Hybrid Approach of Feature Extraction and Vector Quantization in Speech Recognition

Sarthak Manchanda and Divya Gupta

Abstract This paper examines the speech recognition process. Speech recognition has two phases: the front end which comprises of preprocessing of the speech waveform and the back end which comprises of feature extraction and feature matching. In this review, we discuss some feature extraction techniques like MFCC, LPC, LPCC, PLP, and RASTA-PLP. As these techniques have some demerits stated in the paper, we discuss a hybrid approach of feature extraction with some combinations of the above techniques. Feature matching helps in the recognition part of speech recognition. It is done by comparing the feature vectors of the current user to the feature vectors stored in the database. It can be optimized by vector quantization (VQ) in order to speed up the recognition process.

Keywords Feature extraction · Feature matching · Front end
Back end · Vector quantization

1 Introduction

Speech recognition is known as the ability of a machine to detect and identify the words or phrases uttered by the human being and to detect and convert this speech signal into machine read format. This concept of Artificial Intelligence is quite common these days and many advancements are being done toward it [1].

S. Manchanda (✉) · D. Gupta
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India
e-mail: sarthak120895@gmail.com

D. Gupta
e-mail: dgupta1@amity.edu

V. Bhateja et al. (eds.), *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, Advances in Intelligent Systems and Computing 712, https://doi.org/10.1007/978-981-10-8228-3_59

Speech recognition has front end and back end analysis.

Front End:

1. Preprocessing: It is used to increase efficiency for further processes. It helps in making the system more robust. It also helps in generating parametric representation of the speech signal.
2. Framing and Windowing: Most part of a speech signal processing depends on short time analysis done with the help of framing. The signal gets blocked into frames of samples N with a duration ranging between 10 and 30 ms.

Back End:

1. Feature Extraction: It is useful to extract useful information and remove the unwanted and redundant information. Certain features are extracted using this process which helps in speech recognizing. The goal of this technique is to compute the sequence savings of feature vectors which represent the computing signals. Different types of feature extraction techniques are: MFCC, LPCC, LPC, PLP, and RATSA-PLP [2].
2. Feature Matching: It is known as the process in which identification of the speaker is done by comparing the extracted features to the features which are stored in a database. It is basically done by first storing the input signal features in the database and then comparing the stored values to the input values of the unknown speaker. It gives the result as matched or not matched [2].

2 Type of Speech Uttered by Human Beings

There are many types of speech uttered by the human beings which differ in parameters [3]. These types are:

1. Isolated Speech:
   It requires single word at a time. This is one of the best speech types in which there is very less noise on both sides of the window.
2. Connected Speech:
   This type has minimum pause between the words uttered by the human.
3. Continuous Speech:
   This kind of speech has no pause between the words. It is basically what computer dictates and what human beings say the most.
4. Spontaneous Speech:
   This type of speech can be thought of as natural and without trying it before. An ASR system can handle spontaneous speech when it has every word with its meaning in the database.

3 Feature Extraction Techniques

It is the most important part in any ASR system. The meaning of extracting feature is to extract the useful information and remove unwanted information (Figs. 1 and 2).

1. Mel Frequency Cepstral Coefficients(MFCC):
   It is the most common and most popular feature extraction technique used for an ASR. Frequency bands in MFCC are placed in a logarithmical order so that it can approximate the human being system closely than any of the other system (Fig. 3).
2. Linear Prediction Coding(LPC):
   It is a good signal feature extraction method for linear prediction in speech recognition processes (Fig. 4). The basic idea behind LPC is that the speech signal can be approximated as a linear combination of past speech samples stored in the database [4].
3. Linear Prediction Coding Coefficient(LPCC):

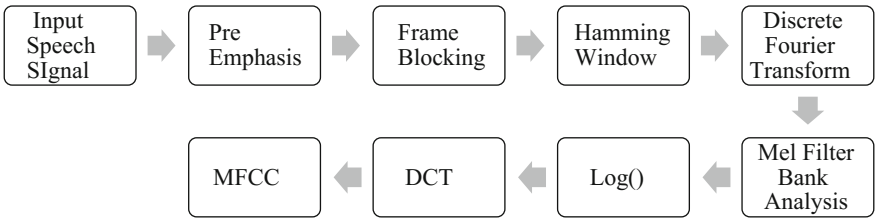

Fig. 1 Mel frequency cepstral coefficients(MFCC)

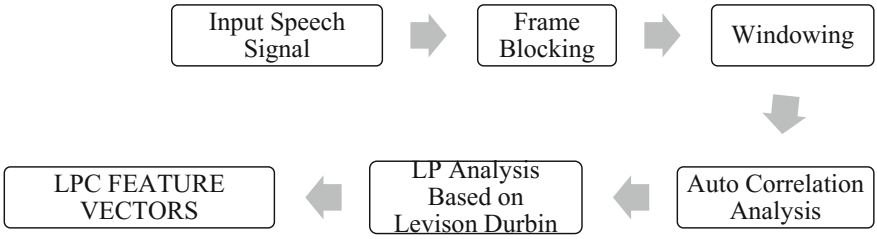

Fig. 2 Linear prediction coding(LPC)

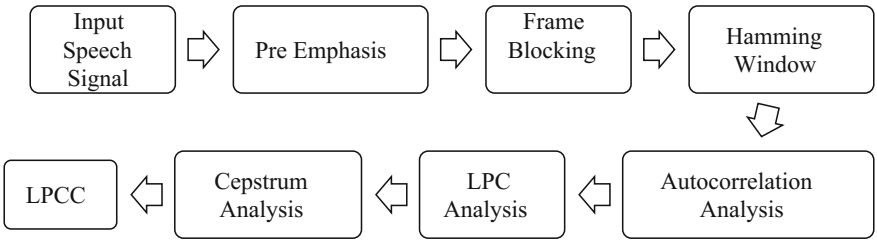

Fig. 3 Linear prediction coding coefficient(LPCC)

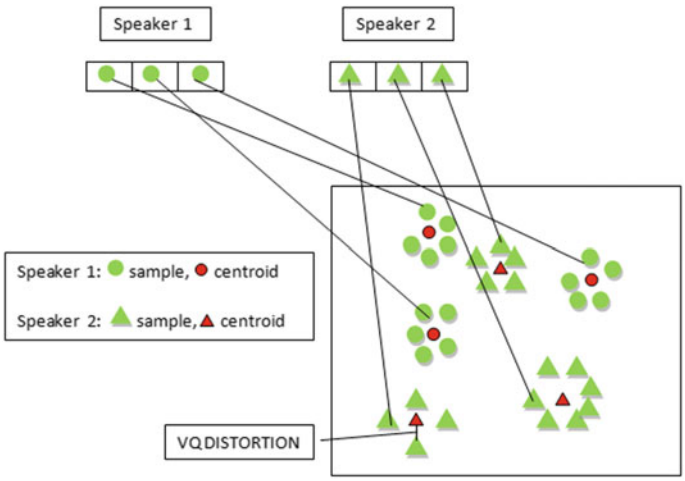

Fig. 4 Vector quantization

This technique works at a low bit by demonstrating the speech signal by a finite number of measure of the signals. It represents a mimic of speech after computing a smooth version of cepstral coefficient using autocorrelation method. In LPCC feature extraction, cepstrum analysis is done on the LPC analysis as seen below [4] (Fig. 3).

Other feature extraction techniques are PLP, RASTA-PLP etc.

4. Hybrid and Robust Technique:
   In order to obtain a new and more effective feature extraction technique, certain hybrid algorithms make a distinction from the previous feature techniques. The previous techniques have some drawbacks which do not make the recognition process accurate and robust. The drawbacks are given in Table 1 [4].

Table 1 Demerits of feature extraction techniques

MFCC	LPC	LPCC	PLP	RASTA PLP
1. It is very much sensitive to mismatch of the channels between the training and testing phases 2. Performance of MFCC is effected by total number of filters used 3. Does not give accurate result in noisy environment	1. It is used in only a linear combination of speech signals	1. It is highly sensitive to Quantization Noise	1. It gives a higher error rate as compared to other techniques	1. It is better than PLP as it does the filtration but still error rate is still found in it

To overcome these drawbacks, a hybrid technique is used. Combination of previous features is taken to make the extraction more robust.

There is a generation of 13 parameter coefficients of these techniques

(MFCC, LPC, LPCC, PLP, and RASTA-PLP). In four experiments, different combinations of these techniques were used to give "39 coefficient parameters" in a single vector:

(a) 13(MFCC) + 13(LPCC) + 13(PLP)
(b) 13(MFCC) + 13(LPCC) + 13(RASTA PLP)
(c) 13(MFCC) + 13(PLP) + 13(RASTA PLP)
(d) 13(LPCC) + 13(PLP) + 13(RASTA-PLP).

4 Feature Matching

While the above feature extraction techniques were used in the front end to extract certain important characteristics, **vector quantization** is used in the back end to generate a correct decision while maintaining a codebook [5, 6].

It is a procedure to identify an unknown speaker by comparing the extracted features from the above approach with the data of the known speaker stored in the database [2].

Vector Quantization(VQ):

VQ is nothing more than "rounding off to the nearest integer." It is also known as the centroid model. It was introduced in the year 1980 and it has its roots from data compression. It has many advantages:

- speed up the recognition process
- for lightweight practical use
- ease of implementation.

Apart from these advantages, it can lead to loss of potential data.

In VQ, the signal vectors from a very large vector scale can be mapped on a finite region of space (Fig. 4). Every region of the space is known as "cluster" and is shown by a center called "codeword". The group of these code words is known as a "codebook" [7].

In the figure are two different speakers on a two-dimensional space. Triangle refers to the vector from "speaker 1" while the circle refers to the vector from "speaker 2". There are two phases in VQ. In the first phase which is also known as the training phase, a codebook is generated which is speaker specific for every single known speaker taking the training vectors in the database. The result of these code words is represented as the centroid of every triangle or circle as shown in the figure by Green Circles for "speaker 1" and Green Triangles for "speaker 2".

In the second phase which is also known as the recognition phase, speech input signal is "vector quantized" using a trained codebook and the total VQ-distortion is computed. The speaker having the codebook with the least distortion is discovered.

5 Conclusion and Future Work

The main objective of this research is to analyze why the hybrid approach is better to use for feature extraction and how can vector quantization help in fast and accurate feature matching. This paper gives the detail about various feature extraction techniques and how hybrid approach is better than these techniques. It shows some of the advantages of the hybrid approach in the table.

On the basis of the review, it is analyzed that the combination of techniques can help in achieving good results in a noisy environment also as variety of filters is used in hybrid approach. This hybrid approach is used for a flexible kind of environment.

For feature matching, vector quantization is used in which optimization is done which speeds up the recognition process. The future work is to develop an ASR System with high accuracy.

References

1. Karpagavalli, S., and Chandra, E "A Review on Automatic Speech Recognition Architecture and Approaches" International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 9, No. 4, (2016) pp. 393–404.
2. Geeta Nijhawan 1, Dr. M.K Soni 2 "Speaker Recognition Using MFCC and Vector Quantisation" ACEEE Int. J. on Recent Trends in Engineering and Technology, Vol. 11, No. 1, July 2014.

3. Akansha Madan, Divya Gupta "Speech Feature Extraction and Classification: A Comparative Review" International Journal of Computer Applications (0975–8887) Volume 90 – No 9, March 2014.
4. Veton Z. Këpuska, Hussien A. Elharati "Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions" Journal of Computer and Communications, 2015, 3, 1–9 Published Online June 2015.
5. Hemlata Eknath Kamale, Dr.R. S. Kawitkar "Vector Quantization Approach for Speaker Recognition" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 3, Special Issue, March-April 2013, An ISO 9001: 2008 Certified Journal.
6. Preeti Saini, Parneet Kaur "Automatic Speech Recognition: A Review" "FLEXIBLE FEATURE EXTRACTION AND HMM DESIGN FOR A HYBRID DISTRIBUTED SPEECH RECOGNITION SYSTEM IN NOISY ENVIRONMENTS" International Journal of Engineering Trends and Technology-Volume 4 Issue 2– 2013.
7. Dipmoy Gupta, Radha Mounima C. Navya Manjunath, Manoj PB "Isolated Word Speech Recognition Using Vector Quantization (VQ)" International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 5, May 2012 ISSN: 2277 128X.
8. Veton Z. Këpuska, Hussien A Elharati "Performance Evaluation of Conventional and Hybrid Feature Extractions Using Multivariate HMM Classifier" Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 5, Issue 4, (Part -1) April 2015, pp. 96–101.
9. Pratik K. Kurzekar 1, Ratnadeep R. Deshmukh 2, Vishal B. Waghmare 2, Pukhraj P. Shrishrimal 2 "A Comparative Study of Feature Extraction Techniques for Speech Recognition System" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 12, December 2014.
10. Pawan Kumar, Astik Biswas, A. N. Mishra and Mahesh Chandra "Spoken Language Identification Using Hybrid Feature Extraction Methods" JOURNAL OF TELECOMMUNICATIONS, VOLUME 1, ISSUE 2, MARCH 2010.
11. H. B. Kekre, Tanuja K. Sarode "Speech Data Compression using Vector Quantization" International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 2, No: 3, 2008.