International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 10, October 2017
www.ijarse.com
IJARSE
ISSN 2319 - 8354

AUTOMATED SPEAKER RECOGNITION METHODS: A CRITICAL REVIEW

Syed Akhtar Imam[1], Priyanka Bansal[2]

[1,2] *Deptt. of Electronics & Communication Engineering, Jamia Millia Islamia, New Delhi, India*

ABSTRACT

In this paper, an overview of state-of-the-art approaches for speaker recognition is presented. Due to the increased scalar of dialogue system applications the interest in that province has grown boomingly in certain years. Nevertheless, there are many open up shots in the field of automatic speaker recognition. The techniques, evaluations, and implementations of various proposed speaker recognition systems are reviewed with distinctive emphasis on issues prerogative to confirmation of speaker. We also describe here our direction for possible improvement to the automated speaker identification.

Keywords- Feature Extraction, Gaussian mixture models, machine learning algorithms, windowing, Mel Frequency Cepstral Coefficients, Vector Quantization

I. INTRODUCTION

The speech signal has an immoderate capability of carrying enlightenment. Speaker recognition, such as speaker identification and speaker verification is based on the fact that one‘s speech cogitates his/her unique characteristics. No two individuals sound tautological because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. Speech signal can be accomplished as a non-evasive biometric that can be collected with or without the person cognition or even transmitted over long distances via telephone. Unlike other types of identification, such as passwords or keys, a man's percept cannot be stolen, forgotten or lost. Speaker recognition permits for a secure method of authenticating speakers.

Speech is a legitimate, quick and convenient way of association with data processing systems. Nowadays, an increasing number of different systems incorporate speech interaction. Personal assistants, oration-supported navigation systems and telephone-based spoken dialogue systems (SDS), abbreviate our daily life. A necessary epithem of speaker recognition technology is forensics. Much of information is exchanged between two parties in telephone conversations, including between criminals, and in late donkey's years, there has been growing interest to integrate automatic speaker recognition to appendix auricular and semi-automatic analysis method.

While speech recognition and semantic interpretation have been the might fields of examination in this area, further knowledge sources about the interaction between humans and machines have been under examination. This research shall allow making spoken dialogue systems more intelligent in the future.

We have concentrated here on the performance of speaker identification systems, supported by foregone employment. Which algorithm should be used to obtain the appropriate level of accuracy? Which speech signal features should be extracted to generate gracious results?

II. SPEAKER IDENTIFICATION AND VERIFICATION SYSTEMS

The goal of a speaker identification procedure is to price out which person is currently speaking. To make the system perfect the sample speech signals of all people enduring access should be collected and narrative correspondingly. This training is usually done offline in advance to the actual deployment of the system. One may differentiate two dissimilar types of speaker recognition systems in terms of the speaker: open-set and closed-set systems. In closed-set identification, the model from the person to be recognized must be available in the existing speech database. Open-set identification allows speech input from a person absent in the database. In this case, the system should identify that the speaker is unascertained.

From the content instant of view, there are text-dependent and text-independent systems. Text-dependent systems approve the speaker identification utterances from a for-defined set only (e.g. passwords, numbers of trust cards or PINs). Text-independent systems do not have this measurement.

In contrast to speaker identification, in speaker verification systems mortal speech is used to verify whether this speaker is the assert person or not [5]. This problem has a chance of frequent issues with the proposition of speaker identification, especially if PINs or passwords are employed. Such systems are widely used in security applications to build a multi-level permission system. Thus at the maximum flat, all speaker recognition systems hold two capital modules: feature extraction and feature matching.

2.1. SELECTION OF FEATURES

Feature extraction is the process that descent a small amount of data from the voice signal that can posterior be used to depict each speaker. Speech signal embraces many features of which not all are serious for speaker discrimination. An ideal feature should have

i) bulky between-speaker variableness and least within speaker variableness

ii) be robust against noise and crookedness

iii) occur often and spontaneously in speech

iv) be unconstrained to measure from conversation signal

v) be difficult to impersonate/mimic

vi) not inclined by the speaker's health or long-term variations in voice.

vii) The numeral of features should also be relatively low

2.1.1. TECHNIQUES FOR FEATURE EXTRACTION

Following are the major techniques for feature extraction method:-

2.1.1.1. LPC

LPC (Linear Predictive coding) analyzes the speech token by estimating the formants, removing their execution from the speech signal, and estimating the earnestness and frequency of the relics buzz. The process of removing the formants is called inverse filtering, and the relic signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is convoked a linear predictor and hence it is called as linear predictive coding. The coefficients of the difference equation (the prediction coefficients) characterize the formants.

2.1.1.2. PLP

PLP (Perceptual linear prediction) binds LPC analysis with psychophysics knowledge of the human auditory system. With respect to LPC, PLP analysis ply three transformations to the speech signal to simulate the perceptual properties of the human hearing. The three psychophysics supported transformations are accurate band analysis, commensurate-loudness pre-emphasis and intensity loudness conversion. Critical band analysis pretended the non-uniform frequency resolution of the auditory system: the earthling hear has a higher crowd separation at moderate frequencies than it does at high frequencies. This is achieved by mapping the frequency scale onto the Bark scale. Equal-loudness ante-emphasis simulates the non-equal sensitivity of hearing at different frequencies. Finally, intensity loudness conversion simulates the non-linear relationship between the amplitude of a sound and its discern loudness using a cube root amplitude compression.

2.1.1.3. MFCC

MFCC (Mel Frequency Cepstral Coefficient) is based on the earthborn peripheric auditory system. The human perception of the frequency filling of sounds for speech signals does not follow a linear scale. Thus for each temper with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the ‘Mel Scale'.The mel frequency scale is a linear frequency duration below 1000 Hz and logarithmic course above 1kHz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing outset, is defined as 1000 Mels.

2.2. FEATURE MATCHING

Feature matching involves the genuine procedure to identify the unascertained speaker by comparing extracted features from his/her voice input with the ones from a set of given speakers.

All speaker recognition systems have to go through two distinguishes phases. The first one is the enrollment sessions or training phase while the assistance one is the operation sessions or testing phase.

In the training phase, each registered speaker has to provide samples of their harangue so that the system can build or train a reference model for that speaker. In plight of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training samples.

During the proof (operational) phase, similar feature vectors are extracted from the trial utterance, and the grade of their match with the reference is obtained using some twinned technique. The level of match is used to arrive at the decision. Some of the techniques for feature matching are discussed as follows:

2.2.1.1. DTW

Dynamic time warping is an algorithmic procedure for measuring the similarity between two sequences which may transmute in delay or speed. According to the DTW techniques proposed by Sadaoki Furui , the training data are used as a commencing platter, and the testing data is time-aligned by DTW. DTW is a method that assigns a computer to find an optimal match between two inclined successions. The average of the two patterns is then taken to yield a novel patter to which a third utterance is era aligned. This process is repeated until all the training utterances have confederated into a single patter. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment process is often used for time series classification.

2.2.1.2. VQ

Vector Quantization method works by graduating a large set of peculiarity (vectors) into groups having approximately the same number of points closest to them. Each assemblage is represented by its centroid point, as in k-means and some other clustering algorithms. The compactness matching the property of vector quantization is powerful, particularly for identifying the density of large and high-dimensioned data. Since data peculiarity is represented by the index of their closest centroid, commonly occurring data have low fallacy and rare data high error.

2.2.1.3. GMM

GAUSSIAN MIXTURE MODELLING assumes vector space to be partitioned into specific components depending on clustering of feature vectors and devise the feature vector distribution in each compositional to be Gaussian. When feature vectors are displayed in d-dimensional feature space after clustering, they some-how counter-fit to the Gaussian distribution. It means each corresponding cluster can be viewed as a Gaussian probability distribution and features belonging to the clusters can be best represented by their probability values. Individual Gaussian classes are interpreted to represents set of acoustic classes. These acoustic classes represent vocal tract information.

2.2.1.4. SVM

Support Vector Machine is a binary star classification system that finds the optimal linear decision surface supported on the concept of structural risk minimization. The decision superficies is a weighted combination of elements of a training set. These elements are invoked support vectors, which characterize the boundary between the two classes. For the intend to characterize the boundary between the two classes, we indigence maximizing the margin. During speaker recognition process, classifying the feature which is descended from the transformation of feature extraction directly will not immediately work when using SVM. It is because SVM only can process fixed-length input, whereas speech signals are non-stationary. Therefore, we need to categorize the feature and desquamation them.

2.2.1.5. HMM

HIDDEN MARKOV METHOD is an orthodox probabilistic model for the sequential or transitory data and it depends upon the fundamental event of the real world: "Future is independent of the past but driven by the present." The HMM is a doubly embedded random process, where final output of the system at a particular instant of time depends upon the state of the system and the output generated by that state. There are two types of HMMs: DHHMs and CDHMMs. These are distinguished by the semblance of data that they operate upon. DHMMs act on quantized data or symbols; on the other hand, CDHMMs act on continuous data, and their issue matrices are the distribution province.

2.2.1.6. NEURAL LOGIC

The characteristic essence MFCC from the training phase is used in a neural network. The neural network is able to represent an intricate pattern that forms the non-linear hypothesis. The feedforward propagation algorithm is implemented to reckon all the activations throughout the network, including the output value of the hypothesis using the initial random weights for prediction. Then backpropagation algorithm for learning the neural network parameter was applied to compute an "error term" that measures how much that node was "responsible" for any errors in the output.

2.2.1.7. GFM

Generalized Fuzzy model is used to solve the problem of ample population speaker identification under noisy conditions. The keynote idea of this approach is to

1) use a decision tree to hierarchically partition the whole population into blocks of small size, and determine which speaker group at the leaf node a speaker under test belongs to, and

2) apply MFCC+GMM to the chosen speaker assemblage for speaker identification. The advantage of this approach is that features that are independent of MFCC are used to partition speakers into groups and only apply MFCC+ GMM to speaker groups at the leaf level. The key defiance in our design is how to achieve a low error probability of decision-tree-based classification. To address this, fuzzy clustering is adopted in constructing the tree for population partitioning, i.e., at each impartial, a speaker may belong to the manifold assemblage. Such redundancy increases the probability of classifying a speaker under test into a correct group/protuberance on the tree.

III. CONCLUSION

In this paper, we made an overview of existing approaches designed for the problem of speaker identification and verification. This overview has shown some open issues in the speaker identification problem. Among them, there is an appropriate choice of the feature set and the modeling algorithm. The low-level features such as cepstral features work well in ideal conditions, but their performance is corrupted in real time situations. Use of high-level information can add complementary knowledge to improve the performance of recognition system. In practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are necessary. There are number of research problems that can be taken up, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores. For this it is important to explore stable features that remain insensitive to variation of speaker's voice over time and are robust against variation in voice quality due to physical states or disguises. The problem of distortion in the channels and background noise also requires being resolved with better techniques.

REFERENCES

[1] R. Togneri, D. Pullella, (2011). “An overview of speaker identification: Accuracy and robustness issues”, Circuits and Systems Magazine, IEEE, 11(2), pp. 23-61, 2011.

[2] Anjali Bala, Abhijeet Kumar, Nidhika Birla -*Voice Command Recgnition using system based on MFCC and DTW”*, International journal of engineering science and technology vol.2 (12), 2010.

[3] P. Geunter, M. Denecke, U. Meier, M. Westphal and A. Waibel, “Conversational speech systems for on-board car navigation and assistance”, in Proceedings ICSLP, 1998.

[4] K. T. Mengistu, A. Wendemuth, “Telephone-Based Spoken Dialog System Using HTK-based Speech Recognizer and VoiceXML”, Fortschritte der Akustik, 33(2), 625, 2007.

[5] R. Togneri, D. Pullella, (2011). "An overview of speaker identification: Accuracy and robustness issues", Circuits and Systems Magazine, IEEE, 11(2), pp. 23-61, 2011.

[6] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72-83, 1995.

[7] J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification", in Proc. ESCA Workshop Automat. Speaker Recognition, Identification, Verification, Apr. 1994, pp. 39-42.

[8] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Esply-Wilson and S. Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition", IEEE Automatic Speech Recognition and Understanding Workshop, 2011.

[9] D. Garcia-Romero, C. Espy-Wilson, "Joint factor analysis for speaker recognition reinterpreted as signal coding using overcomplete dictionaries", Proc. Odyssey Speaker and Language Recognition Workshop, 2010.

[10] D.Garcia-Romero, C.Espy-Wilson,"Analysis of I-vector Length Normalization in Speaker Recognition Systems", Interspeech, 2011.

[11]C. Y. Espy-Wilson, S. Manocha and S. Vishnubhotla, "A New Set of Features for Text-Independent Speaker Identification", ICSLP Interspeech, 2006.

[12] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", IEEE International Conference on Multimedia & Expo, 2005.

[13] D. Chow and W. H. Abdulla, "Robust Speaker Identification Based on Perceptual Log Area Ratio and Gaussian Mixture Models", ICSLP Interspeech, 2004.

[14] Priyanka Bansal and Roma Bharti, " Real Time Speaker recognition system using MFCC and vector quantization technique" , International Journal of Computer Applications, Vol. 117, Issue 1, pp. 25 - 31,2015.

[15] Priyanka Bansal and Syed Akhtar Imam, "Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy", International Conference on Soft Computing Techniques and Implementations, October 2015.