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

Anton A. Yakovenko and Galina F. Malykhina
Peter the Great Saint-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 the voice biometrics. Aim of this study is to examine capacity of the automatic system, based on a physiologically appropriate model of the auditory periphery and self-organizing neural networks, to discriminate voices of different speakers, through the analysis of the provided text-independent speech samples. The idea stems from the human ability to successfully extract a various information from speech in the process of verbal communication in different acoustic conditions, including recognizing the identity of a familiar person by his voice. In the proposed method, speech signal is processed through the computational model of the auditory periphery, which simulates neural responses of auditory-nerve fibers. From the obtained new signal representation feature vectors are formed, by which neural network will be trained to generate voice-clusters of different speakers. Based on the obtained results, one can conclude that the proposed method has demonstrated high quality of unsupervised classification of speakers by their voices.

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 naturally became 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, as well as creating and assessing such models, can play an important role in this context. Based on the fact that neural responses are robust to presence of noisiness in stimuli [4], in present study, the approach to intelligent computational analysis of voice data and its features is considered from the standpoint of perception and psychoacoustics.

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 cluster analysis of unlabeled speech samples. In the domain of automatic speech processing there is a similar task, named voice or speaker recognition, specifically speaker clustering. It represents a task of speech biometry that implies an independent process of recognizing the personality of the unknown speaker by the computational system based on a provided voice sample. In general, speaker recognition can be formulated as identification and verification in a text-dependent or text-independent set of speech data. Regardless of formulation, 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. Within this approach, the mentioned disadvantage is compensated by the universal background model (UBM) – a model used for statistic description of general acoustic space that allows for diversity of voice features and acoustic environments [5]. In order for the UBM to have necessary generalizing qualities, learning requirements are to be introduced that include balance and high diversity of recording conditions, communication channels, language as well as lineup and number of speakers, which imposes severe limitations on the application of this approach.

2 Methodology

On the other hand, 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 fullness of this process. Accordingly, in this paper is proposed an alternative method for extracting features that is based on simulation of neural responses in the model of the auditory periphery, 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. This sort of classification problem deals with the 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 connectionist artificial neural network approach are related to the ability to adapt, generalize and learn without any prior knowledge of the data [6]. The main steps of the proposed method are described below, the general scheme of the proposed architecture is shown in Figure 1.

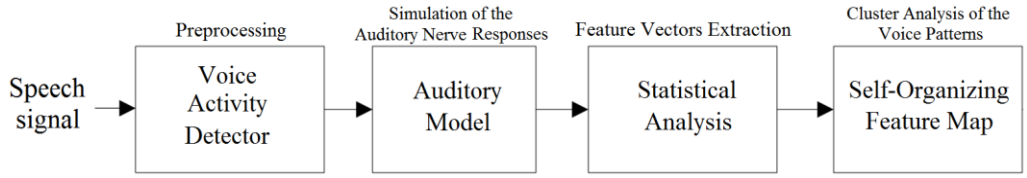

Figure 1: Block diagram of the proposed architecture

2.1 Simulation of the Auditory Nerve Responses

It is known that the auditory periphery converts sound oscillations into responses of neural activity. Waveform is the most common digital representation of a speech signal, which reflects the structure of sound pressure over time. The duration of used speech utterances is 10-15 seconds, due to the fact that even a short sequence of a speech signal contains a complex set of information [7]. Such acoustic stimuli waveforms are fed to the computational model of the auditory periphery. Subsequent stages of modeling reflect the mechanical behavior of the auditory system and the nonlinear processes occurring in the inner ear.

As a tool for this study the Matlab Auditory Periphery (MAP) model, version 1.14, from Essex University, was chosen. This physiologically-based model of the mammalian auditory periphery system has been extensively used by researchers to explore a variety of scientific questions and various applications. The vesicle release mechanism at the auditory nerve-inner hair cell synapse is described in [8]. However, the model has evolved considerably and improve. Its current version is a modified and is stochastic. In general, processing takes place in several stages: filtering the input signal to reproduce the properties of the response of the middle ear, the filtering of basilar membrane (BM) as a bank of the dual resonance nonlinear filters [9], simulation the mechanisms of inner hair cells and firing of the auditory nerve (AN). The version of the model used in this study has some working assumptions. In the mechanism of generation of the AN response, it is assumed that for the occurrence of an action potential, it is sufficient to release a transmitter into the synaptic cleft by a single vesicle. Also, in probabilistic computation of the model, no adjustment is made for the relative refractory period.

In present study, the MAP model simulates the response of ensembles of AN fibers with high spontaneous rate as a probability that represents the average of the ensemble. Responses, in the shape of neural activity image, is represented as firing rate, which is based on the amount of the transmitter in the synaptic cleft:

$$AN_{rate} = \frac{c(t)}{dt},$$

where c(t) is the number of vesicles in the cleft at time *t*.

Absolute refractory period is taken equal 0.75 ms. The probability of spikes occurring at this period can be estimated as follows:

$$P_{fired} = 1 - \prod (1 - P_{t-1}) \quad (1)$$

The right-hand side of equation (1) indicates the product of all probabilities of not firing during the refractory period *[t; t – 1]*. The reduction in the firing probability at the current time interval *t* is proportional to the likelihood that a spike occurred during this period.

Each site on the BM has a response tuned to a particular frequency, termed as the best frequency (BF). The model processes 21 channels with BFs between 250 Hz – 8 kHz. For the input speech signals, at the output, MAP model generates multidimensional matrices containing neural activity images in the form of firing probability for each BF channel, which corresponds to the information processing in the AN. Such transformation is performed for each speech sample of each speaker from the dataset.

2.2 Feature Vectors Formation

Each speaker can be represented from the perspective of perception, as a voice pattern. Obtained in the previous step auditory images are the variations of such patterns. However, resulting matrices are too large, differ in dimension and contain a lot of complex statistical information. Therefore, it is necessary to preprocess the obtained data to form the input space. To reduce the complexity of the data and to extract feature vectors, representing voice patterns, serving as input to the neural network on the next stage of processing, statistical analysis is performed. For these purposes the standard deviation measure that approximates the average distance d_i from the mean $\bar{x}$ was applied:

$$\varsigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}, \tag{2}$$

where $d_i = \left(x_i - \bar{x}\right)$ and $\bar{x} = \frac{\sum x_N}{N} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

The standard deviation is a descriptive measure reflecting the distribution of the data. Measure (2) provides important information about statistical distribution of firing rate at each BF channel for speech utterances of different speakers in vector form. Resulting vectors, containing statistical voice patterns, was created for each utterance from dataset. The size of vectors corresponds to the number of BFs. Collected vectors build up the input data for cluster analysis by neural network. Figure 2 shows an example of obtaining neural activity image for acoustic stimulus and comparison of the resulting distributions of voice patterns for two different speakers.

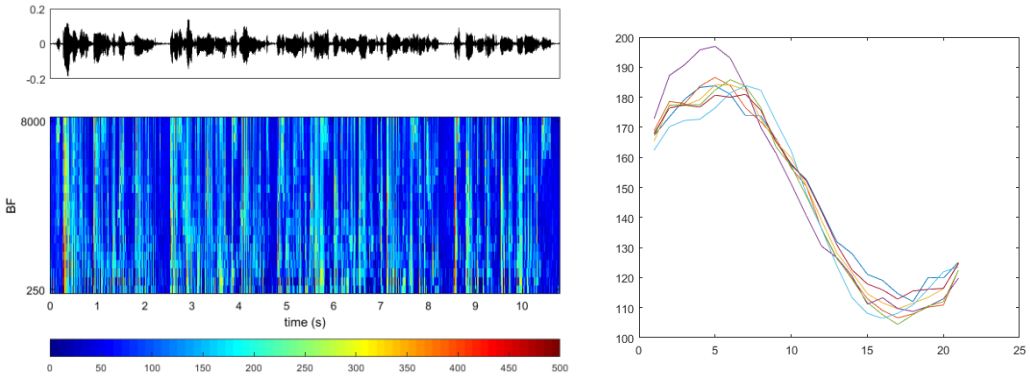

Figure 2: Simulation of neural activity for speech utterance (left panel) and its distribution by 21 BF channels for 7 speech utterances by single speaker (right panel)

2.3 Neural Network Modeling

Self-organizing feature map (SOM) implements a clustering process [6]. As input for the neural network, resulting firing probability vectors are used. To provide the self-organization process more efficient, the compressed data on the base of clustering was offered. In spite of high efficiency the k-

means algorithm cannot discriminate certain clusters correctly (i.e., that goes for samples Lsum, Target, WingNut, Chainlink). SOM algorithm includes the following steps:

1. Obtaining the training vector from input space:
$$\mathbf{x} = [x_1, x_2, ..., x_m]^T$$
2. Algorithm initialization by random selection of a set of weights $\{\mathbf{w}_j(0)\}_{j=1}^{m}$ from the available set of input vectors $\{\mathbf{x}_i(0)\}_{i=1}^{N}$, when $j = 1,2..., m, m -$ total number of neurons in the lattice.
3. Finding the winning neuron at the step *n* using criterion of the Euclidean distance minimization:
$$i(\mathbf{x}) = \arg\min_j \|\mathbf{x} - \mathbf{w}_j\| \quad (3)$$
4. Updating the weights of the winning neuron and its neighbors:
$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(\mathbf{x})} (\mathbf{x}(n) - \mathbf{w}_j(n)), \quad (4)$$
where chosen neighborhood function is defined as:
$$h_{j,i(\mathbf{x})}(n) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right) \quad (5)$$

A set of feature vectors is saved in database for every speaker to form the voice-pattern. During the clustering process a set of vectors of testing speaker compares with vectors from database of reference speaker. A test speech utterance refers to a class of speakers for which it has a minimum distance:
$$\Im(k) = \min_k d(\mathbf{X} - \mathbf{X}^k) = \min_k \left(\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{x}_j^k\| \right), \quad (6)$$
where $\mathbf{X}, \mathbf{X}^k$ are the sets of vectors of current speaker utterance and of *k*-th cluster respectively.

3 Experiments

The purpose of experimental studies is to find out how well biologically inspired computational model of speech perception can learn to classify speakers by their voices based on different unlabeled speech data, by analogy with the natural human ability. The following sections describes the experimental data and the results of tests carried out.

3.1 Data Description

In present study was used ELSDSR corpus of read speech [10], which has been designed to provide speech data for automatic speaker recognition task. The speech signals was recorded in a relatively noise free environment. In total, it consists of 198 examples of speech utterances by 22 speakers: 12 male and 10 female. The source data is provided in the WAV format. The sampling frequency is 16 kHz with a bit rate of 16. Before extraction of the short-time segments for further processing, data was preprocessed by voice activity detector algorithm to remove the silence periods from speech signals. Silence removal and segmentation is based on signal energy, spectral centroid and thresholding criterion [11].

3.2 Performance Evaluation

Preliminary, the data set was normalized by columns of the resulting matrix of the input space. By means of z-score, the standardized data set has mean 0 and standard deviation 1. This measure allows to estimate the distance of a data point from the mean in terms of standard deviation:

$$z = \frac{\left(x - \overline{X}\right)}{S}, \tag{7}$$

where x – data point, $\overline{X}$ – mean value and S – standard deviation of the sample data.

Clustering of different samples was carried out with various parameters of the SOM algorithm to determine their optimum values, including number of sensory neurons, iterations and learning rate. Mean squared error (MSE) is used to evaluate the accuracy performance of the proposed method:

$$MSE = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} \left\| X_i^{(k)} - C_k \right\|^2, \tag{8}$$

where – number of data points in the cluster k, $X_i^{(k)}$ – i-th data point in the cluster. The values of the estimates thus obtained are shown in the table 1.

Number of speakers	Number of clusters	MSE	Time(ms)
2	2	0.02	15
4	4	0.17	21
6	7	0.33	44
10	12	0.48	65
14	11	0.53	79
18	15	0.60	86
22	26	0.75	93

Table 1: Experimental results

4 Conclusion

The paper proposes a biologically inspired speech mining approach. To overcome the complexities associated with speech variability and multidimensionality of the information presented in it, the approach from the perspective of auditory modeling is considered. The task of clustering voices of various speakers is solved, which reproduces the analogy with the human ability to distinguish perceived voices, and to recognize familiar speakers in text-independent context. The allocation of clusters of total input space is performed using self-organizing feature map neural networks, which in turn also reflects biological aspects. Based on the obtained results, it can be concluded about efficiency of the proposed approach, the model is able to extract extralinguistic information contained in the speech flow. Research in this area can be useful not only for computational processing of speech signals, but also for neurophysiology of the human mind, perception and cognition.

References

[1] Lapteva, O. & Wallmannsberger, J., 2011. Mining Speech: A Computational Model of Human Perception. *International Journal of Data Mining and Emerging Technologies,* 1(1), pp. 14-21.

[2] Klatt, D. H., 1989. Review of selected models of speech perception. In: *Lexical Representation and Process.* s.l.:MIT Press, pp. 169-226.

[3] Hansen, J. & Hasan, T., 2015. Speaker Recognition by Machines and Humans. *IEEE Signal Processing Magazine,* 32(6), pp. 74-99.
[4] Johnson, D. H., 1980. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *The Journal of the Acoustical Society of America,* Volume 68, pp. 1446-66.
[5] Reynolds, D. A., 1997. Comparison of background normalization methods for text-independent speaker verification. *Proceedings of the Eurospeech,* pp. 963-66.
[6] Haykin, S., 2009. *Neural Networks and Learning Machines.* 3rd edition ed. Pearson Education, Inc.
[7] Owens, F. J., 1993. *Signal Processing of Speech.* Basingstoke: The Macmillan Press LTD.
[8] Meddis, R., 1986. Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America,* Volume 79, pp. 702-11.
[9] Meddis, R., O'Mard, L. P. & Lopez-Poveda, E. A., 2001. A computational algorithm for computing nonlinear auditory frequency selectivity. *The Journal of the Acoustical Society of America,* 109(6), pp. 2852-61.
[10] Feng, L. & Hansen, L. K., 2005. *A New Database for Speaker Recognition,* Informatics and Mathematical Modelling, Technical University of Denmark.
[11] Giannakopoulos, T., 2009. *A method for silence removal and segmentation of speech signals, implemented in Matlab,* Athens: University of Athens.