained and negative adverbs. Then, we calculate the VA value of the phrase by the formula (4) and (5). According to the formula, we consider that the modification of the adverbs to the emotional words should be superimposed, so the degree of influence should use the way of multiplication to add. For the reason why we first subtract the central value and multiply it by the degree of influence, we think that this action will get the eigenvector of the emotional word. By multiplying the eigenvectors, we will be able to get Valence-Arousal values that are closer to the true emotions.

$$E_v' = D_V * (E_v - C) + C, E_v \in Valence \quad (4)$$

$$E_a' = D_A * (E_a - C) + C, E_a \in Arousal \quad (5)$$



ISBN:978-1-7281-0110-1

### *2.4. The value prediction for Out of Vocabulary*

OOV is an abbreviation for Out of Vocabulary. The timing of using the OOV function is when the word is not established in the thesaurus. We can solve the problem of the word which does not exist in the emotional dictionary through the use of OOV function.

$$X = \frac{\sum_{Seq_n=1}^{n} \frac{\sum_{i=1}^{N} x_i}{N}}{Seq_n}, X \in Valence \quad (6)$$

$$Y = \frac{\sum_{Seq_n=1}^{n} \frac{\sum_{i=1}^{N} y_i}{N}}{Seq_n}, Y \in Arousal \quad (7)$$

We calculate the average VA value of each character in the word and add all the values to average. The average value is the VA value of this undefined word. In addition to the calculation of values not defined in the dictionary, we will design a function in OOV function to be helpful for the calculation of the VA value of the phrase. By detecting the degree and the function of negative adverbs again, we can reduce the calculation of errors and improve the accuracy of the results.

### *2.5. The training for modifiers*

We will use CVAP as the data for our training level and negative adverbs because there are a lot of categories of data in CVAP. We use the type of phrase to train degree words and negative words which is a degree word or a negative word with an emotional word. The training method is using the VA value in the phrases and the collocation of the formulas. We inversely calculate influence of the degree and the negative words in the Valence and Arousal values and then calculate average influence to completing the training of degree words and negative words.

## 3. Experiment

Since the system uses speech input, and used the recognized text for emotion detection. This makes the speech recognition and the accuracy of the emotion detection very important. The following experiments will be conducted on the accuracy of speech recognition and emotion detection.

### *3.1 Speech recognition system accuracy*

The experiments of speech recognition accuracy were conducted using standard Chinese speech recognition experiments. The test data used is 375 Chinese balance sentences that are input into our speech recognition system in sequence. We divide the results into many categories and finally calculate the accuracy of the speech recognition system.

N is the number of words in each Chinese sentence. For example, N = 8 for the sentence "請把這籃兔子送走". After the N value of each Chinese balanced sentence is calculated, the preparation of experimental data is completed. The output is divided into three categories: Deletion, Substitution and Insertion, which will be introduced in the following description. Deletion is the number of words that were not recognized but there are in the correct answer. Substitution is the number of words that were recognized as wrong answer. Insertion is the number of words that were recognized but there are not in the correct answer. After calculating the Deletion, Substitution, and Insertion values of each output, that is, identifying the number of incorrect syllables, the accuracy is calculated according to the formula below to identify the accuracy of the system.

$$\text{Percent Accuracy} = \frac{\mathrm{N} - \mathrm{D} - \mathrm{S} - \mathrm{I}}{\mathrm{N}} * 100\% \quad (8)$$

According to the above experimental data and experimental methods, we obtained the accuracy values of the speech recognition system as shown in Table 1 and Table 2. Base on the experimental results, the recognition accuracy is nearly 80% correct, which is quite good in terms of speech recognition and is an available system.

*Table 1: Result statistical table*

| | N | D | S | I |
|---|---|---|---|---|
| Different homonyms | 2929 | 5 | 478 | 6 |
| Correct identification | 2929 | 5 | 649 | 6 |

*Table 2: Speech recognition system accuracy*

| | Percent Accuracy(%) |
|---|---|
| Different homonyms | 0.833049 |
| Correct identification | 0.774667 |

### *3.2 Emotion detection system experiment*

We measured the accuracy of our proposed Emotion detection method according to the way of accuracy calculation of IJCNLP 2017 competition. This experiment uses the test data released by IJCNLP 2017 to conduct experiments. The experimental data has a total of 750 emotional words and 750 phrases. We enter the test data into the emotion detection system. The system needs to output the Valence and Arousal value of this data. Both values are between 0 and 9. The Valence value represents the degree of positive and negative emotions. The larger the value, the more positive the emotion represented by the data, and the smaller the negative emotion. The arousal value represents the degree of emotional excitement. The larger the value, the more excited the emotion represented by the data, and the smaller the more calm. The Mean Absolute Error (MAE) value and the Pearson Correlation Coefficient (PCC) value are calculated according to the formula for correct test results of the test data. MAE is the average difference between the output of our emotion detection system and the data value of the correct marker. PCC is the correct rate of our output.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - P_i| \quad (9)$$

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{A_i - \bar{A}}{\sigma_A}\right)\left(\frac{P_i - \bar{P}}{\sigma_p}\right) \quad (10)$$

We used the above experimental method to obtain the MAE and PCC values of the output of our emotion detection system, as shown in Table 3 and Table 4. According to our experimental results, the mood detection method proposed by us is not very good, but this result is similar to the baseline and can be regarded as a feasible method. The accuracy in calculating the Valence value is 0.685, which is higher than the baseline. However, the accuracy of calculating the Arousal value is only 0.549, which is lower than the baseline. The future of the emotion detection system should be more accurate toward calculating the Valence value, and modifying the way to calculate the Arousal value makes it more accurate than it is now.

*Table 3: The accuracy of Valence*

| | Valence MAE | Valence PCC |
|---|---|---|
| NCYU-Run1 | 0.9785 | 0.685 |
| Baseline | 1.0175 | 0.6265 |
| NCYU-Run2 | 1.2050 | 0.6665 |

*Table 4: The accuracy of Arousal*

| | Arousal MAE | Arousal PCC |
|---|---|---|
| Baseline | 0.819 | 0.593 |
| NCYU-Run1 | 0.945 | 0.549 |
| NCYU-Run2 | 0.989 | 0.534 |

## 4. Conclusion

This paper presents a speech input instant emotional analysis system that includes CNN-LSTM speech emotion recognition and CKIP Chinese word segmentation as well as a vector-based textual sentiment analysis method that can help user quickly understand the mood tendencies contained in the text. For the experimental results of the accuracy of speech recognition and the accuracy of the price arousal value in this system, the good accuracy of speech recognition can obtain accurate price arousal values. The vector-based text sentiment analysis proposed in this paper can be regarded as a feasible and efficient method.

## 5. Acknowledgments

This project is supported by the special research project of the Ministry of Science and Technology (general research project) Support for "image annotation and sentence generation based on multimodal deep learning" (Project Number: 106-2221-E-415 -021).

## 6. References

[1] Li, Wei, et al. "Improved New Word Detection Method Used in Tourism Field." *Procedia Computer Science* 108 (2017): 1251-1260.

[2] Huang, Minlie, et al. "New word detection for sentiment analysis." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014.

[3] Tang, Duyu, et al. "Learning sentiment-specific word embedding for twitter sentiment classification." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014.

[4] Yao, Yushi, and Guangjian Li. "Context-aware Sentiment Word Identification: sentiword2vec." *arXiv preprint arXiv:1612.03769* (2016).

[5] Wang, Jin, et al. "Dimensional sentiment analysis using a regional CNN-LSTM model." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2016.

[6] Yeh, Jui-Feng, et al. "Dimensional sentiment analysis in valence-arousal for Chinese words by linear regression." *Asian Language Processing (IALP), 2016 International Conference on*. IEEE, 2016.

[7] Saif, Hassan, et al. "Contextual semantics for sentiment analysis of Twitter." *Information Processing & Management* 52.1 (2016): 5-19.

[8] Zheng, Xiaolin, et al. "Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification." *Knowledge-Based Systems* 61 (2014): 29-47.

[9] Maynard, Diana, and Mark A. Greenwood. "Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis." *Lrec*. 2014.

[10] Lan, Man, et al. "Three convolutional neural network-based models for learning sentiment word vectors towards sentiment analysis." *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016.

[11] Hao, Xiaoqing, et al. "Sentiment Diffusion of Public Opinions about Hot Events: Based on Complex Network." *PloS one* 10.10 (2015): e0140027.

[12] Huang, Jia-Yen. "Web mining for the mayoral election prediction in Taiwan." *Aslib Journal of Information Management* 69.6 (2017): 688-701.

[13] dos Santos, Cicero, and Maira Gatti. "Deep convolutional neural networks for sentiment analysis of short texts." *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014.

[14] Gilbert, CJ Hutto Eric. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*. 2014.



ISBN: 978-1-7281-0110-1