ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Real Time Speech Enhancement And Recognition Using Enhanced Non Negative Matrix Factorization

[1]Akarsh K.A., [2]Senthamizh Selvi R. and [3]Suresh G.R.

[1]*P.G. Scholar (Embedded System Technologies), Department of Electronics and Communication Engineering, Easwari Engineering College (Anna University), Chennai, India.*
[2]*Assistant Professor, Department of Electronics and Communication Engineering, Easwari Engineering College (Anna University), Chennai, India.*
[3]*Professor, Department of Electronics and Communication Engineering, Easwari Engineering College (Anna University), Chennai, India.*

ARTICLE INFO

Article history:
Received 20 January 2015
Accepted 02 April 2015
Published 20 May 2015

Keywords:
Voice eXtensible Markup Language (VXML) World Wide Web Consortium (W3C) Text-To-Speech (TTS) Hidden Markov Model (HMM) Automatic Speech Recognition (ASR) Dual Tone Multi Frequency (DTMF) Enhanced Non negative Matrix Factorization (ENMF)

ABSTRACT

Speech enhancement improves the quality and intelligibility of tainted speech signal. Speech recognition technology use Voice XML. In order to create voice-user interfaces VXML of the W3C (World Wide Web Consortium) is used. It uses speech recognition and touchtone (DTMF) for input. Speech recognition technology deals with pre-recorded audio in the front end and text-to-speech synthesis (TTS) from the database are used for output. The text-to-speech synthesis characteristic of advanced Voice XML tools like Web Sphere opens novel perspectives for e-commerce and e-learning. A speech recognition system transforms a human form of speech into the machine form. The human form of speech is typically referred to natural language. Speech recognition process is frequently referred to as Automatic Speech Recognition (ASR). This technology is used to understand human voice and to interact with humans using verbal commands. For example, a telephone banking system interacts with its users and asks for their account numbers and passwords. Such systems speak to the users by means of speech recognition software and can appreciate human voice. Through these systems, users can carry out transactions using speech as an alternative of pressing buttons on the keypad. This work deals with speech recognition application for student's information retrieval in college. The sole purpose of this work is to develop an improved speech recognition application for visually impaired or those who prefer to listen. HMM algorithm is used to develop a speech recognition application. The background noise during the real time speech recognition can be removed by using Enhanced NMF algorithm. The voice enabled application is the best practice of application design, development in software engineering. Voice XML renders content as speech and interacts with the user using speech recognition and speech synthesis technologies.

© 2015 AENSI Publisher All rights reserved.

To Cite This Article: Akarsh K.A., Senthamizh Selvi R., and Suresh G.R., Real Time Speech Enhancement And Recognition Using Enhanced Non Negative Matrix Factorization. ***Aust. J. Basic & Appl. Sci.,*** *9(16): 67-76, 2015*

INTRODUCTION

Voice-enabled applications are accessible in more than a few areas, for example: Voice response amenities are used for various kinds of information over the phone: time, weather, horoscopes, lottery results, sports events, news, exchange rates, and so on. A bank can allow its clientele access their account balances, get information on interest rates and mortgages, calculate loan payments, or transfer funds, all using voice response applications. An application can also call customers to request about transactions such as renewing a Certificate of Deposit. Using a voice response application, brokerage firms can create current stock prices, quotations, and collection balances available over the telephone. Customers can carry out complex transactions without the intervention of an agent. When an agent's advice is required, the application can transfer the call. A voice response application can give information about class schedules, placement details, and course content. Students can register by means of the telephone, and the application that handles the registration procedure can also update the database containing enrolment information. Student's semester details are also provided using IVR. Now-a-days every institution needs computerization. This project work allows the user to know the student attendance and marks quickly through real time speech recognition technology. Interactive Voice Response (IVR) (From Wikipedia, 2015) is a software appliance that accepts

Corresponding Author: Akarsh K.A., P.G. Scholar (Embedded System Technologies), Department of Electronics and Communication Engineering, Easwari Engineering College (Anna University), Chennai, India.

a combination of voice from microphone and DTMF keypad selection and provides appropriate responses in the form of voice, e-mail and perhaps other media voice applications. IVR is a part of a larger voice application that includes database access from MySQL. This work focus towards the description of an integration methodology of web contents steering and spoken driven information accessing using DTMF and VXML technologies. The application will focus on educational institutions (Oviatt, S.L., 1997).

The outline of this paper is organized as follows. Section II describes the speech recognition technology using VXML and Enhanced NMF (ENMF). Section III, describes the evaluation results of the work. Finally, Section IV concludes the paper.

1. Methodology:

This paper gives the solution for the entire computerization of Educational Institutions. The voice response application by the system shall cover the following informational requirements through the voice response:

1. Due status of the student/ Installment of Fees paid in educational institutions.
2. Attendance status for any week, month, day or entire year of the students.
3. Grade/ Mark scored in any exam.
4. Rank in any exam.
5. Rank, Score and percentage in particular subject.
6. Teacher's remarks to students.
7. Timetable for university exams.
8. Test syllabus for the course.
9. Placement details or news in college.
10. Faculty vacancies, if any.
11. Important news for parents like dates for parents-teachers meetings or any other messages regarding the students.
12. Automatic Fee Reminders due dates.
13. Login details of each and every students.

A. Voicexml:

Internet is our main information repository nowadays and web browsing the natural access mechanism to it. However, network browsing is not the preponderance used way to access information in daily life. Mobile devices manage the market and voice is the famous communication channel in this environment. Since spoken conversation based interaction is fundamentally affected by chronological mode of operation, complemented in best cases with some form of asynchronous reaction from the user side, the careful design of spoken dialogue systems starting from their counterparts in web environments is still an open question. Voice XML is becoming a standard for browsing voice contents (Bruce Lucas, 2013). It takes thoughts from HTML, and borrows web information browsing model. Voice XML builds on the basic concept and rules set by XML (VoiceXML, 2015). Interactive applications contain synthesized speech, pre-recorded audio, grammars defining words that could be recognized, and DTMF key input. By saying something or pressing the keypad on the system the user can get different details.

Brief History:

It all started with a research project called Phone Web at AT&T Bell Laboratories. This became the basis of Voice XML (Schnelle-Walka, D., 2010). The project's goal was to create a development tool that combined traditional telephony services with the web based services. Another important issue was that it should also be easy for a person that was not directly involved with the development of the technique to create new services. Later, the company was reorganized into AT&T Labs and Lucent's Bell Labs and they both continued to develop independent versions of their markup languages. There were two other companies that also should be considered as pioneers; IBM's Speech ML and Motorola's Vox ML. These four companies understood that they had to come up with a standard and together they created a forum called Voice XML. The XML is used as the source for this effort, because they thought that this was the direction that web technology was going. The forum grew and in March 2000, Voice XML version 1.0 was released. Four years later, Voice XML is the typical scripting language for accessing Web pages over the real time speech recognition. In March 2004, the World Wide Web Consortium (W3C) has advanced Voice XML 2.0 to the recommendation status, and thereby established as a Web Standard. In December 2010, W3C released Voice XML version 3.0.

Concepts:

A solitary Voice XML document, or a set of VXML documents, called an application, forms an informal finite state machine files. The user can use only a one state, or dialog, at a time in front end coding (Home page for the World Wide Web Consortium for Voice applications, 2015). Each dialog in VXML contains information about the transition to the next link dialog. The transitions are precise by URI's in the commands. The completing of an application ends when a dialog does not identify a successor, or if it has a constituent that openly exits the application. Voice XML has two types of dialogs; forms and menus. The difference between them is that the form defines an interaction from the user that collects standards for a set of appearance item variables, while a menu presents the user with some option and goes to other dialogs based on a choice. A sub dialog is able to be viewed like a function call. It provides a mechanism to invoke a new interaction, and following that recurring to the unique form. Variable instances, grammars, and condition information are saved and are available upon chronic to the calling file. Sub

dialogs can be worn to make a corroboration series that may necessitate a database inquiry to generate a set of mechanism that may be shared among documents in its own application or to create a reusable documentation of dialogs collective among many applications. A process begins when the user starts to interrelate with the Voice XML interpreter context, and ends by a demand from the user, a document, or the interpreter context (HP VoiceXML developer resources, 2015). New documents, grammars and audio files are loaded and processed while the session is still alive.

To build an application using VXML:

When building a voice application, it is useful to define resources like variables, grammars, scripts, and event handlers that can be shared across the application. Variables declared in the application root document allow you to share state across documents that make up your application. Having scripts in your application root document allows you to load the script once for your application. Grammars and event handlers declared in the application root document are active throughout the application. Forms are one of the most important components in Voice XML (Larson, A., 2013). Voice XML forms are analogous to web forms (Beasley, R., 2013). It is used to collect information (voice input) from the user. A form contains items, elements that are used by the main and important loop of the Form Interpretation Algorithm (FIA). The form items can be divided into two different types, input items that can be filled by the user and control items that cannot be filled. The form contains declarations of event handlers, non-form item variables and filled actions. The <filled> element indicates a particular action to perform when some combination of input items are assigned. It is used to collect information (voice input) from the user. VXML can be combining with JSP for human computer interaction.

Figure: 1 shows the structure of VXML [16]. The end user use phone or computer to retrieve information it is connected with Public Switched Telephone Network (PSTN). The VXML gateway provides the connection with internet. The server used is Voxeo server (Bob, C., Edgar, 2012). Voxeo server is mainly used for SIP based voice applications. The end users can see or hear the output data's using computer or mobile phones. The speech is the only way of communication in VXML.

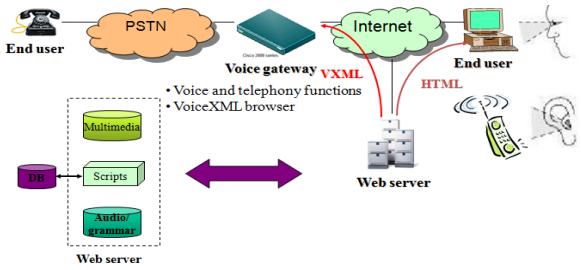

Fig. 1: Voice XML structure.

B. HMM Speech Recognition:

1) Hidden Markov Model (HMM):

This model assumes the system to be a Markov process with unnoticed states. This statistical Markov model is the most popular and successful stochastic approach for speech recognition. This is due to the continuation of elegant and efficient algorithms for both recognition and training. The HMM which is also known as the acoustic model, is required to determine the most likely word succession known some speech data, in conjunction with the language model (Hirsch, H.G. and H. Finster, 2013). The acoustic model is required to give the probability of each possible word sequence, specifically within this mentioned process. From Eqn. (1), matrix of transition probabilities are $X = (x_{ij})$, the matrix of the observation probabilities are $Y = (y_i(v_m))$ and a vector of initial probabilities $\Pi = (\Pi_i)$.

$$M = (X, Y, \Pi) \quad (1)$$

2) HMM for speech recognition:

Hidden Markov Model lie at the heart of all modern speech recognition systems and although the basic framework has not changed significantly in the last decade or more, the detailed modelling techniques developed within this framework have evolved to a state of considerable sophistication (Lawrence, R., 2008). The result has been steady and significant progress in real time speech recognition process. The principal component of a large vocabulary continuous speech recognition model is illustrated in Figure 2. The input audio waveform from a microphone is converted into a sequence of fixed size acoustic vectors $Y_{1:T} = y_1, \ldots, y_T$ in a process called feature extraction. The decoder then attempts to find the sequence of words $w_{1:L} = w_1,\ldots,w_L$ which is most likely to have generated Y, i.e. the decoder tries to find:

$$\hat{w} = \arg\max\{P(W|Y)\} \quad (2)$$

However, since $P(w|Y)$ is difficult to model directly, Bayes rule is used to transform (2) into the equivalent problem of finding:

$$\hat{w} = \arg\max\{p(Y|w)P(w)\} \quad (3)$$

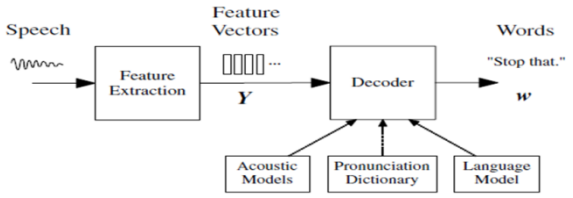

Fig. 2: Architecture of a HMM-based Recognition.

For any given w, the corresponding acoustic model is synthesized by concatenating models to make words as defined by a pronunciation dictionary. The parameters of these models are estimated from training data consisting of speech waveforms and their orthographic transcriptions. The language model is typically an N-gram model in which the probability of each word is conditioned only on its $N-1$ predecessors (Cheng, N., 2011). The N-gram parameters are estimated by counting N-tuples in appropriate text corpora. The decoder operates by searching through all possible word sequences using pruning to remove unlikely hypotheses thereby keeping the search tractable. When the end of the utterance is reached, the most likely word sequence is output. Alternatively, modern decoders can generate lattices containing a compact representation of the most likely hypotheses.

The feature extraction stage seeks to provide a compact representation of the speech waveform. This form should minimize the loss of information that discriminates between words, and provide a good match with the distributional assumptions made by the acoustic models. For example, if diagonal covariance Gaussian distributions are used for the state-output distributions then the features should be designed to be Gaussian and uncorrelated. Feature vectors are typically computed every 10 ms using an overlapping analysis window of around 25 ms. One of the simplest and most widely used encoding schemes is based on Mel Frequency Cepstral Coefficients (MFCCs) (Nilsson, M. and M. Ejnarsson, 2012). These are generated by applying a truncated Discrete Cosine Transformation (DCT) to a log spectral estimate computed by smoothing an FFT with around 20 frequency bins distributed non-linearly across the speech spectrum. The nonlinear frequency scale used is called a mel scale and it approximates the response of the human ear. The DCT is applied in order to smooth the spectral estimate and approximately de-correlate the feature elements. After the cosine transform the first element represents the average of the log-energy of the frequency bins. This is sometimes replaced by the log-energy of the frame, or removed completely.

3) MySQL Database:

The most popular and efficient Open Source SQL database management system is MySQL. It is named after co-founder Monty Widenius's daughter, My. The name of Dolphin in MySQL logo is "Sakila". The official way to pronounce "MySQL" is "My S-Q-L" (not "my sequel"). The SQL part of "MySQL" stands for "Structured Query Language." MySQL is supported, developed, framed and distributed by Oracle Corporation. MySQL is a Relational Data Base Management System (RDBMS), and ships with no GUI tools. MySQL database manages data contained within the databases and is capable of building its own structures. Commands to be entered on the MySQL prompt are preceded by mysql> (Rouillard, J., 2014). UNIX commands are case sensitive whereas mysql commands (except passwords) are not. MySQL database act as backend to store college details in and student's information.

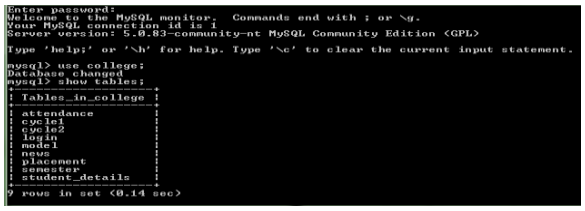

Fig. 3: MySQL database for college information retrieval.

Figure: 3 shows the database content for the college students details. The data's can be processed by using commands. In this we can insert table contents. It is possible to edit the content in MySQL database table. A database is a way of storing data. Databases are specifically designed to be very

efficient in storing and retrieving data. The most common way of structuring data in a database is relational database system. A relational database is a structured collection of tables. Each table consists of rows called records. Columns of the table have keys called fields. Each cell contains data. Most tables have a special column that identifies the rows of the table. The values in this column are called primary keys.

```
mysql> select * from semester;
+--------+---------------------+-----+-----+------+-----+-----+------+------+
| reg_no | name                | DES | EN  | DECS | DIP | AES | MPMC | CGPA |
+--------+---------------------+-----+-----+------+-----+-----+------+------+
| 1001   | AKARSH.K.A          | B   | C   | B    | A   | B   | B    | 8.04 |
| 1002   | Bruntha.M           | C   | D   | A    | C   | B   | A    | 7.26 |
| 1003   | Ganesh Murthy.M.S   | A   | B   | C    | A   | D   | B    | 8.23 |
| 1004   | Ilamathi.M          | B   | A   | A    | C   | A   | A    | 8.52 |
| 1005   | Jarina Raihan.A     | A   | S   | A    | B   | C   | B    | 8.83 |
| 1006   | Kotha Nagarjuna     | C   | C   | E    | B   | E   | D    | 7.24 |
| 1007   | Manikandan.J        | A   | B   | B    | A   | C   | B    | 8.31 |
| 1009   | Mohan kumar.K       | B   | C   | D    | E   | D   | C    | 7.44 |
+--------+---------------------+-----+-----+------+-----+-----+------+------+
```

Fig. 4: MySQL database for student's semester details.

Figure: 4 shows the MySQL database of student semester details. It can be obtained from the command "select * from semester". * indicate all available data's in the database. In the student information retrieval MySQL database is act as the back end tool that stores the student details such as cycle test mark, attendance, login details of student, semester details, college news etc.,

C. Speech enhancement using Enhanced NMF:

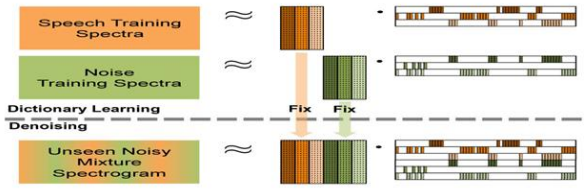

Fig. 5: Block diagram of Enhanced NMF.

The problem of recognizing speech in the presence of background noises remains a difficult one without a satisfactory solution to date (Loizou, P.C., 2013). A large number of algorithms have been proposed in the literature to address the more general problem of mitigating the effect of noise on the speech signal. Many of these attempt to reduce the noise from the speech signal itself. Other techniques modify features derived from the signal to reduce the effect of noise. Figure 5 describes the entire speech enhancement process using Enhanced NMF. The early step is to learn basis vectors from each source (orange and green) as in the dictionary learning. In Enhanced NMF however, the basis vectors for a source are grouped into a pre-defined number of blocks, e.g. in the figure five bases per a block and three blocks per a source, each of which corresponds to a convex cone, or a local dictionary. At the same time the activation matrix H is learned to be block-wise sparse, so that training sample belongs to only one or a very small number of local dictionaries. The speech enhancement algorithm using Enhanced NMF (ENMF) is given below:

1: Set the input values from the speech signal.
2: Output H^S, H^N.
3: Initialize H^S and H^N and with random numbers.
4: **repeat**
5: Update H values.
6: **if** Unsupervised then
7: Update the dictionary
8: **else**
9: Update the source-specific training signals.
10: **end if**
11: **until** Convergence

RESULTS AND DISCUSSION

To analyze and to evaluate the performance of the system, TIMIT database has been used in enhancement of speech signal. Different types of noise database namely Babble, Car, Exhibition, Restaurant, Street, and Train noise respectively. It contains 30 read sentences by different speakers. For assessment, one sentence of male and female speakers is used from the database. The speech data are originally sampled at 44.1 to 16 kHz sampling frequency. The data is appropriately filtered and down sampled to 8 kHz to obtain the narrow band speech which are used in our investigation. Performance of the speech enhancement system is evaluated in this section. Clean speech is manually corrupted at SNR level of -5, 0, 5, and 10 dB. The algorithm's performance is measured through Spectrogram and Spectrum of the signal. The signal enhanced by both algorithms is denoted as

'Enhanced speech'. ENMF algorithm is used to reduce the noise in real time speech recognition.

Figure 6 shows the spectrogram of real time recording speech, with background noise such as fan noise and babble noise. The spectrogram of noisy signals corrupted with fan and babble noise of 0dB SNR.

Figure 7 shows the plot representation of noisy speech affected by different noises. The X-axis denotes the frequency and Y-axis denotes the amplitude values.

Figure 8 shows the waveform of real time recording speech, with background noise such as fan noise and babble noise. It can be sampled at the frequency of 16000Hz and 8000Hz. The gain can be according to the signal.

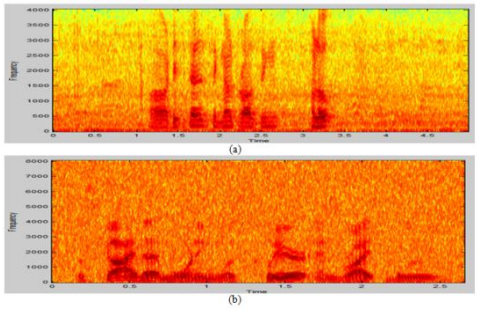

Fig. 6: Spectrogram of real time recording speech signal added with different background noises (a) Fan noise (b) Babble noise.

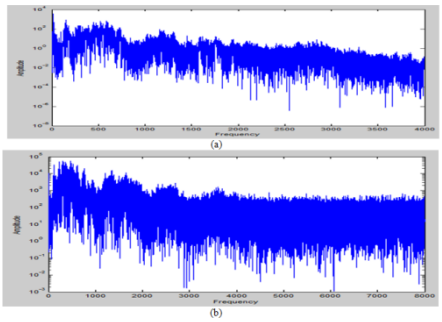

Fig. 7: Plot of real time recording speech signal added with different background noises (a) Fan noise (b) Babble noise.

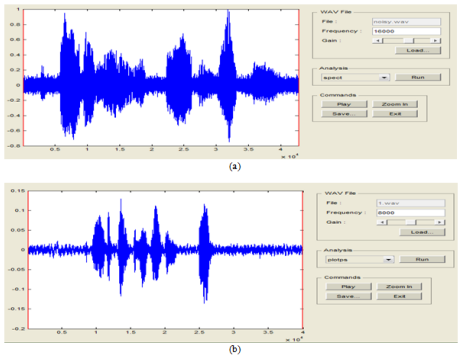

Fig. 8: Waveform of real time recording speech signal added with different background noises (a) Fan noise (b) Babble noise.

Figure 9 shows the spectrogram mesh representation of noisy speech, affected by different noises. The mesh is the 3D representation of speech signal.

Figure 10 shows the spectrogram, plot, waveform and mesh representation of clean speech signal without any noise added in background.

Figure 11 shows the result of the spectrogram, plot, waveform and mesh representation of degraded speech signal using Enhanced NMF algorithm. The proposed Enhanced NMF algorithm gives the better result. It can be revield from the graph, waveform and spectrogram representation of speech signals. The proposed ENMF algorithm for speech signal can

be added with VXML codings to reduce the noise in real time speech recognition.

Figure 11 shows the result of the spectrogram, plot, waveform and mesh representation of degraded speech signal using Enhanced NMF algorithm. The proposed Enhanced NMF algorithm gives the better result. It can be revield from the graph, waveform and spectrogram representation of speech signals. The proposed ENMF algorithm for speech signal can be added with VXML codings to reduce the noise in real time speech recognition.

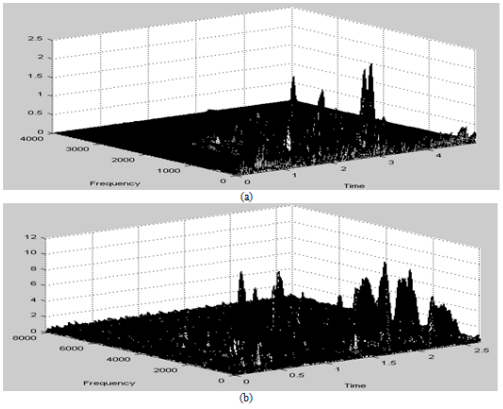

Fig. 9: Spectrogram mesh of real time recording speech signal added with different background noises (a) Fan noise (b) Babble noise.

Figure 12 shows the graph representation of Enhanced NMF algorithm parameters such as SNR VS SDR and SNR VS PESQ. SNR is Signal to Noise Ratio in dB. SNR compares the desired speech signal to the level of background noise. SDR is Source to Distortion Ratio in dB. SDR is defined as the ratio of the power of original modulating audio signal i.e., from a modulated radio-frequency carrier to the residual audio power i.e., noise-plus-distortion powers remaining after the original modulating audio signal is removed. Perceptual Evaluation of Speech Quality (PESQ) provides accurate and repeatable estimates of speech quality degradation occurring through e.g. a telephony network. It compares the audio signal input to a network with the corresponding (degraded) audio signal output from the network. Log Spectral Distance (LSD) value between the enhanced speech signal and the original speech signal is 1.4739dB in the Enhanced NMF method. The spectrogram of improved signal by Enhanced NMF gives the better result in real time speech enhancement technology. For the SDR and SNR, improvements gained by the enhancement systems are shown in below graphs.

The experimental result shows that the Enhanced NMF technique provides high range of noise reduction in speech recognition. It is could be a practical solution to the single-microphone speech recognition. The Enhanced NMF algotithms are analysed by using SNR, SDR, PESQ, LSD parameters. The advantage of this method is HMM algorithm is possible to combine with this proposed algorithm. HMM is used to detect the correct pronunciation of the speaker through the microphone. The speech is detected by the computer. The system performs the desired operation. This work is mainly concentrate on the educational institution student information retrieval system. The student can enter the register number and data of birth using SIP phone in DTMF technology. The system identifies the student from MySQL database. The system can then ask what information does the student needs. For example, if the student needs the semester grade details he/ she can say “semester details” through the microphone the system can identifies and say the semester grades of the particular student. If any noises occur in the background it can be reduced by using Enhanced NMF algorithm for speech [12]. This work paved the way to user friendly human system interaction. It will be very helpful for parents to know their children’s educational performance in colleges and schools. It will also be useful to the educational institutions to check the student details with in a less time. It will save the time and man power. The main problems remain from ASR while mapping the spoken inputs to the Voice XML grammar. Improving full sentence recognition must be a major aim for the next years. The main advantage of VXML is well structured software design process with flexibility for application development, reusable software solutions and limited efforts for initial training.

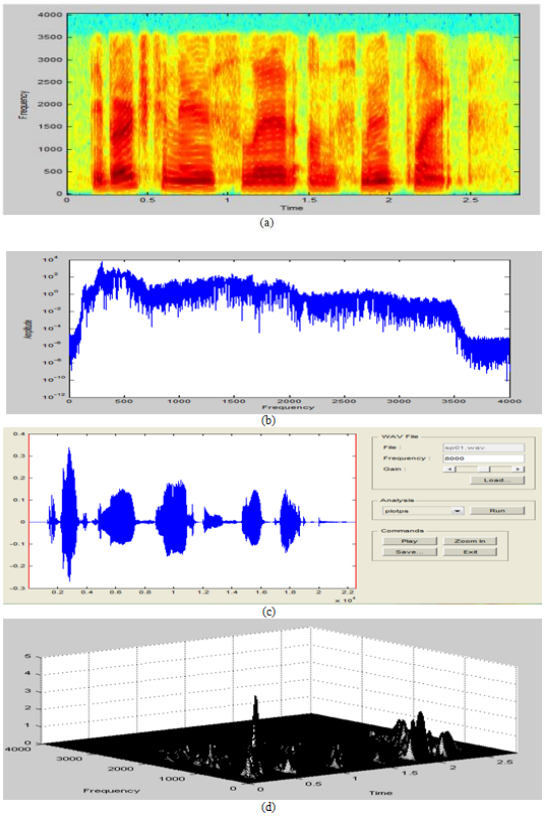

Fig. 10: Spectrogram, Plot, Waveform and Mesh of real time recording clean speech signal.

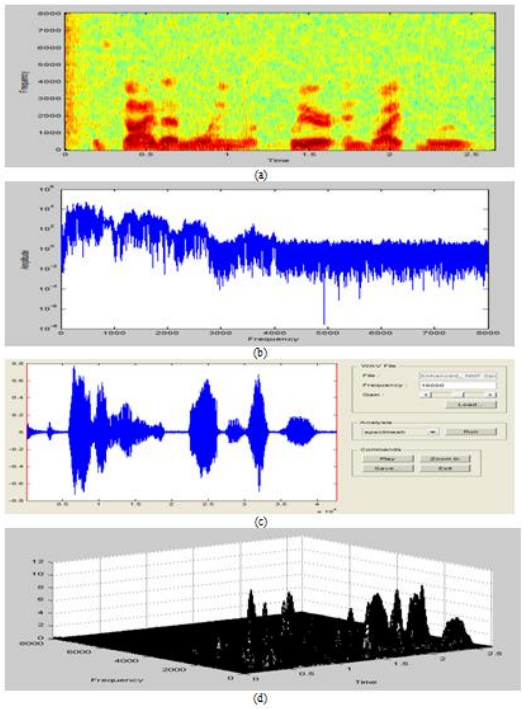

Fig. 11: Spectrogram, Plot, Waveform and Mesh of real time recording degraded speech signal using Enhanced NMF algorithm for real time speech recognition.

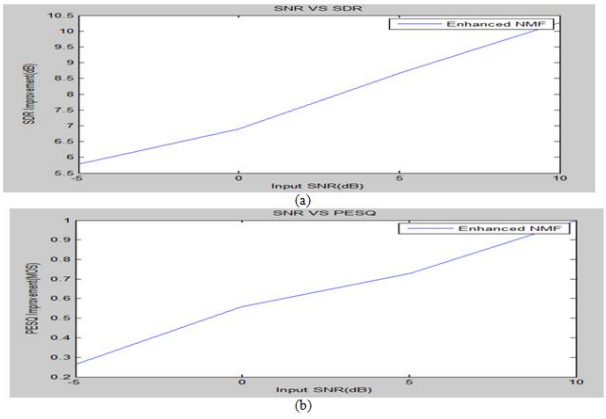

Fig. 12: Graph of Enhanced NMF based speech enhancement for babble noise (a) SNR VS SDR, (b) SNR VS PESQ.

4. *Conclusion:*

The proposed solution consists of a structured dialogue development system. This paper investigated the purpose of Enhanced NMF in speech recognition technology. The techniques used in this paper are Enhanced NMF, HMM, VXML. HMM is mainly used for speech reorganization purpose. Interactive Voice Response application deals with the human computer communication purpose. Speech recognition technology has three types of operations front end operation, back end operation and a middleware. Voice XML can perform the front end operation. MySQL database perform the back end operation and Java Server Page (JSP) is the connecting tool and perform middleware operation. JSP technology helps to create a dynamically generated web page for VXML application. The noise in the real time speech recognition can be reduced using an algorithm named as Enhanced NMF (ENMF). Enhanced NMF algorithm concentrates on the basis vectors of the local dictionaries. HMM algorithm perform viterbi search algorithm to detect the hidden data's. It provides substantial improvement in single microphone speech recognition task. The Enhanced NMF algorithm shows the better results in noise reduction. The automatic speech recognition results in fully computerization in educational institutions. The evaluation of this paper is mainly based on SNR, SDR, PESQ, and LSD. The computational time and manpower is reduced. This application is suitable for single microphone speech recognition. The future work of this paper is to enhance the intelligibility and quality of the real time speech using latest algorithms. The speech recognition application will be implementing in Blackfin Evaluation ADSP processor for fully automatic processing of real time speech. Questions of mispronunciation of, abbreviations, names, accentuations are to solve in the near future.

REFERENCES

Lawrence, R., 2008. Fundamentals of Speech Recognition. NewYork,NY, USA: Pearson Education.

Cheng, N., X. Liu and L. Wang, 2011. "Generalized variable parameter HMMs for noise robust speech recognition," in Proc. ISCA Interspeech' 11, Florence, Italy, 482–484.

Bruce Lucas, 2013. "VoiceXML For Web-based Distributed Conversational Applications", Communications of the ACM, 43(9): 53-57.

Schnelle-Walka, D., 2010. Voi)ceXML - The Open Source VoiceXML Interpreter.

Nilsson, M. and M. Ejnarsson, 2012. Speech Recognition Using Hidden Markov Model Performance Evaluation in Noisy Environment, Blekinge Institute of Technology Sweden.

VoiceXML, 2015. Forum is a global industry organization that works to accelerate the adoption of VoiceXML and adjacent technologies. The reference is taken from the platform certification section of the forum. 'VoiceXML: The Voice eXtensible Mark-up-Language', http://www.voicexml.org.

Home page for the World Wide Web Consortium for Voice applications, 2015. http://www.w3c.org/Voice.

Beasley, R., 2013. Voice Application Development with VoiceXML. USA: Sams Publishing. (ISBN 0-672-32138-6).

From Wikipedia, 2015. description about interactive voice response system http://en.wikipedia.org/wiki/Interactive_voice_response.

Larson, A., 2013. Prentice Hall Professional Technical Reference, VoiceXML: Introduction to Developing Speech Applications.

Bob, C., Edgar, 2012. C M P Books, the VoiceXML Handbook.

Mohammadiha, N., P. Smaragdis and A. Leijon, 2013. "Supervised and Unsupervised Speech

Enhancement Using Nonnegative Matrix Factorization," IEEE Trans. Audio, Speech, and Language Process, 21(10): 2140-2151.

HP VoiceXML developer resources, 2015. : http://www.hp.com/go/voicexml.

Oviatt, S.L., A. DeAngeli and K. Kuhn, 1997. "Integration and synchronization of input modes during multimodal human computer interaction," in Proc. of CHI97, New York, 415-422. ACM Press.

Rouillard, J., 2014. Web services and speech-based applications around VoiceXML, Journal of Networks, 2 (1): 27-35.

Voice Extensible Markup Language (VoiceXML) 3.0, W3C Working Draft 4., 2010. http://www.w3.org/TR/2010/WD-voicexml30-20100304/.

Hirsch, H.G. and H. Finster, 2013. "A new approach for the adaptation of HMMs to reverberation and background noise," Speech Commun., 50(3): 244-263.

Loizou, P.C., 2013. Speech enhancement: theory and practice. New York: CRC Press.