Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 547–552

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

Bio-inspired Approach for Automatic Speaker Clustering Using Auditory Modeling and Self-Organizing Maps

Anton A. Yakovenko and Galina F. Malykhina

Peter the Great St.Petersburg Polytechnic University, Saint Petersburg, Russia
yakovenko_aa@spbstu.ru, malykhina@ftk.spbstu.ru

Abstract

This paper presents a biologically inspired approach to the problem of voice biometrics. The aim of this study is to examine the capacity of the automatic system, based on a physiologically appropriate auditory model and self-organizing neural networks, to distinguish the voices of different speakers. The idea stems from the human ability to successfully extract various information from speech in the process of verbal communication in different acoustic conditions, including recognizing the identity of a familiar person by their voice. Based on the obtained results, one can conclude that the proposed method has demonstrated high-quality unsupervised classification.

Keywords: cluster analysis, speaker clustering, auditory perception, neural activity, self-organization, text-independence, speech mining, MAP, SOM

1 Introduction

Speech is a unique human ability that represents a universal method of communication. With developing technology, automatic speech processing and intelligent analysis has naturally become relevant in various spheres of application. However, a speech signal that contains a set of multi-level information [1] is susceptible to manifold distorting effects [2]. For that reason, many tasks still involve negotiating the variability of speech utterances during automatic processing. Nevertheless, despite any distorting factors, a human can effectively distinguish voice information in the process of verbal communication in a large scope of sound environments, especially if the voice of an individual is familiar to the perceiver [3]. Thus, the question is how computational analysis of speech samples could help to solve the problems of variability of speech and correctly process the necessary information type. Understanding the mechanisms of perception, creating and assessing such models can play an important role in this context. Based on the fact that neural responses are robust to the presence of noisiness in stimuli [4], in this study, the approach to intelligent computational analysis of voice data and its features is considered from the standpoint of perception and psychoacoustics.

10.1016/j.procs.2018.01.083

In everyday life we regularly experience biometric information analysis, which is a unique characteristic of every human being that consists in complex perception of the most distinctive behavioral and physiological modalities. Voice can be identified as one of these modalities. For instance, a human draws conclusions about the personality of the subject, basing their judgment on perceived speech, which is expressed in their ability to discern familiar and unfamiliar voices or to recognize a familiar individual by their voice, regardless of the speech context – in other words, performing a classification of unlabeled speech samples. In the domain of automatic speech processing, there is a similar task, namely voice or speaker recognition, specifically, speaker clustering. It represents a task of speech biometry that implies an independent process of recognizing the personality of an unknown speaker by means of a computational system based on a provided voice sample. This task is based on extraction and analysis of features, speaker model development and comparison as well as decision-making about voice attribution of a specific speech sample. Many modern systems of this kind rely upon most common acoustic features of a speech signal. There are effective methods designed for this approach towards signal representation. However, given that the acoustic representations comprise all kinds of speech information, it is extremely difficult to identify its specific type [1]. Besides, such signal characteristics are by nature greatly susceptible to noises, speech variability and other distortions.

2 Methodology

Perceiving, processing and extracting appropriate information from a variable multilayered ensemble of acoustic data are remarkable features of the auditory system and the brain. The human perceives and assimilates the flux of speech entering at the input of the auditory analyzer in the form of acoustic oscillations, despite their inconsistent nature and the difficulties related thereto. Similarly, it is expected from the computational system to overcome the speech variability factors and the ability to extract features that support the tasks of intelligent speech analysis. Various modern methods supporting extraction of voice features from a complex speech signal demonstrate their efficiency [3]. However, taking into account the physiology of auditory perception, it is obvious that they do not reflect the complexity and integrity of this process. Accordingly, this paper proposes an alternative method for extracting features that is based on simulation of neural responses by the auditory periphery model, which correspond to electrical activity in the process of auditory perception. For unsupervised classification of unlabeled speech data an artificial neural network modeling approach based on self-organization is proposed. The task of this subsystem is to learn the presented speech utterances, transformed into vectors of firing rate probability, and subdivide them into the corresponding clusters by voice attribution. This type of classification problem deals with processing of a set of features that can help to recognize a particular class. Due to the complexity of the audio data and neural activity images, it is necessary to have a nonlinear model with a good generalization ability. The strengths of the artificial neural network approach are related to its ability to adapt, generalize and learn without any prior knowledge of the data [5]. The main stages of the proposed method are described below, the general scheme of the architecture is shown in Figure 1.

2.1 Simulation of the Auditory Nerve Responses

It is known that the auditory periphery converts sound oscillations into responses of neural activity. A waveform is the most common digital representation of speech, as it reflects the