International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
357 – 362

Audio Information Retrieval using Sinusoidal Modeling based Features

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

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

***Abstract*—** In this paper, we propose an approach for searching a speech query word in an audio data base. In this approach, we explore Sinusoidal modeling based features such as Amplitude, Frequency and Phase of the speech signal. Three independent systems are built using these features. Further, the Majority voting logic is used to arrive at a conclusion to locate (time stamp) the query word in the reference utterances. The studies are performed on the TIMIT database. The results show that, Sinusoidal based features can be used for speech processing in place of conventional approaches.

***Keywords*—** *Audio information retrieval, Sinusoidal modeling, Amplitude, Frequency, Phase, Dynamic time warping,*

*****

I. INTRODUCTION

Spoken audio data is available from various sources such as (a) Recordings from radio and television stations (British Broadcasting Corporation (BBC) archives [1], (b) Recordings from talks at conferences (Technology, Entertainment, Design (TED) talks}) [2], (c) Recordings from lectures (MIT OpenCourseWare (MIT OCW) [3], (d) National Programming on Technology Enhanced Learning (NPTEL) [4], and (e) Telephone recordings in a call center, etc. Thus, with the increasing amounts of spoken audio one needs to rely on automatic techniques to find relevant information within them.

The task of a Audio Information Retrieval (AIR) is to find a speech query (or keyword) within an audio. A key aspect of AIR is to enable searching in multi-lingual and multi-speaker audio data. Given the possible applications to deal with the increasing amounts of spoken audio, in this work, we are interested in searching a query within an audio data.

In general, one has to listen to the complete spoken audio to find whether the query exists or not. A solution is to manually transcribe the spoken audio and then use text-based search techniques. However, manual transcription of spoken audio is expensive, time consuming, and difficult for less-spoken languages.

This paper is organized as follows: In Section II, we describe the review of approaches for AIR. The description of databased in the studies is given in Section III. Speech Analysis based on a Sinusoidal Representation is provided in Section IV. In Section V, Sinusoidal Modeling based Features for audio information retrieval are described. Experimental Studies are described in Section VI. Hypothesizing query words in an utterance by using Amplitude, Frequency and Phase based systems is given in Section VII. AIR results are analyzed in Section VIII. Section IX provides the summary and conclusions of the current studies on AIR.

II. REVIEW OF APPROACHES FOR AUDIO INFORMATION RETRIEVAL

A. *Text-based Audio Information Retrieval*

A traditional audio information retrieval approach is to convert spoken audio into a sequence of symbols using Automatic Speech Recognition (ASR) and then perform the text based search. ASR-based technique assumes the availability of large amount of labeled data for training the acoustic and language models. The block diagram of audio information retrieval system using ASR is shown in Fig. 1.

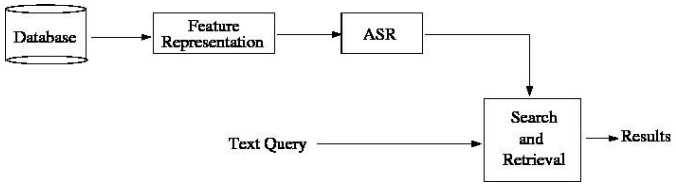

Fig. 1. ASR based Spoken Term Detection (STD) System

AIR using ASR based approach is not scalable for languages where there is no availability or the resources to build an ASR. Thus, there is a need to automate searching of spoken audio.

B. *Speech based Audio Information Retrieval*

To overcome the limitation of text-based search techniques, zero prior knowledge is assumed about the language of the spoken audio. In this paper, we propose speech based AIR. The block diagram of a generic approach for the audio information retrieval system based on speech query is shown in Fig. 2.

 ISSN: 2454-4248
357 – 362

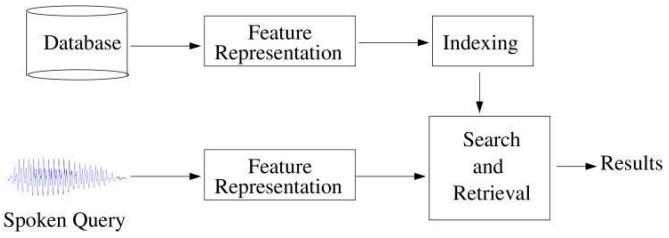

Fig. 2. Block diagram of Audio Information Retrieval system based on speech query [5]

C. *Existing Audio Features*

In a digital recording of speech, the signal is represented by discrete amplitudes as a function of discrete intervals of time. From a statistical point of view, these discrete values of the speech signal may not be used directly by many machine learning algorithms. Thus, there is a need to derive features from the speech signal best suited for the task.

The commonly used acoustic features for speech signal are as follows: (a) Linear prediction cepstral coefficients (LPCC), (b) Mel-frequency cepstral coefficients (MFCC), (c) perceptual linear prediction cepstral coefficients (PLP) and (d) frequency domain linear prediction cepstral coefficients (FDLP) [6, 7, 8, 9].

In general, acoustic parameters such as LPCC and MFCC are susceptible to speaker and environmental conditions. In this paper, we explore Source Filter Model-based features such as Sinusoidal modeling based features to represent the phonetic information present in the speech signal for the task of audio information retrieval [10].

D. *Search Techniques for Acoustic Similarity*

Normally euclidean distance measure is used to find the similarity between the two speech patterns. In this paper, we use Dynamic Time Warping for time alignment and normalization 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 labelled 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]. There are 462 speakers in training (reference) data. Each of the speaker has spoken 10 sentences. In all, there are 4620 utterances in training (reference) data. In a similar way, there are 168 speakers in test data. Each of the speaker has spoken 10 sentences. In all, there are 1680 utterances in test data. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. This data base 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 a Sinusoidal Representation

One of the approaches to the problem of representation of speech signal is to use speech production model in which speech is viewed as the result of passing a glottal excitation waveform through a time-varying linear filter that models the resonant characteristics of the vocal tract. The block diagram of the speech production model is shown in Fig. 3.

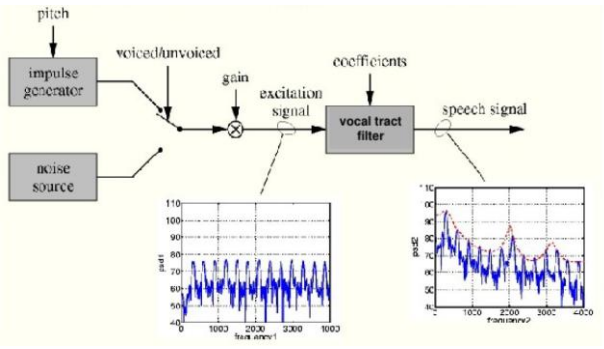

Fig. 3. Source Filter Model for Speech Production.

 ISSN: 2454-4248
357 – 362

In many speech applications it suffices to assume that the glottal excitation can be in one of two possible states, corresponding to voiced or unvoiced speech. In attempts to design high-quality speech coders at the midband rates, generalizations of the binary excitation model have been developed. In this work, the goal is also to generalize the model for the glottal excitation, but instead of using impulses as in multipulse, the excitation waveform is assumed to be composed of sinusoidal components of arbitrary amplitudes, frequencies, and phases. These parameters are estimated from the short-time Fourier transform using a peak picking algorithm. This is called Sinusoidal modeling based approach.

In Sinusoidal modeling based approach, the speech signal is modeled as [10][15][16]:

$$s[n] = \sum_{k=-K}^{K} a_k e^{j2\pi \frac{f_k}{f_s} n}$$

where a_k's are complex amplitudes and f_k's are frequencies. The phase information is obtained from complex amplitudes of a_k's.

Re-synthesis of the speech signal from a_k's and f_k's is obtained from the following equation:

$$s[n] = \sum_{k=1}^{K} 2|a_k| \cos\left(2\pi \frac{f_k}{f_s} n + \angle a_k\right)$$

Furth
er, a metric called Signal-to-Reconstruction-Error Ratio (SREER) to measure the similarity of original and resynthesized speech waveforms s[n] and s^[n] is defined as:

$$SRER = 20 \log_{10} \frac{\text{std}(s[n])}{\text{std}(s[n] - \hat{s}[n])}$$

Where std is the standard deviation of a speech signal.

The speech signal is locally synthesized as:

$$s[n] = \sum_{k=1}^{L_k} A_k[n] \cos(\Phi_k[n])$$

and the instantane
ous amplitude and phase inside a window supported in [0, N] are given by:

$$A_k[n] = A_k[0] + \frac{A_k[N] - A_k[0]}{N} n$$
$$\Phi_k[n] = \Phi_k[0] + \omega_k[0]n + a_k n^2 + b_k n^3$$

where

$$a_k = \frac{3}{N^2}\left(\Phi_k[N] - \Phi_k[0] - \omega_k[0]N + 2\pi M_k\right) - \frac{1}{N}\left(\omega_k[N] - \omega_k[0]\right)$$
$$b_k = -\frac{2}{N^3}\left(\Phi_k[N] - \Phi_k[0] - \omega_k[0]N + 2\pi M_k\right) + \frac{1}{N^2}\left(\omega_k[N] - \omega_k[0]\right)$$

where N is the length of a analysis window, and Fs is the sampling frequency.

Using the following parameters the speech signal is reconstructed.

Fs: Sampling Frequency: 16000Hz

N: Analysis window length: 25ms

S: Analysis step size: 15 ms

L: Number of sinusoidal components: 80

W: Window that will be applied to each frame: Hanning

Consider a speech signal spoken by female speaker in Kannada language. The total duration of the speech signal is 2 minutes 56 seconds and 67 milliseconds. Out of which we have considered the initial duration of 1 second. It has 16000 speech samples. Fig. 4, shows the original, resynthesized and error signal. The Signal to reconstruction error ratio (SRER) is found to be 16.46 dB.

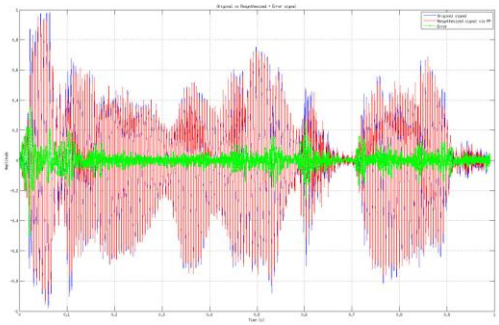

Fig. 4. Original speech signal is represented as blue. The resynthesized speech signal is represented as red. The error signal (difference between the original speech and resynthesized speech signal) is represented by green.

For more clarification, we consider 200 ms duration in the given signal. Typical fundamental frequency of a female speaker will be 200 Hz. Thus time required for one time vocal fold vibration will be 5 ms. Hence in 200 ms speech signal, it is expected that 40 such cycles. This is evident from the Fig. 5.

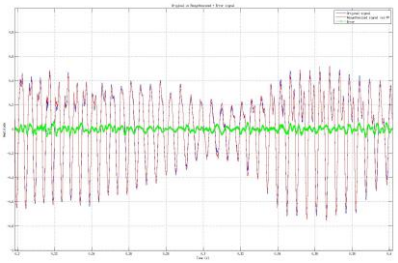

Fig. 5. Original speech signal of duration 200 ms represented as blue. The resynthesized speech signal is represented as red. The error signal (difference between the original speech and resynthesized speech signal) is represented by green.

 ISSN: 2454-4248
357 – 362

V. Sinusoidal Modeling based Features for audio information retrieval

In this section, we propose an algorithm for the extraction of Sinusoidal Modeling based features representation from the speech signal.

1. Consider a frame of size 25 ms and frame shift 15 ms at a time.
2. Consider Hanning for windowing
3. Find the 2048 point Discrete Fourier Transform (DFT) of the current speech frame to get its frequency domain representation. This representation provides the amplitude, frequency and phase values for each DFT coefficient.
4. As DFT is symmetric take only 1024 DFT coefficients.
5. Find 80 peaks from these DFT coefficients which correspond to harmonics of the fundamental frequency.
6. These peaks are identified by peak picking algorithm.
7. Each peak is described by its amplitude, frequency and phase.
8. As we consider 80 peaks per frame, for each frame we get a vector of 80 amplitudes, a vector of 80 frequencies and a vector of 80 phases.

The speech sampling rate of the analog-to-digital converter is assumed to be 16 kHz. The speech samples are divided into overlapping frames. The frame length is 25 msec (400 samples) and the frame shift is 15 msec (240 samples). Each frame is windowed using the Hanning window function. For each frame we get a vector of 80 amplitudes, a vector of 80 frequencies, and a vector of 80 phases.

VI. Experimental Studies

We have extracted the Amplitude, Frequency and Phase features for all the reference utterances of TIMIT database. Further, we have also extracted the Amplitude, Frequency and Phase features for the chosen query words.

A. Evidence of location (time stamp) of query words in reference utterances with respect to Amplitude based DTW system (ABS)

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 Amplitude based features.

B. Evidence of location (time stamp) of query words in reference utterances with respect to Frequency based DTW system (FBS)

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 Frequency based features.

C. Evidence of location (time stamp) of query words in reference utterances with respect to Phase based DTW system (PBS)

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 Phase based features.

VII. Hypothesizing query words in an utterance

The presence of query word in a reference utterance is hypothesized from the time stamp information obtained from Amplitude, Frequency and Phase based DTW systems. The block diagram of the proposed approach for hypothesizing query word in an utterance is shown in Fig. 6.

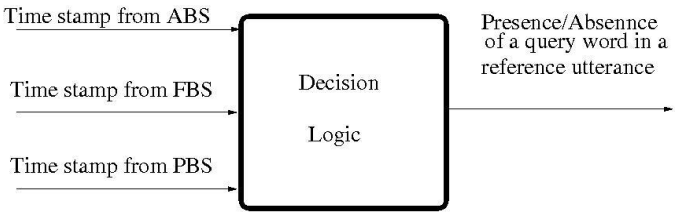

Fig. 6. The block diagram of the proposed approach for hypothesizing query word in an utterance.

We have obtained the hypothesized time stamp information of each of the the query words in all the reference utterances by the three systems namely ABS, FBS and PBS. 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. Decision Logic by Majority Voting

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.

 ISSN: 2454-4248
357 – 362

Table II illustrates the few details on results obtained by the proposed approach. From the table 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 country (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 ocean (row 4), the presence of the query word in the reference utterance is hypothesized correct by both the logics.

(4) For the query word mother (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.

(5) For the query word social (row 6), 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.

TABLE II. TIME STAMP OBTAINED BY THE THREE DIFFERENT SYSTEMS (ABS, FBS, PBS) 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.)	ABS (onset offset) (in sec.)	FBS (onset offset) (in sec.)	PBS (onset offset) (in sec.)	Decision by any one system	Decision by Majority Voting
Water_dr1_faks0_sa1	dr2-feac0-sx75	1.96 2.27	2.01 2.30	1.69 1.98	2.02 2.30	True	True
country_dr2_mccs0_sx389	dr4-feeh0-si1112	0.83 1.31	3.25 3.51	2.58 2.90	3.19 3.51	False	True
ocean_dr2_mcem0_sx228	train-dr3-falk0-sx6	1.79 2.20	1.82 2.20	0.35 0.72	1.79 2.17	True	True
mother_dr4_fdms0_sx48	dr6-mtxs0-si2320	1.47 1.95	2.4 2.64	2.12 2.41	2.26 2.45	False	True
social_dr4_fcrh0_si458	dr3-mtpp0-sx158	0.85 1.28	1.95 2.48	1.89 2.37	2.10 2.50	False	True

VIII. RESULTS AND DISCUSSION

The performance of the AIR system on TIMIT database using sinusoidal modeling based features is given in the Table III.

TABLE III. THE PERCENTAGE OF THE AIR SYSTEM ON TIMIT DATABASE USING SINUSOIDAL MODELING BASED FEATURES.

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	29 (55.76%)	26 (50.00%)
Female	52	35 (67.30%)	35 (67.30%)
Both Male and Female	104	64 (61.53%)	61 (58.65%))

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 61.53%. However 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 Amplitude, Frequency, and Phase 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 58.65%.

Further, we have analyzed our AIR system using the following metrics: (1) True Acceptance (TA), (2) True Rejection, (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	(39/104)=37.5 %

 ISSN: 2454-4248
357 – 362

True Rejection	(25/104)=24.03%
False Acceptance	(22/104)=21.15%
False Rejection	(18/104)=17.30%

IX. SUMMARY AND CONCLUSIONS

In this paper, we have explored the Sinusoidal modeling 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 Amplitude, frequency and Phase features derived from Sinusoidal modeling approach. The studies are performed on TIMIT database. From the studies, it is evident that Sinusoidal modeling based features can be explored for other speech processing applications as well.

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] Robert J. McAulay and Thomas F Quatieri, "Speech analysis/synthesis based on a simumoidal representation," IEEE Trans. Acoust., Speech, and Signal Processing, vol. 34, pp. 744–754, Aug. 1986.

[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] George P. Kafentzis and Yannis Stylianou, "Stationary Sinusoidal Modeling," in Hands on Session: Stationary Sinusoidal Modeling, Summer School, India (GIAN course on Advanced Sinusoidal Modeling of Speech and Applications 2016), December 2016.

[16] Stylianou Ioannis, Harmonics plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, Ecole Nationale Superieure de Telecommunications, Paris, France., 1996.