Information Sciences 297 (2015) 271–282

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Analysis of music/speech via integration of audio content and functional brain response

Xiang Ji [a], Junwei Han [a,*], Xi Jiang [b], Xintao Hu [a], Lei Guo [a], Jungong Han [c], Ling Shao [d], Tianming Liu [b]

[a] *School of Automation, Northwestern Polytechnical University, Xi'an 710072, China*

[b] *Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, GA, USA*

[c] *Civolution Technology, Eindhoven, The Netherlands*

[d] *The University of Sheffield, Sheffield S1 3JD, UK*

ARTICLE INFO

Article history:
Received 23 June 2014
Received in revised form 2 October 2014
Accepted 5 November 2014
Available online 22 November 2014

Keywords:
Audio analysis
fMRI
fMRI-measured feature
Multi-view learning

ABSTRACT

Effective analysis of music/speech data such as clustering, retrieval, and classification has received significant attention in recent years. Traditional methods mainly rely on the low-level acoustic features derived from digital audio stream, and the accuracy of these methods is limited by the well-known semantic gap. To alleviate this problem, we propose a novel framework for music/speech clustering, retrieval, and classification by integrating the low-level acoustic features derived from audio content with the functional magnetic resonance imaging (fMRI) measured features that represent the brain's functional response when subjects are listening to the music/speech excerpts. First, the brain networks and regions of interest (ROIs) involved in the comprehension of audio stimuli, such as the auditory, emotion, attention, and working memory systems, are located by a new approach named dense individualized and common connectivity-based cortical landmarks (DICCCOLs). Then the functional connectivity matrix measuring the similarity between the fMRI signals of different ROIs is adopted to represent the brain's comprehension of audio semantics. Afterwards, we propose an improved twin Gaussian process (ITGP) model based on self-training to predict the fMRI-measured features of testing data without fMRI scanning. Finally, multi-view learning algorithms are proposed to integrate acoustic features with fMRI-measured features for music/speech clustering, retrieval, and classification, respectively. The experimental results demonstrate the superiority of our proposed work in comparison with existing methods and suggest the advantage of integrating functional brain responses via fMRI data for music/speech analysis.

1. Introduction

With the rapid growth of online music/speech tracks, the effective clustering, retrieval, and classification of audio data are highly in demand. Most conventional methods mainly rely on low-level features derived from audio stream. However, their capability is still far from being satisfactory due to the well-known semantic gap between the low-level acoustic features used by machines and the high-level semantics perceived by the human brain's cognitive system. Therefore, the performance of audio analysis approaches still has room for improvement.

[*] Corresponding author. Tel.: +86 18710399679.
E-mail address: junweihan2010@gmail.com (J. Han).

http://dx.doi.org/10.1016/j.ins.2014.11.020

In general, the current studies mainly focus on two aspects to alleviate this problem. In the first aspect, more effective features have been extracted for improving audio representation. For example, Mel-frequency cepstral coefficient (MFCC), spectral centroid, spectral flux, and spectral spread are popular features to describe timbre. Key strength is the feature extracted to describe tonality. Recent surveys of these acoustic features can be found in [14,9]. In practice, these features can represent audio individually, or we could combine them or their statistical measurements such as mean and standard deviation to form a feature vector for audio processing [29]. Recently, Lee et al. [31] proposed a novel audio fingerprint search algorithm for music retrieval where the audio fingerprint is a concise digital summary of an audio object and computed by fast Fourier transform (FFT). Fujihara et al. [15] proposed a new representation of the singing voice via the accompaniment of sound reduction and reliable frame selection in order to improve the performance of automatic singer identification. Ren and Jang [40] performed a spoken language identification technique to tokenize each music piece into a sequence of acoustic segment model (ASM) indices. By representing each music piece as the weighted occurrence frequencies of all the mined time-constrained sequential patterns (TSPs), linear SVM classifiers were employed to facilitate the classification task. Tsunoo et al. [43] calculated rhythmic pattern information and bass pattern information to classify musical genre/style.

In the second aspect, many machine learning algorithms have been proposed to improve the accuracy of audio processing. For example, Esling and Agon [11] proposed an innovative way for content-based audio classification and retrieval, which queries generic audio databases by simultaneously optimizing the temporal evolution of multiple spectral properties. Lo et al. [33] presented an index structure, named Two-Tier Multi-Feature Index, which is more efficient and scalable than existing index structures for music database retrieval. Khunarsal et al. [26] proposed an environmental sound classification algorithm using spectrogram pattern matching along with neural network and K nearest neighbor (KNN) classifiers, which can avoid the problem of filtering less informative and irrelevant frequencies in the classification step. Miotto and Lanckriet [37] added the Dirichlet mixture models to refine the semantic multinomial (SMN) results, where SMN was computed using low-level acoustic features such as MFCC via existing auto-tagging systems. The survey papers [14,9] gave comprehensive summaries of technologies of audio retrieval and classification.

The above methods mainly focus on analyzing music/speech from digital audio content and are considered as "computer-centered". They typically ignore the human brain behaviors in audio understanding. Actually, the human brain is the recipient, analyst and evaluator of the acoustic information [38]. Brain responses to audio data contain abundant information associated with the semantic interpretation of multimedia contents. In contrast to "computer-centered" audio content analysis, the information of brain reaction is implicit and "human-centered". Recently, there is an emerging research trend of applying the implicit information to multimedia analysis. For instance, Hadjidimitriou and Hadjileontiadis [17] extracted features from the electroencephalogram (EEG) [3] responses when subjects were listening to music, and by using these features, the music clips were classified into two categories, i.e., "like" and "dislike". However, EEG usually measures the brain response with the electrodes placed around the brain scalp, which has limitation of the spatial resolution and is unable to comprehensively capture the whole-brain semantic comprehension. In contrast, the development of fMRI technology provides an alternative way to monitor and probe the brain responses to natural stimuli. For example, the study in [13] reported that the acoustic signals perceived by the human brain can be effectively decoded by fMRI signals. Another work in [42] demonstrated that sound category information could be effectively detected with multi-voxel pattern analysis. Leaver and Rauschecker [30] gave some "category-selective" voxels which have greater fMRI signal for a single category of audio than other categories. The above studies provide us the neurological theoretical basis that the acoustic categories can be effectively represented by the brain responses within the auditory and other related brain regions including the attention, emotion, and working memory systems. Besides these studies, the human-centric features derived from fMRI data have already been applied to video/image analysis and achieved encouraging performance [23,19]. For instance, Hu et al. [23] proposed a novel video classification framework by using fMRI-measured features. Han et al. [19,20] applied the fMRI-measured features to video retrieval and abstraction. These studies also motivate us to apply the fMRI-measured features revealing brain responses to improve the capability of audio processing. To the best of our knowledge, this topic has been rarely studied.

In this paper, we adopt fMRI brain imaging to capture the human brain's comprehension and cognition to different types of music/speech stimuli. We propose a novel framework for effective music/speech clustering, retrieval, and classification by integrating the computer-centered acoustic features with human-centered brain response features, where brain response features are derived from fMRI data when participants are listening to the music/speech. Specifically, in the proposed framework, the brain response features are calculated by using the functional connectivity matrix among brain ROIs and feature selection methods. The integration of two types of features is implemented using the proposed ITGP algorithm that can predict the brain response features for samples without fMRI data and multi-view learning algorithms.

The work in this paper is a substantial extension of our previous study in [25] and there are several major novelties and differences as follows.

(1) Instead of using features individually, this paper constructs an effective audio processing framework to integrate the strengths of acoustic features and fMRI-measured features. The motivation to combine these two features is that some audio tracks have higher intra-class similarity and lower inter-class similarity in the fMRI-measured feature space compared with that in the acoustic feature space. Fig. 1(a) shows such a positive exemplar audio track for fMRI-measured features. In this figure, the acoustic feature and the fMRI-measured feature are both extracted from a classical music clip. In the acoustic feature space, the similarities between this sample and all of other audio excerpts are calculated. Then intra-class similarity of this classical music sample is defined as the average similarity between it and