International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0020
ISSN (Online): 2279-0039

International Journal of Engineering, Business and Enterprise Applications (IJEBEA)

www.iasir.net

Fuzzy Clustered Speaker Identification

Prof. Angel Mathew[1], Preethy Prince Thachil[2]
Assisant Professor[1], M.Tech. Student[2]
Department of Electronics and Communication,
Ilahia College of Engineering And Technology, Kerala, INDIA

Abstract*: Speaker identification is a biometric system. In speaker identification the task is to determine the unknown speaker identity by selecting one from the whole population. The key idea is that to use a fuzzy clustering to partition the original large population into subgroups. Clustering is done based on some features of the speeches. For a speaker under test, first conduct the fuzzy clustering based classification. Then apply MFCC + Neural network identification approach to the selected leaf node to determine the speaker identity.*

Keywords*: Fuzzy clustering, MFCC, Neural Networks*

I. Introduction

Identify a person from the sound of their voice is known as speaker identification [1]. There are two types of identification process. They are closed set identification and open set identification. In the closed set identification process set of registered speakers will be there, whereas in the open set the speaker will not be there in the database. In speaker identification, human speech from an individual is used to identify who that individual is. There are two different operational phases. They are training phase and testing phase. In training the speech from verified speaker need to be identified, is acquired to train the model for that speaker. This is carried out usually before the system is deployed. In testing the true operation of the system is carried out where the speech from an unknown speaker is compared against each of the trained speaker models. There are different techniques used for the identification process [2], [3]. In order to accomplish large population speakers and to identify the speakers in the correct group fuzzy clustering approach [4] has been used. Based on the features, the speakers can be separated into different group. At each level of the tree, we use a speech feature to do speaker clustering, i.e., a node (or a speaker group) splits into several child nodes (or speaker subgroups) at its lower level. In this process, speakers with similar speech feature are put into a same child node whereas speakers with dissimilar speech feature are put into different child nodes. Then, each child node contains a smaller population size than its parent node. Thus, at the bottom level, each speaker group at the leaf node has a very small population size and the population reduction is achieved. At the bottom level, we select one and only one speaker group at the leaf node that the speaker belongs to and apply MFCC + Neural Network to the selected speaker group for speaker identification. The advantage of our approach is that 1) we only apply MFCC + Neural Network to the speaker group at the leaf node with a very small population size instead of applying it to the original large population, 2) less computational complexity, and 3) more accurate.

II. Fuzzy Clustering

In large population speaker identification, it's feasible to use hierarchical decision tree for population reduction because human speech does contain many useful features that can be used to cluster speakers into groups. Speaker groups do exist that speakers sharing with a similar speech feature are in a same group whereas speakers having different speech features are from different groups. For example, speakers with different genders can be distinguished by using pitch feature [5]; based on different movement patterns of the vocal cords, different speaker groups could be obtained; Many emerging features which are independent from MFCC may indicate different speaker groups [6]. In summary, human speech has many different attributes and it's feasible to cluster speaker into groups by using various speech features. At each level of our hierarchical decision tree, we try to find different speaker groups by examining a certain attribute of speech. To achieve good performance, features used in our approach for clustering should meet the following requirements: 1) a good feature should be very capable of discriminating different groups of speakers; 2) features used at different levels of the tree should be independent from each other; 3) all features should be robust to additive noise.

A. Feature Description

All features we used fall into the category of vocal source feature. The source-filter model of speech production [7] tells us that speech is generated by a sound source (i.e., the vibration of vocal cords) going through a linear acoustic filter (i.e., the combination of the vocal tract and the lip). MFCC mainly represents the vocal tract

information. The vocal source is believed to be an independent component from the vocal tract and is able to provide some speaker-specific information. This is why we are interested in extracting vocal source features for speaker clustering. The first feature we derived is pitch or fundamental frequency. The rest of five features are all related to the vocal source excitation of voiced sounds. We extract them from the linear predictive (LP) residual signal [8].

B. Feature Extraction

In this section, we will specify how the six features are extracted from the speech signal.

1) Pitch Extraction: Pitch is calculated using cross correlation function. The samples are overlapped. By doing the overlapping samples, no information from the samples will be lost. It uses a 30msec segment and it chooses a segment at every 20msec so it overlaps at every 10mses. In the range of 60 Hz to 320 Hz [9] maximum autocorrelation is found out.

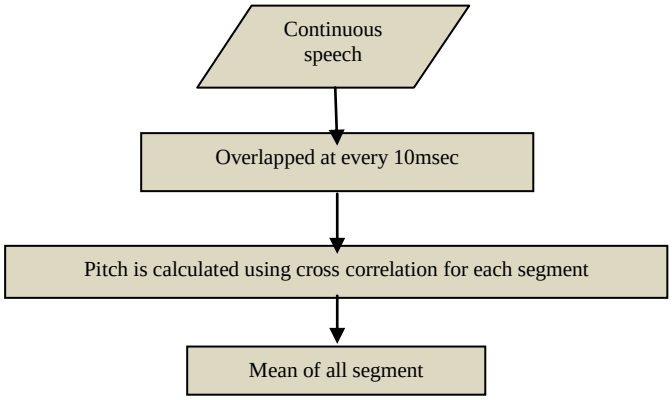

Figure 1: Pitch Feature Extraction

2) Vocal Source Features Extraction: The vocal source features are only derived from voiced speech frames. Given a continuous speech as the input, it is decomposed into short-time frames. The algorithm for vocal source feature extraction is as follows:

Step 1: Read the continuous speech.
Step 2: Speech is segmented into frames.
Step 3: Initialize frame index i = 1.
Step 4: Calculate energy, power and zero crossing.
Step 5: Pre- emphasis and windowing is done.
Step 6: Linear prediction analysis is done.
Step 7: Residual signal is calculated.
Step 8: Positive and negative pulse is detected.
Step 9: Vocal source features such as PAR (peak average ratio), skewness, and pulse width is calculated.
Step 10: If all frames finishes its processing it will terminate else it will jump to step 4.

3) Fuzzy clustering: The algorithm [10] applies to every feature we derived so that it does not specify the feature. We first do feature extraction and obtain the feature. We first calculate the mean and the standard deviation of the feature data. It is fed into Lloyd's algorithm [11] and a partition vector is obtained. The algorithm for fuzzy clustering is as follows:

Step 1: Input number of speeches.
Step 2: Input number of leaf nodes.
Step 3: Feature is extracted.
Step 4: Calculate mean and standard deviation of each feature.
Step 5: Apply Lloyd's algorithm.
Step 6: Initialize cluster index.
Step 7: Apply fuzzy.
Step 8: If cluster size is less than or equal to leaf node it will terminate else it will jump to step 7.

III. MFCC + Neural Network

After obtaining the features, we have to indentify the speaker. In order to identify the speaker MFCC [12] and neural network approach is applied. Since this approach is applied to the last node of the clustered output, the number of speakers will be reduced as compared to the parent node. So that it will function properly.

1) MFCC: MFCC (mel-frequency cepstrum coefficients) is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale'.The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above

1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale.

$F_{mel}=2595 \log_{10}(1+f/100)$

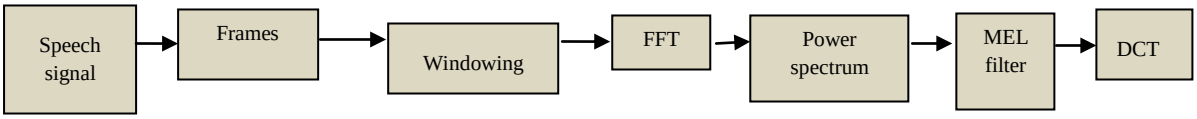

Figure 2: Block diagram of MFCC

The algorithm for MFCC is as follows:
Step 1: Convert time domain into frequency domain
Step 2: Convert speech signal into linear scale
Step 3: Mel frequency scale is linear till 1000Hz
Step 4: Logarithm scale after 1000Hz
Step 5: Power spectrum=$|fft|^2$
Step 6: $F_{mel} = 2595 \log_{10}(1+f/100)$

2) Neural Network: Neural network [13] is a machine that is designed to model the way in which brain performs a particular task or function of interest and network is usually implemented by using electronic components or is simulated on software in a computer.

To achieve good performance neural network employ a massive interconnection of simple computing cells referred to as neurons or processing units. It resembles the brain in two aspects 1) knowledge is acquired by network from its environment through a learning process. 2) Interneuron connection known as synaptic weights are used to acquire knowledge.

The procedure used to perform the learning process is called a learning algorithm, the function of which is to modify the synaptic weights of the network in an orderly fashion to attain a desired design objective.

The algorithm used in the Neural Network is backpropagation algorithm with adaptive learing Rate.the multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as back propagation algorithm.

The network consists of source nodes. The constitute the input layer, one or more hidden layer of computation nodes and an output layer of computation nodes. The input signal propagates through the network in a forward direction, on a layer by layer basis. These neural networks are commonly referred to as multilayer perceptrons.

Two kinds of signals are identified in the multilayer perceptron networks. A function signal is an input signal that comes in at the input end of the network, propagates forward through the network and emerges at the output end of the network as an output signal. An error signal originates at an output neuron of the network and propagates backward through the network.

Back propagation learning consists of two passes through different layers of the network, a forward pass and a backward pass. In the forward pass an input vector is applied to the input nodes of the network and its effect propagates through the network layer by layer. Finally a set of outputs is produced as the actual response of the network. During the forward pass the synaptic weights of the networks are not altered. In the backward pass, on the other hand, the synaptic weights are all adjusted in accordance with an error correction rule. Specifically the actual response of the network is subtracted from a desired response to produce an error signal. This error signal is then propagated backward through the network against the direction of synaptic connection, hence the name error back propagation. The synaptic weights are adjusted to make the actual response of the network move closed to the desired response in a statistical sense. The learning process performed with the algorithm is called back propagation learning.

The adaptive learning rate says that the human brain performs the formidable task of sorting a continuous flood of sensory information received from the environment. New memories are stored in such a fashion that existing ones are not forgotten or modified. The human brain remains plastic and stable.

IV. Conclusion

As the major technique for speaker identification, approach based on MFCC and Neural Network performs well. But as the population increase the performance degrades such as accuracy decreases and computational complexity increases. To improve the performance in the large population fuzzy clustering approach is applied. In this approach it partitions the large population of speakers into very small group and determining the speaker group at the leaf node to which a speaker under test belongs. To this leaf node MFCC and neural network approach is applied.

Reference

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

[2] D. Reynolds, "Large population speaker identification using clean and telephone speech," *Signal Processing Letters, IEEE*, vol. 2, no. 3, pp. 46–48, 1995.

[3] V. Apsingekar and P. De Leon, "Speaker model clustering for efficient speaker identification in large population applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 848–853, 2009.

[4] Yakun Hu, Dapeng Wu, and Antonio Nucci," Fuzzy-Clustering-Based Decision Tree Approach forLarge Population Speaker Identification" *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 762–774, 2013.

[5] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, 2011.

[6] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1097–1111, 2008

[7] X. Huang *et al.*, *Spoken language processing*. Prentice Hall PTR New Jersey, 2001.

[8] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[9] C. Wang, "Prosodic modeling for improved speech recognition and understanding," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[10] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition. i," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 29, no. 6, pp. 778–785, 1999.

[11] Ioannis Katsavounidis, C-C. Jay Kuo, and Zhen Zhang,"A New Initialization Technique for Generalized Lloyd Iteration"IEEE signal Processing Letters,vol. 1,No 10 ,pp.144146,1994

[12] B Milner,X Shao," Prediction of fundamental frequency and voicing from mel- frequency cepstral coefficients for unconstrained speech reconstruction" " *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no.14, pp. 24-33, 2007

[13] T.Poggio, F.Girosi,"Regularization Algorithm for Learning That Are Equivalent to Multilayer Networks"science magazine on vol. 247,, no 4945, pp. 978-982