IMPLEMENTATION OF SPEECH RECOGNITION SYSTEM USING DSP PROCESSOR ADSP2181

[1]KALPANA JOSHI, [2]NILIMA KOLHARE & [3]V.M.PANDHARIPANDE

[1&2]Dept.of Electronics and Telecommunication Engg, Government College of Engg., Aurangabad,India
[3]Dr.BAMU, Aurangabad,India
[1]E-mail: er.joshipankaj@gmail.com

Abstract - While many Automatic Speech Recognition applications employ powerful computers to handle the complex recognition algorithms, there is a clear demand for effective solutions on embedded platforms. Digital Signal Processing (DSP) is one of the most commonly used hardware platform that provides good development flexibility and requires relatively short application development cycle.DSP techniques have been at the heart of progress in Speech Processing during the last 25years.Simultaneously speech processing has been an important catalyst for the development of DSP theory and practice. Today DSP methods are used in speech analysis, synthesis, coding, recognition, enhancement as well as voice modification, speaker recognition, language identification.Speech recognition is generally computationally-intensive task and includes many of digital signal processing algorithms. In real-time and real environment speech recognisers applications, it's often necessary to use embedded resource-limited hardware. Less memory, clock frequency, space and cost related to common architecture PC (x86), must be balanced by more effective computation.

Keywords*-Automatic speech recognition, Digital signal processing, Mel frequency cepstral coefficient*

I. INTRODUCTION

The objective of human speech is not merely to transfer words from one person to another, but rather to communicate, understanding a thought, concept or idea. The final product is not the words or phrases that are spoken and heard, but rather the information conveyed by them. In computer speech recognition, a person speaks into a microphone or telephone and the computer listens. Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation. So speech processing can be regarded as a special case of digital signals processing applied to speech signals. Automatic Speech Recognition technology has advanced rapidly in the past decades.

We know that the heart of every computer is a microprocessor, commonly it's Intel IA-32 (x86) or compatibile. When using an algorithm, e.g. predictive filter for the processing of speech signals, we assume that it is somehow calculated. We simply write a function in Matlab (or Octave): and we are not at all interested, how sophisticated calculation is made the optimization for the platform x86, and therefore that the calculation is as **fast as** possible. And even if it was not, we do not mind, because in Matlab on Intel x86 platform, we usually work out of real time, so we just wait a while to calculate. To keep at disposal with Quad-Core processor running on clock frequency of 3 GHz and with 8 GB of RAM is certainly convenient. There are, however, tasks, in which such an achievement we can't afford, for example for the following reasons: There is a requirement for a low power device, there is a requirement for small size and device portability, there is a small budget or need to minimize the cost of the device. Therefore, it is necessary to choose another hardware platform that is other than the x 86 microprocessor and apparently also other instruments for programs development. For real-time signal processing, it is obviously appropriate to choose a digital signal processor.

II. LITERATURE SURVEY

Every speech recognition application is designed to accomplish a specific task. Examples include: to recognize the digits zero through nine and the words "yes" and "no" over the telephone, to enable bedridden patients to control the positioning of their beds, or to implement a VAT (voice-activated typewriter). Once a task is defined, a speech recognizer is chosen or designed for the task. Recognizers fall into one of several categories depending upon whether the system must be "trained" for each individual speaker, whether it requires words to be spoken in isolation or can deal with continuous speech, whether its vocabulary contains a small or a large number of words, and whether or not it operates with input received by telephone. Speaker dependent systems are able to effectively recognize speech only for speakers who have been previously enrolled on the system. The aim of speaker independent systems is to remove this restraint and recognize the speech of any talker without prior enrolment. When a speech recognition systems requires words to be spoken individually, in isolation from other words, it is said to be an isolated word system and recognizes only discrete words and only when they are separated from their neighbours by distinct interword pauses. Continuous speech recognizers, on the other hand, allow a more fluent

form of talking. Large-vocabulary recognizers are defined to be those that have more than one thousand words in their vocabularies; the others are considered small-vocabulary systems. Finally, recognizers designed to perform with lower bandwidth waveforms as restricted by the telephone network are differentiated from those that require a broader bandwidth input.[4]Digital signal processors are special types of processors that are different from the general ones. Some of the DSP features are high speed DSP computations, specialised instruction set, high performance repetitive numeric calculations, fast and efficient memory accesses, special mechanism for real time I/O, low power consumption, low cost in comparison with GPP. The important DSP characteristics are data path and internal architecture, specialised in stuction set, external memory architecture, special addressing modes, specialised execution control, specialised peripherals for DSP.[6]At the beginning of each implementation process is an important decision: the choice of appropriate hardware platform on which a system of digital signal processing is operated. It is necessary to understand the hardware aspects in order to implement effective optimized algorithms. The above hardware aspects imply several criteria for choosing the appropriate platform: It is preferable to choose a signal processor than a processor for general use. It may not be decisive a processor frequency, but its effectivenes.DSP tasks require repetitive numeric calculations, alternation to numeric, high memory bandwidth sharing, real time processing. Processors must perform these tasks efficiently while minimizing cost, power consumption, memory use, development time. To properly select a suitable architecture for DSP and speech recognition systems, it is necessary to examine well the available supply and to become familiar with the hardware capabilities of the "candidates". In the decision it is necessary to take into account some basic features, in which processors from different manufacturers differ. Most DSPs use fixed-point arithmetic, because in real world signal processing the additional range provided by floating point is not needed, and there is a large speed benefit and cost benefit due to reduced hardware complexity. Floating point DSPs may be invaluable in applications where a dynamic range is required. To implement speech recognition different algorithms like Linear predictive coding, Mel Frequency Cepstral coefficient, HMM can be utilized. Here is an attempt to implement the speaker independant speech recognition system with small vocabulary and isolated words based on MFCC algorithm using a Fixed point DSP processor ADSP2181. Advantags of MFCC method are it is capable of capturing the phonetically important characteristic of speech, band limiting can easily be employed to make it suitable for telephone application.[5]

III. FUNCTIONAL DESCRIPTION

Speech Recognition is the process of converting an acoustic signal, captured by microphone or telephone to a set of words. The main requirement for speech recognition is extraction of voice features which makes distinguish different phonemes of language. The second part is matching of parameters for recognition purpose. Voice is a pressure wave, which is afterwards converted into numerical values in order to be digitally processed.Fig.1 gives the theme of the system.

A microphone allows the pressure sound p(t) to be converted into an electrical signal xc(t). Then a sampler at T time intervals (i.e. at a sampling frequency f=1/T) yields voltage values xc(nTc)=x(n), and finally an analog to digital (A/D) converter quantizes each x(n), n=0,1,….. N-1 into a specific number.[4]

This project is an implementation of Speech Recognition algorithm on a fixed point digital signal processor(DSP). Digital signal processors are designed to be especially efficient when executing algorithms common to real time signal processing like speech processing. DSPS are designed to efficiently handle high precision, high throughput arithmetic operations that must be executed in typical signal processing algorithm.The work is based on the detailed analysis of Mel Frequency Cepstrum Coefficient (MFCC) algorithm.Digital Signal Processing approaches the problem of speech recognition in two steps:Feature Extraction followed by Feature Matching.

Windowing-Traditional methods for spectral evaluation are reliable in the case of a stationary signal (i.e. a signal whose statistical characteristics are invariant with respect to time). For voice, this holds only within the short time intervals of articulator stability, during which a short time analysis can be performed by "windowing" a signal x1(n) into a succession of windowed sequences x1(n), 1 = 1, 2 ...T, called frames, which are then individually processed [3,4]

$$x1(n)=x1(n-t.Q), \quad 0\leq n \leq N,\ 1 \leq t \leq T \tag{1}$$

$$x1(n)=w(n)x1(n) \tag{2}$$

Where w(n) is the impulse response of the window. Each frame is shifted by a temporal length Q. If Q = N, frames do not temporally overlap while if Q < N, N -Q samples at the end of a frame x1(n)are duplicated at the beginning of the following frame x1(n). Fourier analysis is performed through the

Fourier transform that for discrete time signal x1(n) is :

$x1(e^{jw}) = \sum x1(n).\ e^{-jwn}$ for n=0 to N-1 (3)

In ASR, the most-used window is the Hamming window whose impulse response is a raised cosine impulse:

w (n)=0.54-0.46cos(2n/n-1) for n=0, 1…. N-1 (4)
= 0 otherwise

The side lobes of this window are much lower than the rectangular window (i.e. the leakage effect is decreased) although resolution is appreciably reduced. This is because the Hamming main lobe is wider.The Hamming Window is a good choice in speech recognition, because a high resolution is not required, considering that the next block in the feature extraction processing chain integrates all the closest frequency lines.

Once sampling frequency fc is fixed, the spectral resolution is inversely proportional to the sequence length N.A narrow-band spectrum is the one obtained when resolution is high, while a wide-band one is obtained when the resolution is low.

Moreover, larger windows (about 70 ms) have a higher frequency resolution. This allows identification of each single harmonic. However, in such a case, fast transitions in the spectrum (as for instance the pronunciation of stop consonants) are not detected. Narrow windows have been proposed to estimate the fast varying parameters of the vocal tract; while large windows are used to estimate the fundamental frequency. A 20-30 ms long window is generally a good compromise.

Spectral Analysis- The standard methods for spectral analysis rely on the Fourier transform of $x_1(n)$: $X_1(e^{jw})$, Computational complexity is greatly reduced if $X_1(e^{jw})$ is evaluated only for a discrete number of w values.The characteristics of the vocal tract may be estimated by the period gram of $X_1(n)$ that is simply the square magnitude of the DFT: $[X_1k)]^2$.

Filter bank processing -Especially analysis reveals those speech signal features which are mainly due to the shape of the vocal tract. Spectral features of speech are generally obtained as the exit of filter banks which properly integrate a spectrum at defined frequency ranges. A set of 24 band-pass filter is generally used since it simulates human ear processing.

Filters are usually no-uniformly spaced along the frequency axis. As a rule, the part of the spectrum which is below 1KHz is processed by more filter- banks since it contains more information on the vocal tract such as the first formant. The frequency response of the filter banks simulates the perceptual processing performed within the human ear there fore such filtering is called perceptual weighting. In ASR, the most widely used perceptual scale in recognition is the Mel scale whose filter-bank characteristics are outlined. The central frequency of each Mel filter bank is uniformly spaced before 1 KHz

Log Energy Computation-The previous procedure has the role of smoothing the spectrum, performing a processing that is similar to that executed by the human ear. The next step consists of computing the logarithm of the Square of magnitude of the coefficients. Because of the logarithm algebraic property which brings back the logarithm of a power to a multiplication by a scaling factor.

Mel frequency cepstrum coefficient computation (MFCC)- The final procedure for the Mel Frequency cepstrum coefficient computation (MFCC) consists of performing the inverse DFT on the logarithm of the magnitude of the filter bank output.

$y1^{(m)}(k) = \sum \log \{ | Y_t (m)| \}.\cos (k(m-1/2)\prod/m)$ k=0,…….L (5)

The procedure has great advantages. First note that since the log power spectrum is real and symmetric then the inverse DFT reduces to a Discrete Cosine Transform (DCT).

Technical Specification of the System:1)Processor ADSP2181-16-bit fixed point CORE operating at 5V,Internalmemory-DM-16KWords(16bits),PM-16KWords(24bits),External memory- DM-16K Words (16-Bits)PM-16k Words (24-bits),Clock-24.576MHz 2)UART-16C550 (19200 Baud rate used) 3) CODEC-HD44233 4)Power Supply-SMPS 5V,500mA with EMI filter.

General Specifications of the System**:**1)Speaker Independent Programmable Speech recognition system.2)Small vocabulary, isolated word system 3)System based on Mel Frequency Cepstral Coefficient algorithm.4)Implemented using fixed point DSP processor ADSP2181.

The proposed system is operated in two phases1.Training phase 2.Testing phase:In training phase the system is trained for a particular word. The speech samples are converted into MFCCs and are stored in database.In testing phase the word uttered is recognised by the system.The speech samples are converted into MFCCs and are compared with the database. Recognition is observed on the display by having the serial number of the word that is spoken.In the proposed system, to convert the speech signal to its electrical equivalent mono type of microphone is used. The output of a microphone is sampled with 8KHz sampling frequency. This sampling freq. is generated from the serial clock of the DSP processor. Serial clock freq. is generated from the clock out freq.

of the processor..To generate the sampling freq. SCLK is divided .CODEC IC HD44233 converts the input analog signal to its discrete time signal.The output is digital in magnitude .This discrete signal is received by the processor through UART ST 16C550.Serial port 1 of ADSP2181 is used to receive the data.For every utterance 2000 samples are taken. Samples are taken in Rx register. Everytime when Rx receives a sample, count is decremented and Rx is saved in DM. This takes place until the count 2000 comes to zero.Once all the samples are taken into data memory, processing of the samples starts by taking first 256 samples .One frame consists of 256 samples. The signal of 256 samples is windowed by Hamming window. The frames are overlapped by 100 samples. The overlapping is done to ensure that all the speech events influence the block calculation.The windowed frame is undergone 256 point FFT. Here DIT algorithm is used. The magnitudes are important as they carry the information related to speech. From the magnitude spectrum power spectrum is estimated.It is passed through a series of 20 mel spaced bandpass triangular filters. The lower cut off freq of the first triangular filter is kept at 100Hz and the upper cut off freq. of the last triangular filter is at 4KHz.Upto 1KHz the bandwidth of the filters is kept100Hz and the scale along freq. axis is linear but after that Log scale is used on the freq. axis. It is a MEL scale. This resemble human hearing system. By using a bank of triangular filters we get spectrum of a spectrum. The energy in a single bin is calculated.As per frame 20 filters are used we get 20 such numbers for a single frame. We get the spectrum mapped on the MEL scale. This is a MEL spectrum. Then the log power spectrum is calculated by taking the log of the sums. It is real.To convert that to again time domain Discrete Cosine Transform is carried out.It is a conversion of a MEL spectrum to MEL Cepstrum. For a single utterance 240 MFCCs are estimated.This whole procedure is executed for a single utterance in the training and testing phases. For recognition, the MFCCs of the utterances in the training phase are compared with the MFCCs of the utterance in the recognition mode The best match is the identified word.

IV. PERFORMANCE ANALYSIS

Wide range of problems in accuracy arise when common automatic speech recognition systems are tested under operating conditions different from those existing in the laboratory when the acoustic models are trained. Also systems designed to work with many speakers there is a worsening of recognition performance when some changes in environmental parameters.The sources of variation are categorized like noise, distortion, articulator effects and pronunciation. [1,2] The proposed system is tested in various conditions like1)Noise level-The system is tested in Silent room where the noise level is very low,Living room where there is considerable amount of noise like the noise of a telephone ring or that of a type-writer and Noisy room where the noise level is very high like the exhibition hall.The system is also tested with varied levels of noise in Training and Testing phase.2) Distance between a speaker and a microphone-The system is tested with variation in microphone distance in training and testing phase.3)Different Speakers- The system is trained for a single word by a particular speaker and tested it for the same word by different male and female speakers. Table I gives the accuracy of the implemented system under various conditions.

Table I System accuracy under various conditions of noise,speakers,microphone dist.

Training	Testing	Speaker	Mic. Distance	Accuracy %
L	L	Same	Less than 5cm	95
M	M			85
H	H			65
L	H			75
L	L	Different	Less than 5cm	85
L	L	Same	Less than 5cm	95
L	L		More than 10cm	50

L,M,H are the Low,Medium,High levels of Noise

On the basis of performance analysis carried out for the proposed system,it is concluded that the accuracy of the system is high when the system is trained and tested in the silent room,with microphone distance of less than 5cm.The accuracy is moderate with different speakers when the system is trained by the third person. It is further concluded that the accuracy of this system degrades because of the change in the noise level in the training phase and testing phase. Also the accuracy is much low when the distance between the speaker and the microphone is greater than 5 cm. The accuracy is affected when different speakers are using the system which is trained by some different speaker.

Thus it is concluded that wide range of problems in accuracy arise when Common Automatic Speech Recognition System are tested under operating conditions different from those existing in the laboratory when the acoustic models are trained. Also systems designed to work with many speakers there is a worsening of recognition performance when some changes in environmental parameters.

V. ACKNOWLEDGMENT

I feel great pleasure in submitting this paper "Implementation of Speech Recognition System using DSP Processor ADSP 2181**".** I express my sincere thanks to my dissertation guide Prof. N. R. Kolhare for guiding me at every step in making of this project. She motivated me and boosted my confidence and I must admit that the work would not have been completed without her guidance and encouragement.

REFERENCES

[1] Performance of Speech Recognition Devices Acoustics, Speech, and Signal Processing 1989 IEEE International conference.

[2] Performance of Isolated word Recognition System Acoustics, Speech and Signal Processing IEEE International Conference

[3] Speech processing edited by Chris Rowden , Mc Graw Hill Publications

[4] Speech Recognition by Hachette and Ricotti, Wiley Publication,PP 122-132

[5] ADSP 2181 Hardware Manual and Software Manual

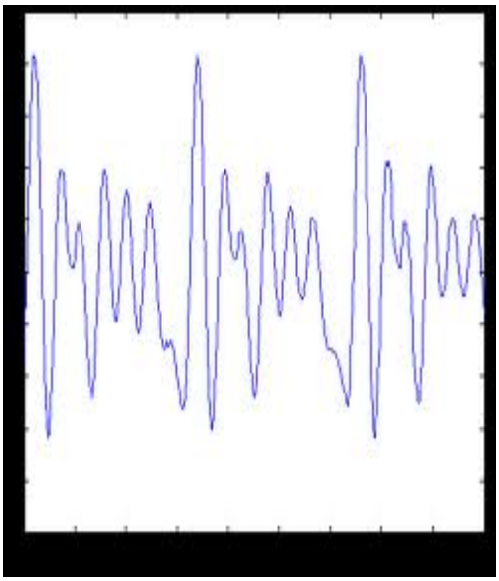

Fig.1.Speech signal in time domain

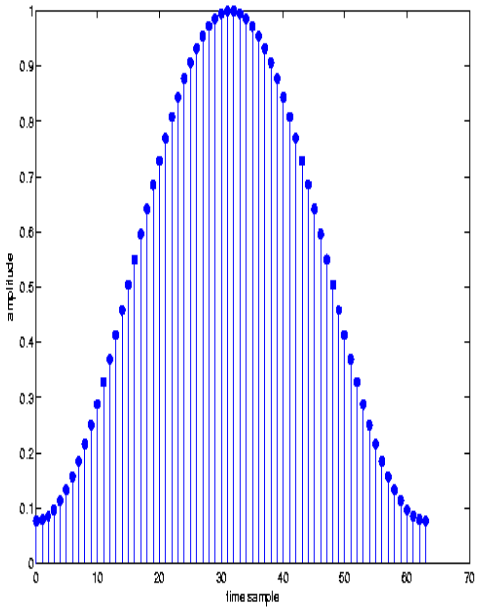

Fig. 2.Hamming window

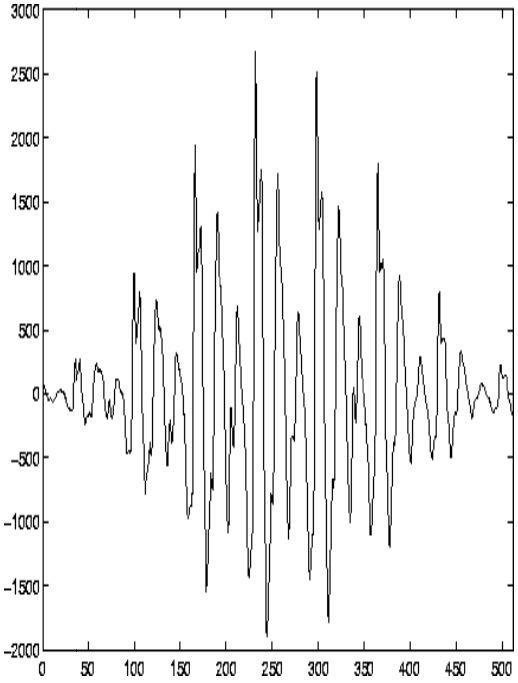

Fig.3.Windowed speech

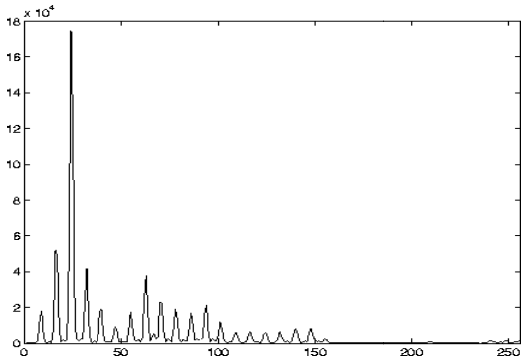

Fig. 4 : Magnitude spectrum

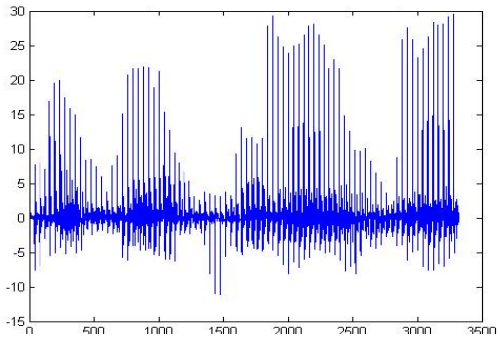

Fig.6.MFCC after DCT

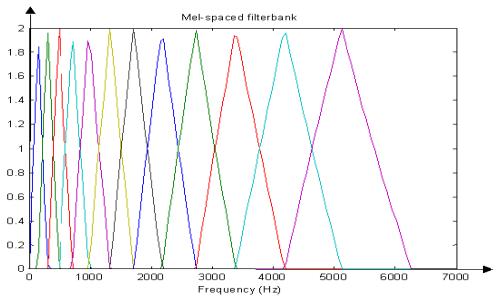

Fig.5.Mel spaced filter bank