# Emotion Estimation by Joint Facial Expression and Speech Tonality Using Evolutionary Deep Learning Structures

Chih-Che Chung, Wei-Ting Lin, Rong Zhang, Kai-Wen Liang, and Pao-Chi Chang
*Department of Communication Engineering*
*National Central University*
Taoyuan City, Taiwan
pcchang@ce.ncu.edu.tw

***Abstract*—This work proposes an emotion recognition system by adopting facial expression and speech tonality on deep learning networks. Both convolutional neural networks and long short term memory networks are used for feature training. The two features can be trained together to acquire higher accuracy. Moreover, Structure Evolution which is inspired by the Genetic Algorithm is added to optimize the parameters in the model. The experimental results show the joint model optimized by Structure Evolution surpasses the single model by at least 10% and outperforms the state-of-the-art work over 1%.**

***Keywords—emotion estimation, deep learning, speech, facial expression, Structure Evolution***

## INTRODUCTION

Audiovisual emotion recognition has been a popular research topic for years. There has been various of works in visual pattern recognition for facial emotional expression recognition, as well as in signal processing for audio-based detection of emotions, and lots of multimodal approaches combining these cues. Thanks to the development of artificial intelligence, researches based on deep learning achieve better performance on this issue that made it possible to create real affective systems in reality. We now view applications across many domains, including robotics, HCI, healthcare, and multimedia.

When it comes to acquiring neural network with the highest accuracy, we often set different parameters (e.g. kernel size, numbers of kernels, numbers of layers) in our model during training process through try and error which is usually time-consuming and a game of luck or, more often, we manage to have our data better preprocessed; nonetheless, the outcome is not always under our expectation.

Being eager for a systematic and automatic network-updated method, we are inspired by Darwin's Evolution Theory and Genetic Algorithm. We design a system that every species goes through fitness, selection, crossover and selection stages. After survival of the fitness over generation competition, we can get the model having the highest accuracy, we name the method as Structure Evolution.

In our work, we estimate emotion based on facial expression with Convolutional Neural Network (CNN), speech tonality with both Long Short Term Memory Network (LSTM) and CNN and the integration of both through model set out by try and error and model sieved out by Structure Evolution. The results show that we are able to obtain a model with higher accuracy than models that are trained by a single dataset. Furthermore, Structure Evolution is found to be helpful in optimizing parameters in neural network and sieving out better performed model.

## I. STRUCTURE EVOLUTION

We apply the concept of Darwin's Evolution Theory to help us finding the model with the highest accuracy. In Structure Evolution, each neural network is considered as an individual, i.e., species, and parameters among neural network are regarded as gene. An individual is supposed to undergo Fitness, Selection, Crossover and Mutation stages, as shown in Fig. 1. We are then able to keep evolving better individuals, i.e. neural network with the highest accuracy, over generations.

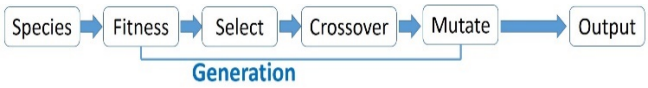


Fig. 1. Flow chart of structure

In our case, genes in CNN include numbers of layers in convolutional layer, numbers of kernels, kernel size, numbers of layers in dense layer, numbers of neurons in dense layer, activation function and optimizer; while, genes in LSTM include numbers of layers, numbers of neurons, numbers of layers in dense layer, numbers of neurons in dense layer, activation function and optimizer. In the beginning, we need to set out parameters table, e.g. as set in Table 1 and Table 2.

Table 1. Parameters table of CNN

| |
|---|
| # of layers in convolutional layer: [1, 2, 3, 4, 5, 6, 7, 8] |
| # of kernels: [4, 9, 16, 25, 36, 49, 64, 81] |
| Kernel size: (3,3) (5,5) (7,7) |
| # of layers in dense layer: [1, 2, 3, 4, 5, 6, 7, 8] |
| # of neurons in dense layer: [4, 9, 16, 25, 36, 49] |
| Optimizer: 'rmsprop', 'adam', 'sgd', 'adagrad', 'adadelta', 'adamax', 'nadam' |
| Activation function: 'relu', 'elu', 'sigmoid', 'tanh' |

Table 2. Parameters table of LSTM

| |
|---|
| # of layers in LSTM: [1, 2, 3, 4] |
| # of kernels: [25, 36, 49, 64, 81] |
| # of layers in dense layer: [1, 2, 3, 4, 5, 6, 7, 8] |
| # of neurons in dense layer: [25, 36, 49, 64] |
| Optimizer: 'rmsprop', 'adam', 'sgd', 'adagrad', 'adadelta', 'adamax', 'nadam' |
| Activation function: 'relu', 'elu', 'sigmoid', 'tanh' |





978-1-7281-2575-5/19/$31.00 ©2019 IEEE

In our initial setting, we randomly generate 20 individuals based on parameters table and there are 20 generations in total.

In Fitness stage, each individual is trained by dataset and accuracy is calculated. Each individual has the possibility of 40% to survive.

Due to the fact that individuals with better symptoms are more likely to yield fitter offspring. In Select stage, we select two of the highest accuracy models as "parent", as Fig. 2 and Fig. 3., and adopt multi-point crossover model with 60% crossover rate in Crossover stage to generate new model, as Fig. 4.

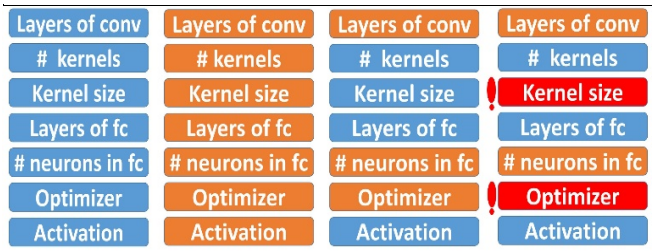


Fig. 2. Parent1 Fig. 3. Parent2 Fig. 4. New model Fig. 5. Child

To simulate the randomization in nature, mutation is required. In our system, algorithm would randomly decide whether parameters in new model would change in Mutation stage which, therefore, generate a child, i.e., a new individual, for next generation, as Fig. 5. The "!" in Fig. 5 marks the changed parameter.

## II. Emotion Estimation on Facial Expression

A CNN model is used to train the image features for the sake of emotion estimation using facial expressions. The adopted dataset is FER-2013 [1] which contains 28709 training images with 7 categories of happy, angry, sad, neutral, fear, surprise and disgust of emotions. Through the optimization of Structure Evolution, our model can achieve 61.48% accuracy that beats to the model without Structure Evolution for 60.13% accuracy, as shown in Fig. 6. and Table 3.

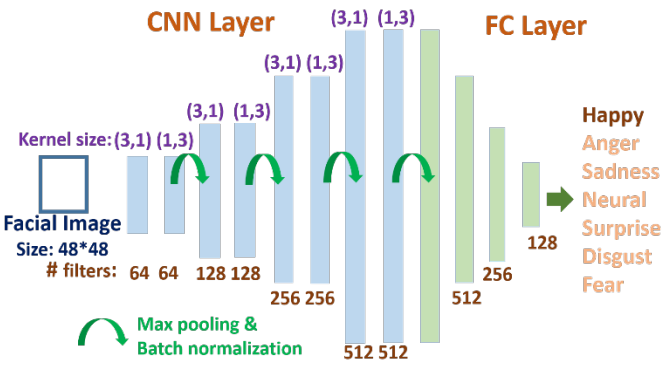


Fig. 6. Emotion recognition model on facial image (without Structure Evolution)

Table 3. Parameters of emotion recognition model on facial image (with Structure Evolution)

| |
|---|
| Setting: There are 20 individuals in each generation and 20 generations in total.<br><br>CNN:<br># of layers in convolutional layer: 8<br># of kernels: 49<br>Kernel size: (5,5)<br># of layers in dense layer: 2<br># of neurons in dense layer: [49, 49]<br>Optimizer: Adam<br>Activation function: Elu |

## III. Emotion Estimation on Speech Tonality

### A. Neural network and preprocessing

We utilize the advantages of LSTM structure to process speech features. The feature chosen is the Mel-scale Frequency Cepstral Coefficient (MFCC) of the speech signal. We set (a) 40 logMel filter-banks, (b) 78 frames per speech file, and thus we can get the MFCC metric with the size of 78*40 for each speech file, as shown in Fig. 7.

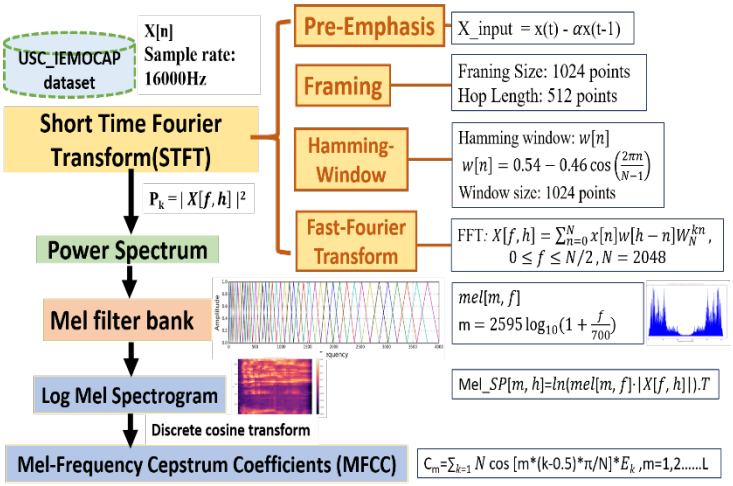


Fig. 7. Mel-scale Frequency Cepstral Coefficient

### B. Speech dataset and results

We adopt the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [2] database provided by University of Southern California in our model which is one of the largest open-sourced emotion detection dataset. It consists of approximate 12 hours of audio-visual data, including facial recordings, speech and text transcriptions.

We consider total 4861 speech data of happy, anger, sadness and neutral, four categories of emotions in total. The accuracy of our own model which combines CNN and LSTM is 56.4%, as shown in Fig. 8, and the accuracy of model optimized by Structure Evolution is 60.18%, as shown in Table 4.

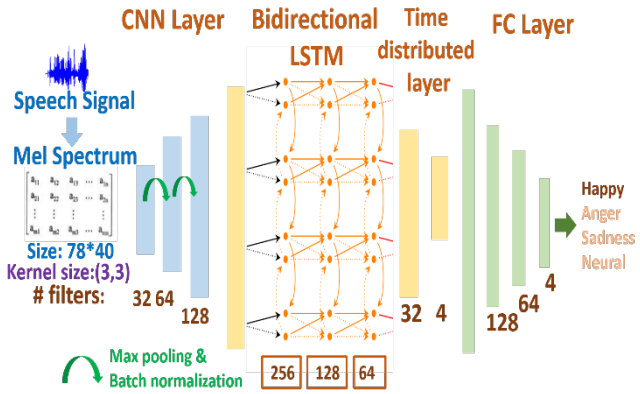


Fig. 8. Emotion recognition model on speech data (without Structure Evolution)

Table 4. Parameters of emotion recognition model on speech data (with Structure Evolution)

Setting: There are 20 individuals in each generation and 20 generations in total.

LSTM:
# of layers in LSTM: 3
# of kernels: 64
# of layers in dense layer: 8
# of neurons in dense layer: 49
Optimizer: 'adam'
Activation function: 'relu'

## IV. Emotion Estimation by Joint Facial Expression and Speech Tonality

In addition to building CNN and LSTM models for facial expression and speech separately, we combine the two outputs from each model by concatenating them and feed them into a final fully connected layer. As predicted, the combination of the two features can achieve higher accuracy on both cases, with and without Structure Evolution, for 72.16% and 65.44% accuracy respectively. Fig. 9. is the joint model generated by Structure Evolution.

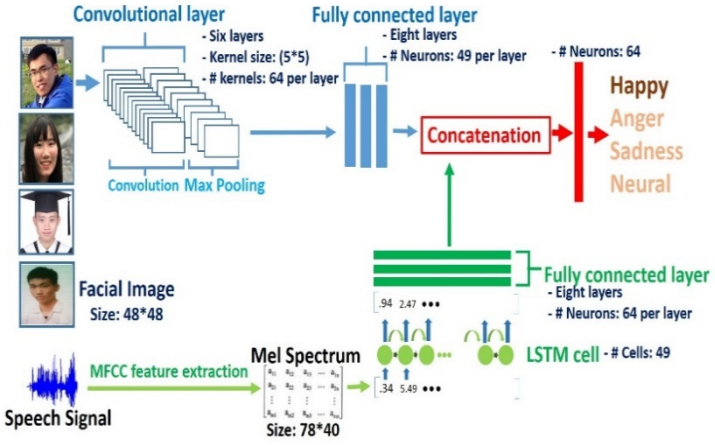


Fig. 9. Neural network of facial image and speech estimate model (with Structure Evolution)

## V. Discussion

### A. Accuracy of emotion estimation

As shown in Table 5., no matter adopting the model found out by ourselves or Structure Evolution, models all have better performances on integrated data.

Furthermore, adopting the neural networks generated by Structure Evolution are shown to improve the performance of each model on the single set of image, speech and integrated data.

Table 5. Accuracy comparison

| Accuracy / Trained Dataset | Without Structure Evolution | With Structure Evolution |
|---|---|---|
| Facial image | 60.13% | 61.48% |
| Speech | 56.40% | 60.18% |
| Image and speech | 65.44% | 72.16% |

### B. Time complexity of Structure Evolution

When making use of Structure Evolution, only after training all the individuals in a generation can we get the model with the highest accuracy. As we set larger number of individuals in each generation, it would increase the biodiversity in each generation and is more likely for us to get model with higher accuracy; nonetheless, it would also take more time to train individuals in each generation cycle. Thus, the time complexity of Structure Evolution is in direct proportion with the number of individuals and generations in initial setting.

### C. Size of neural network

The screenshot taken from one of the generations was shown in Fig 10. We saved each neural network in a generation into files with its accuracy shown in the file name. E.g. GACNN_60.18433_h5: the accuracy of the neural network is 60.18433%. It is obvious to find the size of neural network may not in direct proportion with its accuracy.

Many people tend to consider the more complicated the neural network is, the higher accuracy it may results. We are able to dispel the myth in the case, and, most importantly, it indicates the Structure Evolution is effective to get the model with simpler model with higher accuracy which we may not sense when we device models on our own.

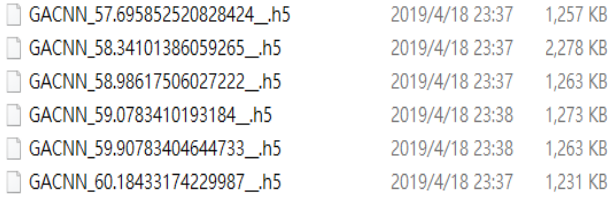


Fig. 10. The example screenshot taken from one of the generations

### D. Comparison of our model with related works

Table 6. Comparison of our work with others'

| Method | Data | Accuracy |
|---|---|---|
| ICON [7] | IEMOCAP database of audio and textual features | 63.5% |
| Dialogue RNN [8] | IEMOCAP database of audio and textual features | 64.5% |
| Multi-modal [9] | IEMOCAP database of speech, text, and motion-capture data | 71.04%. |
| Structure Evolution | IEMOCAP database of audio, FER-2013 facial image | 72.16% |

We also compare our model with others' work in Table 6., and our models have excellent which outperforms the state-of-the-art work over 1%.

### E. The advantages of using Structure Evolution

- Structure Evolution is a systematic and automatic network-updated method based on Darwin's Evolution Theory. It is able to optimize parameters in neural network without manual setting which is always time-consuming.

- Structure Evolution is especially applicable to data which features are irregular, unclear, or hard to tell. When we are not able to analyze or capture features in our data, not to mention understand the roles of layers in our network; at this time, devising network is just a game of luck. Nevertheless, Structure Evolution seems to provide us a way to find the fittest neural structure in this situation.

We can conclude that Structure Evolution are helpful in finding optimized structures which lead to higher accuracy. Furthermore, we can have a model with higher accuracy by training the model with integrated data.

## VI. Conclusion

Emotion estimation plays a crucial role in humans' lives. In this work, we have obtained a joint model trained by facial image and speech which exceeds the single model by 10% and surpasses the state-of-the-art work over 1%. Furthermore, Structure Evolution is a systematic and automatic network-updated method which is helpful in finding optimized structure leading to higher accuracy. For future work, we look forward to improve the accuracy by expanding the datasets and accomplish real-time speech and face emotion recognition.

## References


[1] I. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, et al, "Challenges in Representation Learning: A report on three machine learning contests," *arXiv* 2013.

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

[3] K.F. Man, K.S. Tang, and S. Kwong, "Genetic algorithms: concepts and applications [in engineering design]," *IEEE Transactions on Industrial Electronics ( Volume: 43 , Issue: 5)*, 1996.

[4] Eiben, A. E., "Genetic algorithms with multi-parent recombination". *PPSN III: Proceedings of the International Conference on Evolutionary Computation. The Third Conference on Parallel Problem Solving from Nature:* 78–87, 1994.

[5] O.Arriaga, P.G. Ploger, and M. Valdenegro, "Real-time Convolutional Neural Networks for Emotion and Gender Classification," *arXiv:1710.07557v1 [cs.CV] 20* Oct. 2017.

[6] H. Zhiyan and W. Jian, "Feature Fusion Algorithm for Multimodal Emotion Recognition from Speech and Facial Expression Signal," *MATEC Web of Conferences,* Jan. 2016.

[7] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R.Zimmermann, "ICON: Interactive Conversational Memory Network for Multimodal Emotion Estimation," *Conference on Empirical Methods in Natural Language Processing,* 2018.

[8] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A.Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Estimation in Conversations," *CoRR(abs/1811.00405*), 2018.

[9] S. Tripathi and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," *CoRR(abs/1804.05788),* Apr. 2018.