Dance to Music Expressively: A Brain-inspired System Based on Audio-semantic Model for Cognitive Development of Robots*

Dengju Li[1], Rui Yan[1*], Xiaoliang Xu[2], and Huajin Tang[1]

[1] Neuromorphic Computing Research Center,
College of Computer Science, Sichuan University, Chengdu, China.
{kevinleeex,huajin.tang}@gmail.com, *ryan@scu.edu.cn
[2] School of Computer Science and Technology, Hangzhou Dianzi University,
Hangzhou 310018, China. xxl@hdu.edu.cn

Abstract. Cognitive development is one of the most challenging and promising research fields in robotics, in which emotion and memory play an important role. In this paper, an audio-semantic (AS) model combining deep convolutional neural network and recurrent attractor network is proposed to associate music to its semantic mapping. Using the proposed model, we design the system inspired by the functional structure of the limbic system in our brain for the cognitive development of robots. The system allows the robot to make different dance decisions based on the corresponding semantic features obtained from music. The proposed model borrows some mechanisms from the human brain, using the distributed attractor network to activate multiple semantic tags of music, and the results meet the expectations. In the experiment, we show the effectiveness of the model and apply the system on the NAO robot.

Keywords: Cognitive robot· Brain-inspired system· Emotional model· Semantic representation.

1 Introduction

With the development of robotics, a growing number of social robots have entered people's lives. Many robots play the role of human beings, such as caring for the elderly, teaching, assisting the treatment of autistic children [3, 4]. Although the intelligence level of the robot is gradually improving, in the field of cognitive development, how to get robots to have compatible cognitive abilities as humans, to interact naturally with humans, or to respond quickly in changing environments, still face significant challenges and difficulties [13, 2]. In recent years, the research on the cognitive development of robots has attracted wide attention from scholars [1, 12, 9, 5], and these studies have demonstrated that

* This work was supported by the National Natural Science Foundation of China under grant number 61773271 and the National Key R&D Program of China under grant number 2017YFB1300201.

emotion, memory, and biological plausibility play essential roles in the cognitive development of robots.

Music processing is a whole-brain phenomenon [15], while the Limbic system plays a vital role for associating the auditory perception with meaning and memory, and guide behavioral responses to music, which is consists of the hippocampus, amygdala, cingulate cortex, and hypothalamus. The hippocampus remembers songs and related experiences and contexts. The amygdala is mainly responsible for emotional responses to music, while the prefrontal cortex and cingulate cortex participate in behavioral decision evoked by music. The coordination of our brain regions allows us to dance to music and to feel and express our emotions. To enable cognitive robots to perform similar functions, we design a simple brain-inspired system based on the proposed audio-semantic model. In the aspect of obtaining the labels of music, our work's idea is different from the methods like [11, 14] based on the multi-label classification. Ours draws on some mechanisms of the human brain, using the distributed attractor network to activate multiple semantic tags.

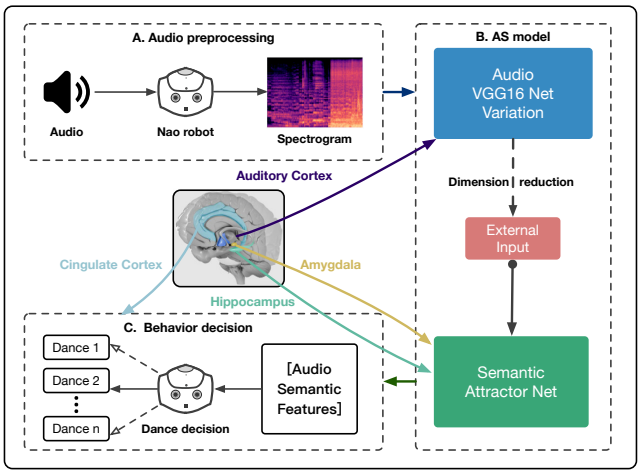

Fig. 1. System architecture.

2 System Design

This section shows the overall architectural design of the system, as shown in Fig. 1. In this paper, we apply the proposed system on the NAO robot, which is a widely used programmable humanoid robot designed by Aldebaran Robotics.

Part A is the preprocessing module of audio. For input music, the 6 seconds time window is used to intercept music into segments, and we apply the short-time Fourier transform (STFT) to each music segment to obtain its mel-scale

spectrogram, and regard it as the input of the audio-semantic model. Part B is the proposed audio-semantic model for mapping the high-level auditory pattern of music to the corresponding semantic features, which will be described in detail in section 4. Part C is the behavioral decision module of the system, roughly corresponding to the functional role of prefrontal cortex and cingulate cortex. It is used to make related dance decisions after receiving the semantic features of music. We make decisions by comparing the cosine similarity with the semantic features of different dance types, and then randomly select a dance for display in that type.

3 Data Acquisition

GTZAN music genre dataset is utilized as the input source of music, which contains blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock in 10 genres, each of which contains 100 music clips with a length of 30 seconds. We use 6 seconds time window, 1.5 seconds offset to intercept each song, and obtained a total of 17,000 samples of 6 seconds in length. Each sample gets a mel-scale frequency spectrogram through STFT. Because we use CNN network to process audio like images, we copy the transformed data into three channels, and then divide the training set and validation set by 7:3 to train the audio network. In order to complete the designed experiment and extract semantic commonness from the original features of music, it is necessary to tag music with corresponding semantic labels consists of emotions, characteristics, and contexts (ECC). Thus a song can be represented in semantic vector space, as shown in Fig. 2.

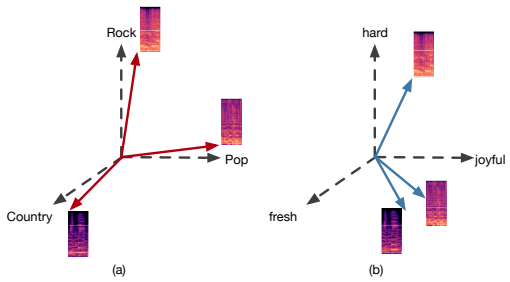

Fig. 2. Representation space for music semantics. (a) shows the representation space of music genre labels. The vectors of music with different genres in this space are orthogonal relations. (b) shows representation space of music ECC features. The vectors of music, which express similar emotions, are closer in this space.

We develop a multi-user online tagging system. For the songs that need to be labeled, we do not show any visible features to the participates, who are required to label the songs only by listening. In order to reduce the extra workload, we

only randomly select 10 songs of each category, 100 songs in total, and label the content of the first 12 seconds for each song, the songs are intercepted to 6 seconds with 2 seconds offset as training data, in a total of 400. Moreover, 12 to 18 seconds of each song as testing data with the same tags, in a total of 100. We provide 50 tags in three categories: Emotion (such as *fresh, joyful, sad*), Characteristic (such as *fast-pace, guitar, piano*), and Context (such as *dinner, morning, working*), named ECC[3] tags.

Participants are three males and three females, a total of 6 non-music professionals aged 20 to 28 years old. The language of tags is the native language of each participant, which is later uniformly translated into English. Each annotator labels the same 100 song segments with a length of 12 seconds in random order. Finally, for each labeled song, the tag with term frequency equal or greater than three will be included as the ECC semantic feature of the song.

4 Model and Method

Audio-semantic model is the core of the system. The model consists of two parts, which are responsible for mapping high-level auditory perceptual patterns to the activation pattern representing semantic features of music. The structure is shown in Fig. 3.

4.1 VGG16 Network for Audio Processing

In order to process music, we construct a deep convolutional network based on the VGG16 as shown in Fig 3 part A. The original network structure consists of 16 layers including the convolution (*Conv*) layer and the full connection (*FC*) layer. We remove the original *FC* layers after *Conv13* layer, replace with two lower dimension *FC* layers with $ReLU$ activation function, and add the Dropout with the probability of 0.5. Finally, the output layer of 10 nodes is added, which corresponds to the ten genre labels of music. The pre-trained VGG16 on ImageNet dataset is used to do transfer learning for our task. We freeze the parameters of the first five layers of the network and then carry out fine-tuning. We use Adam optimizer and set the learning rate to 0.001, the batch size to 128 and training epochs to 30. The network is trained with the data set described in section 2. The VGG16 is regarded as the music feature extractor. Thus we do not need the classification result of the VGG16 network, and only keep the output value of the penultimate layer (FC15) without activation function. In order to get data used for driving the semantic attractor network in part B, we feed the labeled 100 pieces of the 6-second song into the trained audio network. For 100 samples' FC15 output, PCA is used to reduce original 512 dimensionalities to 60, to reduce computation and aggregate effective features. The dimensionality-reduced data is used as the audio input of the semantic attractor network for training.

[3] The full ECC tags can be obtained via https://github.com/kevinleeex/DTME

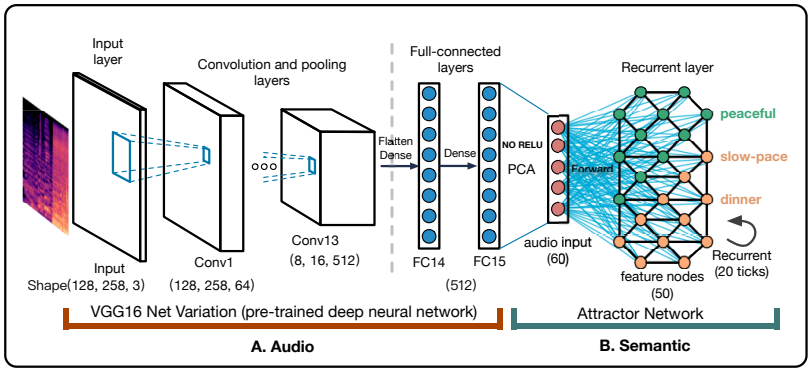

Fig. 3. Illustration of audio-semantic model structure.

4.2 Attractor Network for Semantics

Music can activate brain mechanisms related to semantic processing, as language does [8]. Research on concept processing using a feature-based distributed semantic model shows that statistical structural similarity of semantic and conceptual features between objects can explain a series of behavioral and neuroimaging data [6, 10]. In this paper, we use attractor networks to obtain the semantic features corresponding to music. Attractor network [7] is a dynamic recurrent network which evolves into the stable pattern over time. We use the given ECC feature tag set to construct attractor points, in a total of 50. Emotional features correspond to the amygdala, while characteristic and context features correspond to the hippocampus. The audio input nodes are fully forward connected to the attractor network, and the internal nodes of the attractor network are connected with each other. We train the attractor network to learn the corresponding binary patterns from the input music, where '1' represents the presence of this feature, and '0' represents the absence. Moreover, we use the cross-entropy of the desired activation pattern and actual activation pattern as the loss function and use back-propagation through time (BPTT) with 20 time-ticks iteration for training. The neuronal computation process is described in [6]. We use the first 12 seconds with 6 seconds offset labeled data of music clips to train the attractor network, and training will stop until 95% nodes' activation value reaches more than 0.7. The network uses AdaGrad optimization method with learning rate $\eta = 0.02$, which can dynamically adjust the learning rate and is more suitable for sparse pattern learning. The input of each epoch is a random sample sequence, and weight will start adaptation after five time-ticks.

5 Experimental Results

5.1 Model Results

For the audio part, under the task of the music genre classification, the VGG16 variant network achieves accuracy of 95% in the training set and accuracy of

92% in the validation set, and it performs much better than the traditional method such as *Decision Tree, Logistic Regression, Random Forest, and SVM* with manually extracting features (including spectral centroid, spectral roll-off, zero-crossing rate, RMS, and onset strength).

For the semantic part, the average activation rate of the model reaches 95% after 110 epochs. We use 12 to 18 seconds of 100 songs for testing and end up with an average activated rate of about 71%. On the one hand, some sudden changes in the music style may affect the testing result. On the other hand, due to the subjectivity of music evaluation, it is impossible for human beings to evoke precisely the same emotions and memories, even when facing the same song. Therefore, it is reasonable to some extent for robots to make different choices from some of us.

5.2 System Results

We integrate the trained AS model into the system and deploy the system on the NAO robot. We program some dance clips for the robot with its developer kit. Music and dance often need to express the same emotions. In this paper, dance clips are classified into four types, and corresponding semantic features are tagged with ECC feature set, see Fig. 4. Then, we select four types of songs from the GTZAN dataset, and randomly select a song in each type and ensure that the song did not participate in the training of semantic attractor network. Each song captures the first 6-second segment, and then the segments are spliced into a 24-second testing clip[4], as shown in Fig. 5 (a).

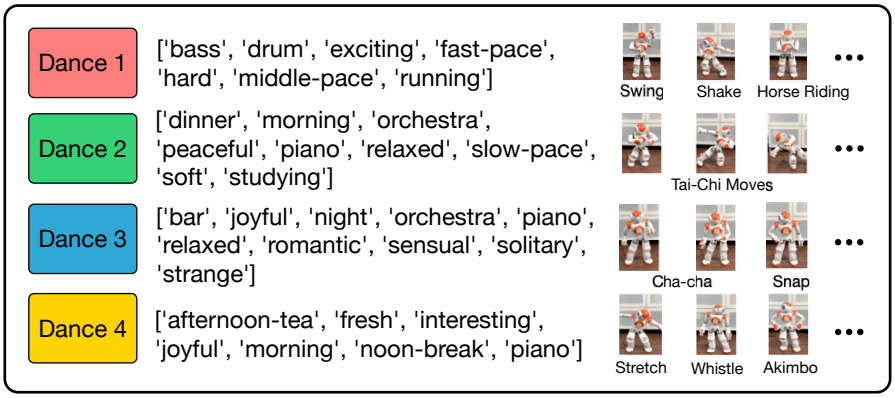

Fig. 4. Illustration of robot dance clips in different styles with semantic features.

Scanning the testing clip with the 6 seconds time window and 2 seconds offset to get the corresponding semantic features and make behavioral responses. Each

[4] Readers can download a copy for listening via https://github.com/kevinleeex/DTME/blob/master/assets/testing_clip.wav

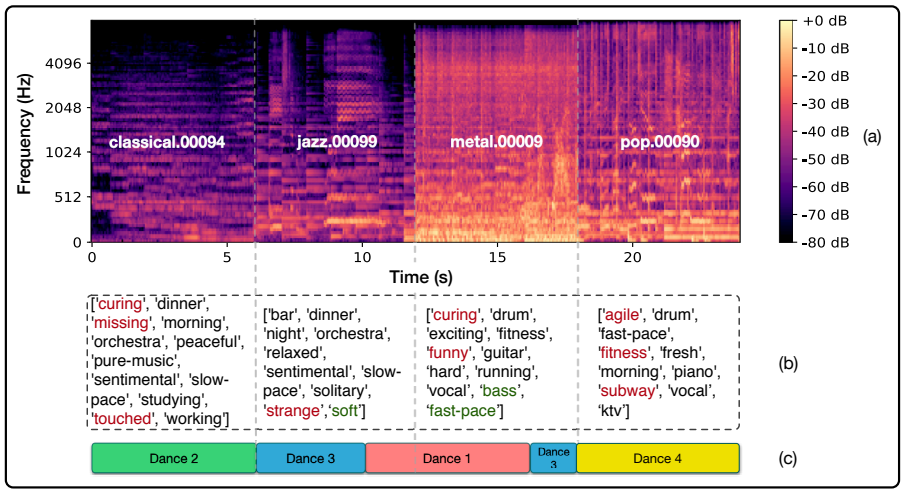

Fig. 5. Illustration of the results on testing clips. (a) shows the mel-spectrogram of the testing music clips. (b) denotes the associated semantic features related to the music. (c) shows the dance decisions made by the robot.

test segment is fed to the system to obtain the corresponding output, as shown in Fig. 5 (b) and (c). Fig. 5(b) shows the semantic output corresponding to the four segments with an interval of 6 seconds. The tags marked in black and red are the output value of the model, the reds are the wrong output tags, the greens are the correct tags but do not appear, and the blacks are the consistent output with the provided by the annotators. (c) shows the dance decisions made by comparing the cosine similarity with the emotion and memory features evoked by music, and the dance with the highest similarity score will be performed.

6 Conclusion

This work inspired by the functional structure of Limbic system of human brain constructs a system for cognitive development of robots, which based on the proposed model of audio mapping to semantics, to realize that robots can evoke emotions and related memories from music. Moreover, behavioral decisions are made through the similarity comparison between music semantics and dance semantics in their bag-of-words vector representation, and then the robot will dance to the music as feedback. The working principle of the model is described in detail, and the experimental results show the effectiveness of the model. However, the audio part of the proposed model lacks some biological plausibility, and we can consider using the spiking neural network (SNN) to enhance the biological plausibility of the AS model, while it is suitable to capture the spatial-temporal patterns and is advantaged in dealing with sound coding and learning [16]. Furthermore, the decision module can be constructed more sophisticated according to specific application scenarios.

References

1. Aly, A., Griffiths, S.S., Stramandinoli, F.: Metrics and benchmarks in human-robot interaction: Recent advances in cognitive robotics. Cognitive Systems Research **43**, 313–323 (2017)
2. Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., Yoshida, C.: Cognitive developmental robotics: A survey. IEEE Transactions on Autonomous Mental Development **1**(1), 12–34 (2009)
3. Broekens, J., Heerink, M., Rosendal, H.: Assistive social robots in elderly care: a review. Gerontechnology **8**(2), 94–103 (2009)
4. Cabibihan, J., Javed, H., Ang, H.M., Aljunied, S.M.: Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. International Journal of Social Robotics **5**(4), 593–618 (2013)
5. D. Fischl, K., B. Cellon, A., C. Stewart, T., K. Horiuchi, T., Andreou, A.: Socio-emotional robot with distributed multi-platform neuromorphic processing : (invited presentation). pp. 1–6 (03 2019)
6. Devereux, B., Clarke, A., Tyler, L.K.: Integrated deep visual and semantic attractor neural networks predict fmri pattern-information along the ventral object processing pathway. Scientific Reports **8**(1), 10636 (2018)
7. Hinton, G.E., Shallice, T.: Lesioning an attractor network: Investigations of acquired dyslexia. Psychological review **98**(1), 74 (1991)
8. Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., Friederici, A.D.: Music, language and meaning: brain signatures of semantic processing. Nature Neuroscience **7**(3), 302–307 (2004)
9. Masuyama, N., Islam, M.N., Seera, M., Loo, C.K.: Application of emotion affected associative memory based on mood congruency effects for a humanoid. Neural Computing and Applications **28**(4), 737–752 (2017)
10. Nishida, S., Nishimoto, S.: Decoding naturalistic experiences from human brain activity via distributed representations of words. NeuroImage **180**, 232–242 (2017)
11. Oramas, S., Nieto, O., Barbieri, F., Serra, X.: Multi-label music genre classification from audio, text, and images using deep features. arXiv preprint arXiv:1707.04916 (2017)
12. Tang, H., Huang, W., Narayanamoorthy, A., Yan, R.: Cognitive memory and mapping in a brain-like system for robotic navigation. Neural Networks **87**, 27–37 (2017)
13. Tikhanoff, V., Cangelosi, A., Metta, G.: Integration of speech and action in humanoid robots: icub simulation experiments. IEEE Transactions on Autonomous Mental Development **3**(1), 17–29 (2011)
14. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music by emotion. EURASIP Journal on Audio, Speech, and Music Processing **2011**(1), 4 (Sep 2011)
15. Warren, J.D.: How does the brain process music. Clinical Medicine **8**(1), 32–36 (2008)
16. Xiao, R., Yan, R., Tang, H., Tan, K.C.: A spiking neural network model for sound recognition. In: Sun, F., Liu, H., Hu, D. (eds.) Cognitive Systems and Signal Processing. pp. 584–594. Springer Singapore, Singapore (2017)