Multiple Microphone Speaker Recognition System for Second Language Based on Biomimetic Pattern Recognition with Big Data Fusion

Guangming Xian*
School of Software, South China Normal University, Foshan Guangdong 528225, China
*Corresponding author

Abstract. The speaker recognition problem for second language based on biomimetic pattern recognition with big data fusion by multiple microphone has been addressed in this paper. Biomimetic pattern recognition is a new machine learning algorithm which can be used to study the geometric characteristics of a large number of sample points in high-dimensional space. Machine learning is an important way for computer to realize intelligence. The development of artificial intelligence can not be separated from the support of machine learning. Big data Unifield architecture of Hadoop system integrates machine learning and big data processing. Several speaker feature extraction methods of big data are described. Cepstrum and ∆ceptrum can offer a significant computational advantage in the eigenvalue problem in result of enhancing correct rate. The big data fusion framework of neural network multiple microphone verification system is presented; in particular, a neural network constructed by multiple value neural network algorithm of big data fusion is also proposed for adjusting the parameters. The experimental results of big data fusion using multiple microphone speaker recognition system for second language illustrate the effectiveness of the proposed biomimetic pattern recognition method.

Keywords: Speaker recognition; Second language; Biomimetic pattern recognition; Machine learning; Big data fusion.

1. Introduction
Speaker recognition (SR) [1-3] is also called voiceprint recognition (VPR), which distinguish the speaker through speech [4-5]. We can identify and authenticate identity by using technology of speaker recognition [6-7].

The difference of speaker recognition and speech recognition is that speaker recognition does not focus on symbol and semantic content information containing in speech signal but focus on personal characteristics containing in the speech signal in order to achieve the purpose of speaker recognition. Speaker recognition has a deep technical background. The origin of human language is a complex physiological process between the human language central and vocal organ. There are differences between two people's voiceprint. The size and morphological aspects of vocal organ, tongue, teeth, throat, lung, and nasal cavity of each person has a lot of differences when people are talking. Each person's voice sound characteristics both have stability and invariance. They are not absolute and immutable. This variation is derived from the physiology, pathology, psychology, simulation,
camouflage, and also associated with environmental disturbance. Because each person's speech organs are not all the same, under normal circumstances, people can still distinguish voices from different people and judge whether the voice belong to the same person.

The basic principle of speaker recognition with big data fusion is to construct a mathematical model for each one by analyzing the acoustic and auditory. And then we perform an exact match for the model and the actual input speech by the computer. We can identify the speaker according to the result of matching. The principle is closely related to the speaker's physiological and behavioral characteristics. The biological characteristics are existing on spectrum surface (i.e., channel characteristics) and are also present in the source of the sound or several discontinuous voice segments. Big data audio feature can be effectively extracted from one of these properties to construct mathematical model. The associated data can be stored in the database. SR server performs data retrieval and accurately matches in database.

Multiple microphone speaker recognition system with big data fusion is the combining of speech data derived from different microphones such that the resulting information is more accurate, complete or dependable than when these microphones are used individually [8-9].

In this article, firstly, a brief introduction of the speaker recognition with big data fusion is given. Secondly, big data feature selection of speaker recognition for second language [10-12] is introduced. Thirdly, the big data fusion function of multiple microphone speaker recognition system is illustrated [13-15]. Fourthly, experimental big data fusion comparison results of different features and big data fusion [16-18] experiments using multiple microphone speaker recognition [19-21] system can be obtained [22-23]. Finally, we draw the conclusion of this paper.

2. Feature Selection of Speaker Recognition with Big Data Fusion

Speaker recognition with big data fusion is extracting the speaker’s personality traits from a speaker voice. Based on the personal feature analysis and recognition, we can achieve the purpose of speaker identification or confirmation for big data fusion.

Speaker recognition system with big data fusion (as shown as figure 1) consists of big data preprocessing feature extraction, big data fusion pattern matching and judgment.

![Figure 1. Sketch of speaker recognition system with big data fusion.](image)

Establishment and application of a speaker recognition system with big data fusion can be divided into two stages, namely the training stage and the stage of recognition. In the training phase, each user of the system speaks some training corpus. The system according to the training corpus, establish each user template or model parameters reference set by training and learning. In the recognition stage of big data fusion, the characteristic parameters extracting from the unknown speaker voice signal was compared to the reference parameter set and model template in the training process. According to the comparison, the user of the reference model with the nearest distance is regarded as input language speaker. Speaker verification feature parameters derived from the input speech is compared to reference volume. If both the distance is less than a specified threshold, then the confirmation, otherwise reject.

Big data feature extraction is one of the most important parts in speaker recognition. It is the process of extracting to the speaker's personality characteristics from speech signal.
Although it is not entirely clear that which parameter can reflect the speaker’s feature. Two aspects are generally included: difference of vocal organs to generate voice (congenital) and movement differences of pronunciation organs while sound-producing (acquired).

The former is mainly manifested in the voice frequency spectrum structure, mainly containing the spectral envelope characteristic information which reflects the vocal tract resonance and antiresonance characteristic. It also includes the spectral detail structural feature information which reflects the sound characteristics of vocal fold vibration. Representative characteristic parameters are cepstrum and pitch parameter (static characteristic). The latter pronunciation habits differences are mainly in the voice frequency spectrum structure on the time changing, containing the main parameters of dynamic characteristics, representative characteristic parameter including cepstrum and the linear regression coefficient of the pitch (dynamic characteristic), namely the difference value of cepstrum (Δcepstrum) and the difference value of pitch (Δpitch). In speaker recognition of big data fusion, spectral envelope characteristics especially cepstrum are often used.

According to the above analysis, in the ideal case, selecting big data feature should satisfy the following criteria:

1) Different speakers can be effectively distinguished and relatively stability can be remained when the same speaker's voice changes.
2) Big data features can be easily extracted from the speech signal.
3) Not easy to be imitated.
4) Try not to vary with the time and space.

3. The Big Data Fusion Function of Multiple Microphone Speaker Recognition System

The “image” being mapped to the big data feature space which is information being preprocessed from the same target collected by multiple microphone. And the image is multiple sample points. The distribution of multiple sample points partly reflects the distribution of speaker in the big data feature space [24-25].

Determining the coverage range of such sample in feature space is the effective step of speaker verification with big data fusion.

Multiple weights neurons can be regarded as complex shape cover area in big data feature space according to a plurality of weight vectors. The generic expression of a multiple weight neural property should be

\[ Y = f[\emptyset(W_1, W_2, \cdots, W_m, X)] \]  

where \( W_1, W_2, \cdots, W_m \) is m weight vectors; X is the input vector; \( \emptyset \) is computing function determining by multiple weight vector neurons (A plurality of vector input and a scalar output). And f is the nonlinear transfer function.

In speaker verification process of big data fusion, the question to be answer is “Yes” or “No”. So step function can be utilized, i.e.

\[ f(x) = \begin{cases} 1, & \text{if } x \leq k \\ 0, & \text{if } x > k \end{cases} \]  

We suppose that big data feature space is the n-dimensional real space \( \mathbb{R}^n \). The vector equation can be expressed as follow.

\[ \emptyset(W_1, W_2, \cdots, W_m, X) = k \]  

where \( k \) is a constant which can be regard as a trajectory of X vector in feature space \( \mathbb{R}^n \) determining by m weight vector constructed by \( W_1, W_2, \cdots, W_m \). This trajectory is n-1 dimensional hypersurfaces (or hyperplane) in \( \mathbb{R}^n \). The trajectory separates \( \mathbb{R}^n \) into two parts: in one side \( \emptyset \geq k, Y = 1 \) and in other side \( \emptyset < k, Y = 0 \).

If \( \emptyset(W_1, W_2, \cdots, W_m, X) = k \) is a closed hypersurface, then a limited coverage area forms in the feature space.

If the input vector X in the coverage area, neuron output \( Y = 1 \). Else if outside the coverage, then \( Y = 0 \).
The “image” being mapped to feature space and being collected by multiple microphone (the number of microphone is m) is vector \( S_1, S_2, \cdots, S_m \). Weight vectors \( W_1, W_2, \cdots, W_m \) are determined by \( S_1, S_2, \cdots, S_m \). That is to say

\[
W_1 = w_1(S_1, S_2, \cdots, S_m) \\
W_2 = w_2(S_1, S_2, \cdots, S_m) \\
\vdots \\
W_m = w_m(S_1, S_2, \cdots, S_m)
\] (4)

Therefore \( \emptyset(W_1, W_2, \cdots, W_m, X) = k \) inevitably can be rewritten into

\[
\emptyset(S_1, S_2, \cdots, S_m, X) = k
\] (5)

Using biomimetic pattern recognition principle with big data fusion and high dimensional space geometry analysis method, we can find out the suitable function for solving specific recognition problem. And when \( X = S_1 \) or \( X = S_2, \cdots, \) or \( X = S_m \), formula (5) can be established. Feature space status covered by neuron is determined by common sample. It plays the role of fusing speech information provided by samples collected by m multiple microphone.

The above is based on neuron and the following hypothesis is based on there have p neurons. Each neuron weight is determined by sample collected once. Each output of neuron is \( Y_1, Y_2, \cdots, \) or \( Y_p \).

\[
Y_1 = f[\emptyset(S_{11}, S_{12}, \cdots, S_{1m}, X)] \\
Y_2 = f[\emptyset(S_{21}, S_{22}, \cdots, S_{2m}, X)] \\
\vdots \\
Y_p = f[\emptyset(S_{p1}, S_{p2}, \cdots, S_{pm}, X)]
\] (6)

Since the input vectors are collected by a plurality of microphones and X also has m vectors which are represented by \( X_1, X_2, \cdots, X_m \), then

\[
Y_{11} = f[\emptyset(S_{11}, S_{12}, \cdots, S_{1m}, X_1)] \\
Y_{12} = f[\emptyset(S_{21}, S_{22}, \cdots, S_{2m}, X_2)] \\
\vdots \\
Y_{1p} = f[\emptyset(S_{p1}, S_{p2}, \cdots, S_{pm}, X_p)] \\
Y_{21} = f[\emptyset(S_{11}, S_{12}, \cdots, S_{1m}, X_1)] \\
Y_{22} = f[\emptyset(S_{21}, S_{22}, \cdots, S_{2m}, X_2)] \\
\vdots \\
Y_{2p} = f[\emptyset(S_{p1}, S_{p2}, \cdots, S_{pm}, X_p)] \\
\vdots \\
Y_{m1} = f[\emptyset(S_{11}, S_{12}, \cdots, S_{1m}, X_1)] \\
Y_{m2} = f[\emptyset(S_{21}, S_{22}, \cdots, S_{2m}, X_2)] \\
\vdots \\
Y_{mp} = f[\emptyset(S_{p1}, S_{p2}, \cdots, S_{pm}, X_p)]
\] (7)

The speaker is determined by the discriminant function as below.

\[
R = \bigcup_{j=1}^{m} \bigcup_{i=1}^{p} Y_{ji}
\] (8)
where R represents that none of the discriminant input vector $X_1, X_2, \ldots, X_m$ is in the union range covered by p neuron $\emptyset$ in big data feature space [26-29]. The union range covered by p neuron $\emptyset$ can form very complex geometry shape in big data feature space. And it can satisfy speaker verification [30-31] with big data fusion [32-36].

4. Experimental Results

We adopt the big data Unifield architecture of Hadoop system, which integrates machine learning and big data processing, and adds machine learning layer in the flow processing layer. The flow layer not only uses the model, but also contains the continuous training of the model. It is applicable to the situation that there is a large amount of data to be analyzed, and machine learning is convenient and has a very large demand or planning.

4.1. Experimental Big Data Fusion Comparison Results of Different Features

Cepstrum and pitch feature is the features commonly used which can get good recognition with big data fusion. The following table gives the experimental big data fusion comparison results of cepstrum and pitch features.

The big data fusion experimental results (table 1) show that the using cepstrum and $\Delta$cepstrum can get better recognition performance than pitch and $\Delta$pitch. Because the stable cepstral coefficients are relatively easy to extract using pitch and $\Delta$pitch. Compared to cepstrum and $\Delta$cepstrum, pitch and $\Delta$pitch only exists in the part of voiced sound. Accurate and stable pitch feature is difficult to extract by pitch and $\Delta$pitch.

Table 1. Experimental big data fusion comparison results of different features

| Selected Feature | Correct rate (%) |
|------------------|------------------|
| cepstrum         | 92.2%            |
| $\Delta$cepstrum | 87.5%            |
| pitch            | 32.8%            |
| $\Delta$pitch    | 25.0%            |

4.2. Big Data Fusion Experiments Using Multiple Microphone Speaker Recognition System

Neural network multiple microphone verification system for big data fusion use four microphones (m=4) and five neurons (p=8) to verified one speaker. The big data fusion framework can be shown as figure 2.

Figure 2. The big data fusion framework of multiple microphone speaker recognition system.

Our big data fusion experiments are performed by 8 people and each person is collected 8 times before combination microphone. We can get a total of 8 groups and each group has 4 kind of sound. Neural network for big data fusion is constructed by multiple value neural network algorithm and k is the adjustable distance constant. We make many times of cross certification to verify the speaker. At different distance constant k, the correct rate, false acceptance rate and false rejection rate can be shown as table 2.
To illustrate the effect of improving correct rate via enhancing coverage area by using multiple $∅$ function neurons. We use the same big data fusion verification method which the neuron network covering the space complexity is simplified. The big data fusion algorithm confirms the speaker with only one (namely $p=1$) neuron and cross validation experiment is performed. The experimental result of big data fusion can be shown as table 3.

From the above experimental results, we can get

(1) If $m = 3$, $p = 8$ and $k^2 = 275$ are selected, the big data fusion recognition correct rate is 96.9%, false acceptance rate is zero and false rejection rate is 3.1%. We can make up for it via verification again. It is illustrated that the proposed biomimetic pattern recognition with big data fusion is preferable for multiple microphone speaker recognition in second language.

(2) We can improve the recognition effect of big data fusion by increasing the number of function neuron namely increasing the complexity covered by neuron.

(3) In order to ensure false acceptance rate equal to zero, we should maintain the distance constant value of $k$ in the range of 250–300 while increasing the value of $p$.

Table 2. Experimental result of big data fusion with eight $∅$ function neurons

| $k^2$ | Correct rate | False acceptance rate | False rejection rate |
|-------|--------------|------------------------|----------------------|
| 200   | 92.2%        | 4.5%                   | 7.8%                 |
| 225   | 90.6%        | 2.0%                   | 9.4%                 |
| 250   | 93.8%        | 0                      | 6.2%                 |
| 275   | 96.9%        | 0                      | 3.1%                 |
| 300   | 93.8%        | 0                      | 6.2%                 |
| 325   | 93.8%        | 2.7%                   | 6.2%                 |
| 350   | 85.9%        | 5.8%                   | 14.1%                |
| 375   | 82.8%        | 6.9%                   | 17.2%                |
| 400   | 81.3%        | 9.4%                   | 18.8%                |

Table 3. Experimental result of big data fusion with only one $∅$ function neuron

| $k^2$ | Correct rate | False acceptance rate | False rejection rate |
|-------|--------------|------------------------|----------------------|
| 300   | 84.6%        | 5.8%                   | 15.4%                |
| 325   | 82.1%        | 3.2%                   | 17.9%                |
| 350   | 84.8%        | 2.5%                   | 15.2%                |
| 375   | 81.3%        | 0.6%                   | 18.7%                |
| 400   | 80.8%        | 2.2%                   | 19.2%                |
| 425   | 80.4%        | 2.5%                   | 19.6%                |
| 450   | 76.1%        | 8.1%                   | 23.9%                |
| 475   | 75.0%        | 10.7%                  | 25.0%                |
| 500   | 71.9%        | 18.1%                  | 28.1%                |

5. Conclusions

Our results have shown that cepstrum and $Δ$cepstrum big data feature introduced better performance than the pitch and $Δ$pitch with respect to the percentages of recognition. The experimental results have proved that using biomimetic pattern recognition with big data fusion for multiple microphone speaker recognition in second language provides an excellent recognition rate, that we can improve taking care of the following:

(1) When parameters $m = 3$, $p = 8$ and $k^2 = 275$ are selected, the best classification performance in the experiment using biomimetic pattern recognition with big data fusion for multiple microphone speaker recognition in second language is obtained.

(2) Increasing the number of function neuron is a very effective method for improving the recognition effect of big data fusion.

(3) We also maintain the distance constant value of $k$ in the range of 250–300 while increasing the value of $p$ to ensure false acceptance rate equal to zero.
There are still many challenges in speaker recognition for second language with big data fusion technology. For example, speaker recognition with big data fusion under noise environment still needs to be broken through. Signal processing is good at dealing with linear problems, machine learning methods such as biomimetic pattern recognition are good at dealing with nonlinear problems. But the actual problem must be the superposition of linear and non-linear, so it must be the fusion of the two in the future to better solve the problem of speaker recognition for second language with big data fusion under noise environment.

Acknowledgments
The author acknowledges the support of the South China Normal University and South China University of Technology.

References
[1] Sadic S and Gulmezoglu M B 2011 Common vector approach and its combination with GMM for text-independent speaker recognition Expert Systems with Applications vol 38 pp 11394-11400
[2] Muller F and Mertins A 2011 Contextual invariant-integration features for improved speaker-independent speech recognition Speech Communication vol 53 pp 830-841
[3] Lung S Y 2007 Efficient text independent speaker recognition with wavelet feature selection based multilayered neural network using supervised learning algorithm Pattern Recognition vol 40 pp 3616-3620
[4] Li J J and An D 2011 Embedded speaker recognition system design and implementation based on FPGA Procedia Engineering vol 29 pp 2633-2637
[5] Mary L and Yeegnanarayana B 2008 Extraction and representation of prosodic features for language and speaker recognition Speech Communication vol 50 pp 782-796
[6] Lung S Y 2010 Improved wavelet feature extraction using kernel analysis for text independent speaker recognition Digital Signal Processing vol 20 pp 1400-1407
[7] Alexander A and Botti F 2004 The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications Forensic Science International 146S pp 95-99
[8] Ruhm K H 2007 Sensor fusion and data fusion-mapping and reconstruction Measurement vol 4 pp 145-157
[9] Huang G S and Wang S W 2009 A data fusion scheme for building automation systems of building central chilling plants Automation in Construction vol 18 pp 302-309
[10] McDonald J L 2006 Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners Journal of Memory and Language vol 55 pp 381-401
[11] Mayberry R I and Lock E 2003 Age constraints on first versus second language acquisition: Evidence for linguistic plasticity and epigenesist Brain and Language vol 87 pp 369-384
[12] Serrano R and Llanes A 2011 Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe System vol 39 pp 133-143
[13] Verhoeven L and Steenge J 2011 Assessment of second language proficiency in bilingual children with specific language impairment: A clinical perspective Research in Developmental Disabilities vol 32 pp 1798-1807
[14] Cumming A and Riazi A 2000 Building models of adult second-language writing instruction Learning and Instruction vol 10 pp 55-71
[15] Remi A van C 2011 Developing second language sociopragmatic knowledge through concept-based instruction: A microgenetic case study Journal of Pragmatics vol 43 pp 3267-3283
[16] Ramachandrana R P and Farrelle K R 2002 Speaker recognition-general classifier approaches and data fusion methods Pattern Recognition vol 35 pp 2801-2821
[17] Sambhoos K and Bowman C 2011 A case study with design of experiments: Performance evaluation methodology for level 1 distributed data fusion processes Information Fusion vol 12 pp 93-104

[18] Aluja-Banet T and Daunis-i-Estadella J 2007 GRAFT a complete system for data fusion Computational Statistics & Data Analysis vol 52 pp 635-649

[19] Mechelen van I and Smilde A K 2010 A generic linked-mode decomposition model for data fusion Chemometrics and Intelligent Laboratory Systems vol 104 pp 83-94

[20] Gigli G and Bosse E 2007 An optimized architecture for classification combining data fusion and data-mining Information Fusion vol 8 pp 366-378

[21] Shasharina S and Eger R 2004 Data grid for fusion simulations and experiments Computer Physics Communications vol 164 pp 134-137

[22] Rochman D and Koning A J 2010 Exact nuclear data uncertainty propagation for fusion neutronics calculations Fusion Engineering and Design vol 85 pp 669-682

[23] Ashraf S and Brabyn L 2012 Image data fusion for the remote sensing of freshwater environments Applied Geography vol 32 pp 619-628

[24] Sabatier S B and Trester M R 2019 Measurement of the impact of identical twin voices on automatic speaker recognition Measurement vol 134 pp 385-389

[25] Jokinen E and Saiedi R 2019 Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task Computer Speech & Language vol 53 pp 1-11

[26] Xian G M and Zeng B Q 2012 Using combination of lifting wavelet and multiclass SVM based on global optimization class strategy for fault pattern identification Journal of Vibroengineering vol 14 pp 1616-1622

[27] Xian G M and Qu J P 2009 Computer-aided experiment of using real-time SALS image processing technique for visual characterization flow field of polymer melts Polymer Bulletin vol 62 pp 345-354

[28] Xian G M 2010 Mechanical failure classification for spherical roller bearing of hydraulic injection molding machine using DWT-SVM Expert Systems with Applications vol 10 pp 6742-6747

[29] Liu X Z and Zhu R B 2020 Intelligent data fusion algorithm based on hybrid delay-aware adaptive clustering in wireless sensor networks Future Generation Computer Systems vol 104 pp 1-14.

[30] Wang Y and Wang P 2018 Position synchronization for track geometry inspection data via big-data fusion and incremental learning Transportation Research Part C: Emerging Technologies vol 93 pp 544-565

[35] Yan Z and Liu J 2018 Big data fusion in Internet of Things Information Fusion vol 40 pp 32-33

[36] Liu J and Li T R 2020 Urban big data fusion based on deep learning: An overview Information Fusion vol 53 pp 123-133