DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5305

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

SEARCHING OF SPEECH QUERIES IN AN AUDIO DATABASE USING MEL-FREQUENCY CEPSTRAL COEFFICIENTS AND GAUSSIAN POSTERIORGRAMS BASED FEATURES

B. N. Veerappa
Department of Studies in CSE
UBDT College of Engineering
Davangare – 577004, Karnataka, India

Sudarshana Reddy H. R.
Department of Studies in E&E Engineering
UBDT College of Engineering
Davangare – 577004, Karnataka, India

Abstract: In this paper, we propose to use Mel-frequency cepstral coefficients (MFCC) and Gaussian Posteriorgrams (GPG) features to develop an Audio information retrieval (AIR) system. Using this AIR system we search speech queries in an audio database. In our proposed approach, we develop three independent systems based upon MFCC and GPG features to obtain the time stamp evidence for the location of speech queries in the reference utterances. Further, the Majority voting decision logic is used to arrive at a conclusion to locate (time stamp) the query word in the reference utterances. We use TIMIT database to conduct our proposed studies.

Keywords: MFCC; Gaussian Posteriorgrams; Dynamic time warping; Audio information retrieval

I. INTRODUCTION

The task of Audio Information Retrieval (AIR) is to find a speech query within an audio database. Spoken audio data is available from various sources. For example,

(a) Recorded speeches in parliament and public speeches
(b) Recordings from radio and television stations such as British Broadcasting Corporation (BBC) archives [1], and All India Radio, Door Darshan channels etc.
(c) Recordings from talks at conferences (Technology, Entertainment, Design (TED) talks}) [2]
(d) University course wares such as recordings from lectures (MIT Open Courseware (MIT OCW) [3], National Programming on Technology Enhanced Learning (NPTEL) [4]
(e) Recordings done in court hearings.
(f) President and Prime Minister address to the nation

There is an alarming increase in the amount of audio data and hence there is a need to develop automatic and robust approach to search the required audio information within the given audio database.

One of the straight forward approaches is to listen to the entire speech utterance to verify whether the keyword to be searched is present or not. In order to achieve this, one need to manually transcribe the entire speech utterance and then make use of text-based search methods. The drawbacks of the manual transcription of speech utterances are tedious, time consuming and highly expensive.

Th organization of this paper is as follows: The review of approaches for AIR is described in Section II. The database used in the studies is described in Section III. MFCC and Gaussian posteriorgrams based representation of speech is provided in Section IV. Experimental details on the studies on AIR is described in Section V. Hypothesizing query words in an utterance by using MFCC and Gaussian Posterior gram based systems is explained in Section VI. Analysis of results is discussed in Section VII. Final section provides the summary and conclusions from the current studies on AIR.

II. EXISTING APPROACHES FOR AUDIO INFORMATION RETRIEVAL

A. *ASR based Audio Information Retrieval*

A conventional approach to audio information retrieval is to convert speech utterance into a sequence of text symbols using an Automatic Speech Recognition (ASR) system. Then carryout the text based search. But ASR-based approach requires the large amount of labeled data for training the models. The block diagram of AIR system using ASR is shown in Fig. 1.

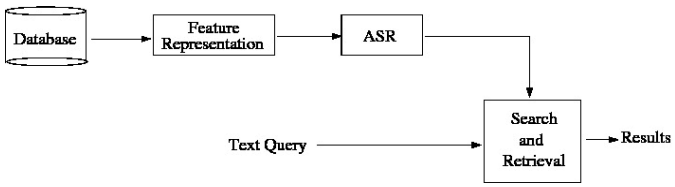

Figure 1. ASR based Spoken Term Detection (STD) System.

AIR using ASR based approach is not scalable for many languages where there is no availability of labeled data or the proper resources to build an ASR. Thus, there is a need to automate searching of speech utterance.

B. *Speech based Audio Information Retrieval*

To overcome the drawback of ASR based search techniques, no prior knowledge about the speech utterance language is assumed. In this paper, we propose an AIR approach based on speech data [5]. The block diagram of proposed approach for the audio information retrieval system based on speech query is shown in Fig. 2.

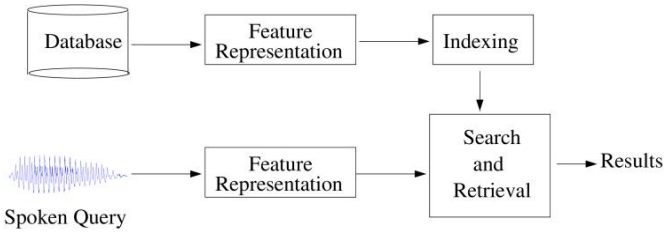

Figure 2. Audio Information Retrieval system based on speech query [5].

C. Commonly used Audio Features

In a digital system, the speech signal is represented by discrete amplitude values as a function of discrete time intervals. From a statistical point of view, these discrete speech samples are not directly used by many machine learning approaches. The information lies in the sequence of samples rather than individual samples themselves. Therefore it is necessary to extract the features from the speech signal which are best suited for a particular task.

The following are the commonly used acoustic features for processing of speech signals: (a) Linear prediction cepstral coefficients (LPCC), (b) Mel-frequency cepstral coefficients (MFCC), and (c) Perceptual linear prediction cepstral coefficients (PLP) [5, 6, 7, 8, 9,10].

The main limitations of LPCC and MFCC are that they are susceptible to speaker and environmental conditions. In this paper, we explore dynamic features of MFCCs and Gaussian Posteriorgrams. The dynamic features such as first derivatives of MFCCs called deltas and second derivatives of MFCCs called accelaraton contains dynamics of vocal tract features. Further Gaussian posteriorgrams smoothen the static and dynamic features. These features are explored for the proposed task of audio information retrieval.

D. Search Techniques for Acoustic Similarity

Normally euclidean distance measure is used to find the similarity between the two speech patterns. In our studies, we have used Dynamic time warping (DTW) algorithm to match the query word at an appropriate location in the reference utterance.

Dynamic time warping (DTW) algorithm is used for alignment of two time series data of unequal lengths. Initially the DTW is used for template based speech recognition [21]. It tries to align two sequences of feature vectors by warping the time axis iteratively until an optimal match between the two sequences is obtained. Dynamic Time Warping for time alignment and normalization is to compensate for variability in speaking rate in reference-based speech systems [11].

III. DESCRIPTION OF DATA BASE

For the studies of Audio information retrieval, we require large amount of labeled data. The data base should have wave files, prompt sentences, word level transcription for time stamp. The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech processing systems [12][13] [14]. Sampling frequency of each of the speech utterance is 16,000 Hz and number of bits per sample is 16.

The, TIMIT data base contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers distributed over eight different dialectical regions of the United States. Out of 630 speakers, 462 speakers are used for training (reference) data. In a similar way, 168 speakers are used in test data. In all, there are 4620 utterances in training data and 1680 utterances in test data. This database is used in our AIR studies.

A. Selection of Query Words (Keywords)

To measure the performance of Audio information retrieval system, choice of query words is very important. We have examined all the words occurring in 2343 prompt sentences of TIMIT database. The query words need to be selected in such a manner that, they should not be part of other words. Based on this factor, we have arrived at the following 5 query words.

(a) water
(b) country
(c) ocean
(d) mother
(e) social

The frequency of occurrence of these query words in the reference utterances (training data) are given in Table I.

Table I. Details of frequency of occurrence of query words

Sl. No.	Query Word	Frequency of occurrence
1	water	18
2	country	2
3	ocean	15
4	mother	4
5	social	13
	Total	52

IV. SPEECH ANALYSIS BASED ON MFCC AND GAUSSIAN POSTERIORGRAMS REPRESENTATION

The first step in any automatic speech processing system is to extract features. These features mainly identify the components of the speech signal which represent mainly the linguistic information. In this regard, we have explored the following two approaches for the representation of speech signal for the task of AIR

(a) Mel-frequency cepstral coefficients (MFCC)
(b) Gaussian posteriorgrams

A. Mel-frequency cepstral coefficients

The phonemes generated by a human are filtered by the shape of the vocal tract which include tongue, teeth etc. This shape determines the type of the phoneme sound to be produced. If we can determine the shape of the vocal tract accurately, then it is possible to identify the type of phoneme that is produced by the corresponding shape of the vocal tract. The shape of the vocal tract system is manifasted in the envelope of the magnitude spectrum of the short time Fourier analysis. The MFCC features accurately represents the envelope.

Generally, the MFCC features are used in many speech processing tasks [15]. But, we explore them for the task of AIR. Following are the main steps used in the extraction of MFCC features [16].

1. Consider a short segment of speech frame.
2. Calculate the Discrete Fourier Transform (DFT).
3. Find the magnitude spectrum of the DFT.
4. Pass the magnitude spectrum through the mel filterbanks (typically 26).
5. The filterbank energies are obtained by summing the energy in each of the filter.
6. The log filterbank energies are obtained by applying the logarithm to all filterbank energies.
7. To reduce the dimension of the log filterbank energies take the Discrete Cosine Transform (DCT) of the log filterbank energies.
8. Select only the first 13 DCT coefficients.
9. These 13 coefficients are known as Mel-Frequency Cepstral Coefficients.

The above steps are depicted in the Fig. 3.

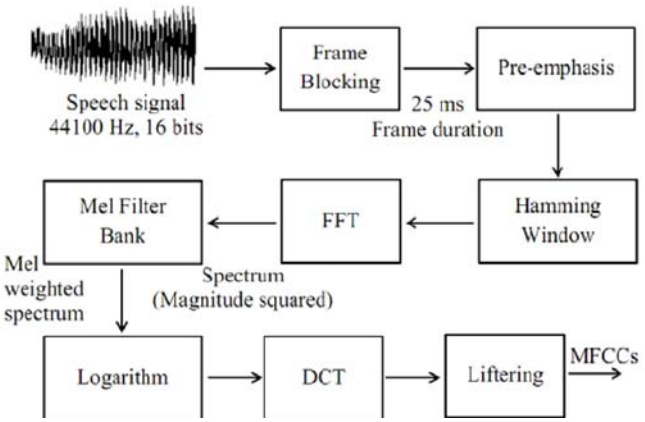

Figure 3. Block diagram for obtaining MFCC features.

B. MFCC, Delta and Accelaration Coefficients

The MFCC (static) feature vector describes only the power spectral envelope of a single frame. But, the speech is produced from vocal tract system which is dynamic in nature. Thus speech is also having information in the dynamics. That is, in the trajectories of the MFCC coefficients over a period of time. Thus by calculating the MFCC trajectories (first and second derivatives) and then appending them to the original feature vectors represents the vocal tract shape better. The dynamic features derived from static features also contain dynamics of speech. The performance of a speech processing systems can be greatly enhanced by adding time derivatives to the basic static parameters.

We have derived dynamic features such as 13-dimensional delta (first order derivative) and 13-dimensional (acceleration) features. Thus each speech frame is represented by 39-dimensional (13-MFCC + 13-Delta + 13-Acceleration) feature vector.

C. Gaussian Posteriorgrams based Features

Just as phonetic posteriorgram described in [20], a Gaussian posteriorgram is a probability vector representing the posterior probabilities of a set of Gaussian components for a speech frame. Gaussian posterior features have been widely used in speech recognition systems [17], [18], [19]. Formally, if we denote a speech utterance with n frames as S = (f1, f2 , , fn), where n is the number of frames in the speech utterance S.

Then the Gaussian posteriorgram (GP) is defined by:

$$GP(S) = (q1 , q2 , .., qn)$$

Each qi vector is calculated by:

$$qi = (P(GPC1|fi), P(GPC2|fi), ... , P(GPCm|fi))$$

where GPCi represents i-th Gaussian component of a Gaussian Mixture Model (GMM) and m denotes the number of Gaussian components. Typical value of m is 128. Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering. They are also used intensively for the estimation of density estimation.

TheGaussian posteriorgrams are extracted as follows:

Step1: In the first step, a GMM is trained by using all the feature vectors of training utterances. Then use this GMM to produce a raw Gaussian posteriorgram vector for each of the frames of training utterances.

Step 2: In the second step, for each of the posteriorgram vector smoothing technique is applied.

V. EXPERIMENTAL STUDIES

We have considered the following three types features for the studies of AIR:

(a) 39-dimensional MFCC-Delta-Accelaration features derived from 13-dimensional MFCC (static) features

(b) 128-dimensional Gaussian Posteriorgram features derived from 13-dimensional MFCC (static) features

(c) 128-dimensional Gaussian Posteriorgram features derived from 39-dimensional MFCC (static) features

We have extracted the above three types of features for all the reference utterances of TIMIT database. Further, we have also extracted the above three types of features for the chosen query words. For short time analysis, we have considered a frame of size 25 ms and frame shift 15 ms at a time.

A. Evidence of location (time stamp) of query words in reference utterances with respect to 39-dimensional MFCC-Delta-Accelaration based DTW system (Only_M39)

Dynamic time warping approach is applied to derive the warping path which provides the best alignment of the query word and all the training utterances in order to obtain the location (time stamp) of query in a reference utterance with respect to 39-dimensional MFCC-Delta-Accelaration features.

B. Evidence of location (time stamp) of query words in reference utterances with respect to 128-dimensional Gaussian Posteriorgrams derived from 13-dimensional MFCC based DTW system (GPG128_M13)

In a similar way, Dynamic time warping approach is applied to derive the warping path which provides the best alignment of the query word and all the training utterances in order to obtain the location (time stamp) of query in a reference utterance with respect to 128-dimensional Gaussian Posteriorgrams derived from 13-dimensional MFCC features.

C. Evidence of location (time stamp) of query words in reference utterances with respect to 128-dimensional Gaussian Posteriorgrams derived from 39-dimensional MFCC-Delta-Accelaration based DTW system (GPG128_M39)

Finally, Dynamic time warping approach is applied to derive the warping path which provides the best alignment of the query word and all the training utterances in order to obtain the location (time stamp) of query in a reference utterance with respect to 128-dimensional Gaussian Posteriorgrams derived from 39-dimensional MFCC-Delta-Accelaration based features.

VI. HYPOTHESIZING QUERY WORDS IN AN UTTERANCE

The presence of query word in a reference utterance is hypothesized from the time stamp information obtained from Only_M39, GPG128_M13, and GPG128_M13 based DTW systems. The block diagram of the proposed approach for hypothesizing query word in an utterance is shown in Fig. 4.

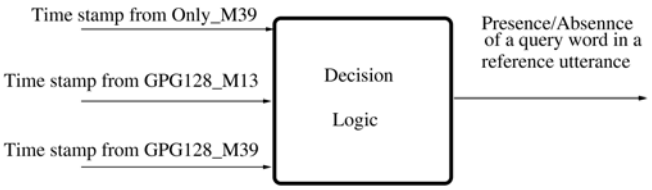

Figure 4. The block diagram of the proposed approach for hypothesizing query word in an utterance.

We have obtained the hypothesized time stamp information of each of the query words in all the reference utterances by the three systems namely Only_M39, GPG128_M13 and GPG128_M39. The presence of query word in the reference utterance is hypothesized by the following decision logics:

A. Decision by Any One System

Whenever the hypothesized time stamp of any one of the systems overlaps with the ground truth (time stamp in the reference utterance), then it is assumed (True) that the corresponding reference utterance has the query word.

B. Majority Voting Decision Logic

Whenever the hypothesized time stamp of at least any two systems overlaps with each other, then it is assumed (True) that the corresponding reference utterance has the query word.

Table II illustrates the few details on the results obtained by the proposed approach.

Table II. Time stamp obtained by the three different systems (Only_M39, GPG128_M13, GPG128_M39) with respect to few query words and few reference utterances. The ground truth of the query word in the reference utterance is also provided. The decision by any one system and decision by majority voting are obtained to study the overall performance of AIR system.

Query word	Reference utterance	Ground Truth (onset offset) (in sec.)	Only_M39 (onset offset) (in sec.)	GPG128_M13 (onset offset) (in sec.)	GPG128_M39 (onset offset) (in sec.)	Decision by any one system	Decision by Majority Voting
water_dr1_mdab0_sa1	dr2-feac0-sx75	1.96 2.27	1.25 1.39	2.37 2.53	2.19 2.36	True	True
ocean_dr2_mcem0_sx228	dr5-mdas0-sx6	1.62 1.94	0.44 0.79	0.48 0.69	0.48 0.77	False	True
mother_dr2_mcem0_sx48	dr7-fkde0-si1141	0.19 0.87	2.51 2.81	1.95 2.26	1.89 2.1	False	True
social_dr4_fcrh0_si458	dr7-fjsk0-si1052	0.21 0.75	1.83 2.35	1.84 2.35	1.83 2.3	False	True

From the Table II,the following observations are made:

(1) For the query word water (row 2), the presence of the query word in the reference utterance is hypothesized correct by both the logics.

(2) For the query word ocean (row 3), the presence of the query word in the reference utterance is hypothesized correct only by the Majority voting where as the decision by any one system fails.

(3) For the query word mother (row 4), the presence of the query word in the reference utterance is hypothesized correct only by the Majority voting where as the decision by any one system fails.

(4) For the query word social (row 5), the presence of the query word in the reference utterance is hypothesized correct only by the Majority voting where as the decision by any one system fails.

VII. RESULTS AND DISCUSSION

The performance of the AIR system on TIMIT database using 39-dimensional MFCC-Delta-Accelaration features, 128-dimensional Gaussian Posteriorgrams derived from 13-dimensional MFCC features, 128-dimensional Gaussian Posteriorgrams derived from 39-dimensional MFCC-Delta-Accelaration based features is given in the Table III.

Table III. The percentage of the AIR system on TIMIT database using 39-dimensional MFCC-Delta-Accelaration features, 128-dimensional Gaussian Posteriorgrams derived from 13-dimensional MFCC features, 128-dimensional Gaussian Posteriorgrams derived from 39-dimensional MFCC-Delta-Accelaration.

Query word spoken by	Frequency of Occurrence of query words in the reference utterances)	Correct hypothesis by decision by any one system (in percentage)	Correct hypothesis by Majority Voting (in percentage)
Male	52	33 (63.46%)	43 (82.69%)
Female	52	37 (71.15%)	45 (86.53%)
Both Male and Female	104	70 (67.31%)	88 (84.61%))

It is observed from the Table III that, the correct hypothesis of the query words in the reference utterances by using any one system logic is 67..31%. Also, in reality the ground truth (time stamps) of query words in the reference utterances will be unknown.

Thus we have developed three independent systems based on 39-dimensional MFCC-Delta-Accelaration features, 128-dimensional Gaussian Posteriorgrams derived from 13-dimensional MFCC features, 128-dimensional Gaussian Posteriorgrams derived from 39-dimensional MFCC-Delta-Accelaration features. By using Majority voting logic, it is possible to hypothesize the occurrence of query words without the knowledge of ground truth and is observed to be 84..61%. This result is encouraging for the task of AIR.

Further, we have analyzed our AIR system using the following metrics: (1) True Acceptance (TA), (2) True Rejection (TR), (3) False Acceptance (FA), and (4) False Rejection (FR). These metrics are defined as follows:
(1) True Acceptance (TA): Decision by Any One System logic is True and Decision Logic by Majority Voting is also True.

(2) True Rejection (TR): Decision by Any One System logic is True and Decision Logic by Majority Voting is False.

(3) False Acceptance (FA): Decision by Any One System logic is False and Decision Logic by Majority Voting is True.

(4) False Rejection (FR): Decision by Any One System logic is False and Decision Logic by Majority Voting is also false.

The details of the above metrics is provided in the Table IV.

Table IV. Performance of AIR system based on acceptance and rejection ratios.

Metric	Ratio
True Acceptance	(62/104)=59.61%
True Rejection	(8/104)=7.69%
False Acceptance	(26/104)=25%
False Rejection	(8/104)=7.69%

VIII. SUMMARY AND CONCLUSIONS

In this paper, we have explored the Mel-frequency cepstral coefficient (MFCC) based features and Gaussian Posteriorgram based features for audio information retrieval. There is no prior knowledge about the location of the query words in the reference utterances. Therefore, it is necessary to arrive at a conclusion about the existence of a query word in a reference utterance by using multiple evidences. In this regard we have built three independent AIR system by using 39-dimensional MFCC-Delta-Accelaration features, 128-dimensional Gaussian Posteriorgrams derived from 13-dimensional MFCC features, 128-dimensional Gaussian Posteriorgrams derived from 39-dimensional MFCC-Delta-Accelaration features. The studies are performed on TIMIT database. From the studies, it is evident that MFCC based features and Gaussian Posteriorgram based features are useful for the task of audio information retrieval.

IX. REFERENCES

[1] British Broadcasting Corporation. http://www.bbc.co.uk/archive/.

[2] TED (Technology, Entertainment and Design). http://www.ted.com/

[3] MIT Open Course Ware. http://ocw.mit.edu/index.htm.

[4] National Programme on Technology Enhanced Learning. https://onlinecourses.nptel.ac.in/.

[5] Gautam Mantena, Query-by-Example Spoken Term Detection on Low Resource Languages. PhD thesis, Language Technologies Research Center, International Institute of Information Technology, Hyderabad, India, 2014.

[6] I. Szöke, Hybrid word-subword spoken term detection. PhD thesis, Brno University of Technology, 2010.

[7] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in Proc. of HLT-NAACL, pp. 129–136, 2004.

[8] J. Makhoul, "Linear prediction: A tutorial review," Proc. of IEEE, vol. 63, pp. 561 – 580, April 1975.

[9] L. Rabiner, B.-H. Juang, and B. Yegnanarayana, Fundamentals of Speech Recognition. Prentice-Hall, Inc., 1993.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal of Acoustical Society of America, vol. 57, pp. 1738–1752, Apr. 1990.

[11] L. R. Rabiner and B. -H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersey: PTR Prentice Hall, 1993.

[12] Fisher William M, Doddington George R, Goudie Marshall, and Kathleen M, "The DARPA speech recognition research database: Specifications and status," pp. 93-99, 1986.

[13] https://catalog.ldc.upenn.edu/ldc93s1.

[14] F. W. F. J. P. D. Garofolo J, Lamel L and D. N., "Darpa, timit acoustic-phonetic continuous speech corpus cd-rom," National Institute of Standards and Technology, 1990.

[15] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 28, pp. 357–366, Aug. 1980.

[16] Steve Young et al., The HTK Book (for HTK Version 2.2). Cambridge: Entropic Ltd, 1999.

[17] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in Proc. of ASRU, 2009.

[18] H. B. G. Aradilla and M. Magimai-Doss, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in Proc. of ICASSP, 2009.

[19] J. V. G. Aradilla and H. Bourlard, "Using posterior-based features in template matchingfor speech recognition," in Proc. of Interspeech, 2006.

[20] W. S. T. Hazen and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in Proc. of ASRU, 2009.

[21] Sakoe,H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. on Acoust., Speech, and Signal Process., ASSP 26, 43-49 (1978).