Wang, Q., Dong, H., Zhu, G., Sun, H., Jaques, J., Piccirilli, A. B. & Dutta, N.K. 2006. All-optical logic OR gate using SOA and delayed interferometer. Optics Communication **260**: 81-86.

Vlachos, K., Pleros, N., Bintjas, C., Theophilopoulos, G. & Avramopoulos, H. 2003. Ultrafast time-domain technology and its application in all-optical signal processing. IEEE Journal of Lightwave Technology **21**: 1857-1868.

Zhang, M., Zhao, Y., Wang, L., Wang, J. & Ye, P. 2003. Design and analysis of all-optical XOR gate using SOA-based Mach-Zehnder interferometer. Optics Communication 223: 301-308.

Zhang, S. & Karim, M. A. 1998. One-step optical negabinary and modified signed-digit adder. Optics & Laser Technology **30**: 193-198.

Zohar, S. 1970. Negative radix conversion. IEEE Transactions on Computers **c-19**: 222-226.

Zoiros, K. E., Kalaitzi, A., & Koukourlis, C. S. 2010. Study on the cascadability of a SOA-assisted Sagnac switch pair. Optik - International Journal for Light and Electron Optics **121**: 1180-1193.

Submitted: 10/09/2013
Revised: 02/12/2013
Accepted: 30/12/2013

Journal of Engg. Research Vol. 2 - (2) June 2014 pp. 81 - 99

تطوير أنظمة التعرف على الكلام للصم في اللغات المحلية

*سي. جيالاكشمي، **في. كريشنامورتهي، ***أ. ريفاتهي

*قسم هندسة الكمبيوتر – كلية ترتشي- ترتشي- تاميلنادو- الهند

**قسم هندسة الكمبيوتر – كلية الهندسة – جامعة آنا- جيناي- الهند

***قسم هندسة الكمبيوتر- كلية ساراناثان الهندسية- ترتشي- تاميلنادو- الهند

الخلاصة

هذه الورقة تقدم أداء نظام التعرف على الكلام بالنسبة للأطفال الطبيعيين والأطفال الصم. مع أن التجويف الأنفي والسمعي للصم يبدو طبيعيا إلا أنهم لا يستطيعون النطق بسبب عدم قدرتهم على السماع، وذلك لأن فهم اللغة والنطق بها مرتبط بعمليات في الدماغ لذلك فإن الشخص المصاب بضرر في السمع أو بنشاط الدماغ بسبب في الولادة أو سبب طارئ يجد صعوبة في التحدث، هؤلاء يصنفون على أنهم صم أو ضعيفي السمع بناء على القدرة على السماع.

إن التشخيص المبكر للصم يساعد الشخص المعنى على إصدار أصوات بواسطة العلاج بالتحدث، إذا تم تشخيص الصمم في مرحلة متأخرة فإنه من الصعب جعل حديث المصاب مفهوما لدى الآخرين، لذلك أصبح من المهم إيجاد طريقة لجعل لغة هؤلاء المصابين مفهومة وخاصة باللهجات المحلية.

في هذه الورقة تم تطوير نظام للغة التاميل باستخدام خاصية MFCC في الإستخلاص في المقدمة وأداة HTK المتكاملة في نهاية النظام حيث تم تقييم هذا النظام بالمقارنة بين محادثة أشخاص طبيعيين وأشخاص صم وكانت نسبة التعرف على الكلمات 92.4 % بالنسبة للأشخاص الصم و98.4 % بالنسبة للأشخاص الطبيعيين، مع أنه من الصعب للمستمعين غير المعتادين على سماع حديث الأشخاص الصم إلا أن النظام يمكن استخدامه للتعرف على المحادثة بين الأشخاص الصم فيما بينهم.

Development of speech recognition system in native language for hearing impaired

C. JEYALAKSHMI*, V.KRISHNAMURTHI** AND A.REVATHI***

**Department of ECE, Trichy Engineering College, Trichy, Tamilnadu, India.*

***Former HOD, Department of ECE, College of Engineering, Anna university, Chennai, Tamilnadu, India.*

****Department of ECE, Saranathan College of Engineering, Trichy, Tamilnadu, India.*

lakshmi.jeya67@yahoo.com, profvkmurthi@yahoo.co.in, revathidhanabal@rediffmail.com

** Corresponding Author.*

ABSTRACT

This paper presents the performance of the speech recognition system with reference to children with normal hearing and children with hearing impairment. Though the nasal and oral cavities of the hearing impaired are perfect, they cannot produce sounds since they cannot hear anything. The reason is that the ability to understand language and speech production is coordinated by the brain. So a person with a problem in the ear or damage in brain activities due to an accident, stroke or birth defect may have problems in producing speech. They are classified as profoundly deaf and hard of hearing, based on the degree of hearing ability. Early detection of deafness would enable the hearing impaired to produce sounds by speech therapy. If deafness is detected at a later stage, it is difficult to make the speech of the hearing impaired understandable. So, it is necessary to develop the system for recognizing their speeches, especially in the native language. In this paper, a system is developed for Tamil language by using, Melfrequency cepstral coefficient feature extraction at the front end and Hidden Markov Model tool kit at the back end. System is evaluated and the comparison is done between the speeches of normal speakers and the hearing impaired. Recognition accuracy is 92.4% for hearing impaired speeches and 98.4% for normal speeches. Though it is difficult for the unfamiliar listeners to understand the hearing impaired speeches, this system can be utilized for recognizing the speeches of Hearing impaired by others.

Keywords: Deaf or hearing impaired (HI); hidden markov model tool kit (HTK); mel-frequency cepstral coefficients (MFCC); perceptual linear prediction (PLP); sub harmonic-to-harmonic ratio (SHR).

INTRODUCTION

Most people with any hearing impairment are adults of 65 years of age and above and children. Hearing impairment ranks third on the list of chronic health conditions of older adults, after arthritis and hypertension (Craig Newman *et al.*, 2004). As the population continues to age and to live longer, the number of people with hearing loss

will continue to rise. Hearing loss is the loss of ability to hear pure tones across the range of audio frequencies important for understanding speech.

There are approximately 70 million profoundly deaf, hard-of-hearing and speech-impaired people worldwide and there are so many causes for deafness. When we speak, we must coordinate many muscles from various body parts and systems, including the larynx, the teeth, lips, tongue, mouth and the respiratory system. Some people with speech problems such as articulation disorders may have hearing problems. Even mild hearing loss may have an impact on how a person reproduces the sounds they hear. People with problems during birth such as cleft palate can produce speech with the help of normal human beings. When a person has a cleft palate, there is a hole in the roof of the mouth, which affects the movement of air through the oral and nasal passages. The vocal tract begins at the opening of the vocal cords or glottis and ends at the lips. The vocal tract consists of the pharynx and the mouth or oral cavity. The nasal tract begins at the velum and ends at the nostrils. As the air is expelled from the lungs through trachea, the tensed vocal cords tend to vibrate due to air flow. The airflow is chopped into quasi-periodic pulses, which are then modulated in frequency in passing through the pharynx which is the throat cavity, the mouth cavity, and possibly the nasal cavity. Depending on the positions of the various articulators i.e. jaw, tongue, velum, lips, and mouth, different sounds are produced (Rabiner & Juang 1993).

A person having problem in any of the two cavities cannot speak due to the disability. Even though the two cavities are in proper condition, some of them are not able to produce sound. The reason behind this is that the hearing impaired does not know how to produce a particular sound. Figure 1 (Rabiner & Juang 1993) indicates the inner parts of the ear. Hearing begins when sound waves that travel through the air reach the outer ear or pinna and the sound waves then travel from the pinna through the ear canal to the middle ear, which includes the eardrum (a thin layer of tissue) and three tiny bones called ossicles. When the eardrum vibrates, the ossicles amplify these vibrations and carry them to the inner ear.

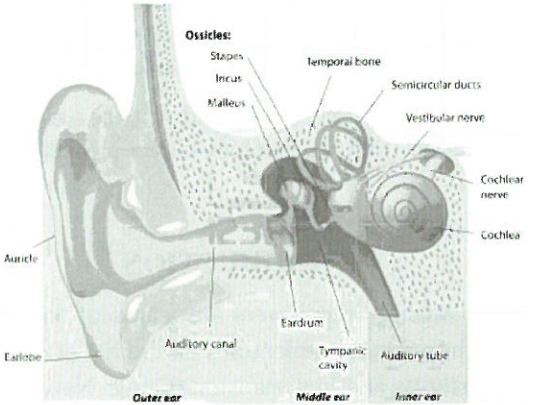

Fig.1. Hearing mechanism of human

The inner ear is made up of a snail-shaped chamber called the cochlea, which is filled with fluid and lined with thousands of tiny hair cells. When the vibrations move through this fluid, the tiny hair cells translate them into electrical nerve impulses and send them to the auditory nerve, which connects the inner ear to the brain. When these nerve impulses reach the brain, they are interpreted as sound. The cochlea is like a piano so that specific areas along the length of the cochlea pick up gradually higher pitches. Disease, damage, or deformity of the cochlear hair cells is a common cause of hearing impairment or deafness. Loudness of sounds in the range 0–70 dB is classified as normal hearing sensitivity, mild hearing loss, moderate hearing loss, moderately severe hearing loss respectively. 70–90 dB is classified as severe hearing loss and > 90 dB is classified as profoundly deaf. Mild to severe hearing loss is known as hard of hearing.

There are so many problems faced by the hearing impaired. The three main issues are education, employment and communication problems. Since their speech cannot be understood by others, they cannot behave socially like normal people in the society. On most of the occasions, even their parents and teachers cannot recognize their speech. With regard to employment, they may not get a job in the first trial. At the same time, other staff members will have to be trained to understand deaf culture, as well as encouraged to learn some basic signs, to help the hearing impaired employee in the workplace, so that they do not feel lonely at work. It is important to create a good atmosphere at the work place, so that they may feel quiet at home. Parents of the deaf children should be able to understand the needs of their children and admit them in a place congenial to them. Making the children go for education becomes a problem to their parents from kinder garden to college level. Hearing loss can seriously affect a person's quality of life. It limits the ability to perform certain functions, such as listening in noisy and crowded places. It restricts participation in social activities and can lead a person to withdraw from situations that require communication with others, including spouse, family members and friends. The psychosocial effects of hearing loss should not be underestimated. People with hearing loss often show signs of withdrawal, cognitive loss, depression, social isolation or psychosis. Even younger adults with mild hearing loss have reported a variety of psychosocial problems affecting life. Family members, friends and coworkers may experience frustration, impatience, anger, pity and guilt, when interacting with a person with hearing loss. These reactions are sources of stress in the relationship. It is difficult for the doctors to diagnose the health problems of the hearing impaired people quickly.

Many hearing-impaired teens read lips and use American Sign Language, cued speech, or other sign languages, and in some cases an interpreter may be available to translate spoken content in the classroom. Some teens may attend a separate school or special classes offered within a public school. Universities are exclusively available to offer education at the college level to the hearing impaired people. At home, devices

such as closed-captioned TVs, lights that flash when the doorbell or phone rings and telephones with digital readout screens are some of the telecommunication devices to assist the hearing impaired (Valerie Henderson-Summet *et al.*, 2007). Hearing aids and lip-reading are more effective in face-to-face communication (Dr.Colin Brooks 2000).

Hearing aids come in various forms that are fitted inside or behind the ear and make sound louder. They are adjusted by the audiologist so that the sound coming in is amplified enough to enable the person with a hearing impairment to hear it clearly. Sometimes, the hearing loss is so severe that the most powerful hearing aids cannot amplify the sound enough. In those cases, a cochlear implant may be recommended. Cochlear implants are surgically implanted devices that bypass the damaged inner ear and send signals directly to the auditory nerve. A small microphone behind the ear picks up sound waves and sends them to a receiver that has been placed under the scalp. This receiver then transmits impulses directly to the auditory nerve. These signals are perceived as sound and allow the person to hear.

There are two major drawbacks in the development of speech processing aids for the hearing impaired. The first is a lack of basic knowledge as to how speech is acquired, produced and perceived. Thus, even with the sophisticated electronic instrumentation of today, perceptual aid substantially superior to a good quality conventional hearing aid is not available (Harry Levitt 1973). But all the devices are designed for enabling the hearing impaired to understand the normal speech. In this scenario, if a system is developed for recognizing the hearing impaired speech in real time, the communication difficulties between the hearing impaired and the normal will be minimized. In this paper, recognition of hearing impaired speech is carried out by the use of Hidden Markov Model (Rabiner & Juang 1993) with MFCC features. The reasons behind the popularity of the method are the inherent statistical framework, the ease and availability of training algorithms for estimating the parameters of the models from finite training sets of speech data and the flexibility of the resulting recognition system in which one can easily change the size, type, or architecture of the models to suit particular word, sounds (Juang & Rabiner 1991).

DEAF SPEECH DATA BASE GENERATION

The performance of speech recognition is usually evaluated in terms of accuracy and speed. In general, speech recognition is a very complex problem since the same words spoken by the same person will not be equal in different time periods. Vocalizations with regard to hearing impaired children vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume and speed. We can informally check whether the children are hard of hearing or profoundly deaf. As per the observations of the speech therapist, normal children should be able to hear the sounds அ, இ, உ, ஸ, ழ within a radius of 5 metres, when the teachers are speaking at normal speaking rate

and intensity. If not, they are hard of hearing. For this study, we have taken speech samples of 10 deaf children in the age group of 10-14 years from Maharishi Vidhya Mandir service centre for the hearing impaired, Tiruchirappalli. The deaf children are able to follow only one language, mostly their native language because they cannot follow the different facial expressions and throat vibrations. Like a Speech therapist, we have used the tactile method to record the speeches of hearing impaired, i.e. we made them touch our throat by their hands to feel the vibration when pronouncing a particular word and asked them to see our facial expressions and listen to the sound. They had to visually see the words also, by writing them on the board. Among the ten children seven were profoundly deaf and three were hard of hearing. These ten children were admitted in the HI school after verifying the clinical report given by the audiologist. Based on the clinical report, therapist used to follow the training method for the children appropriately. According to the clinical report of the children, they have been classified as profoundly deaf and hard of hearing. In this work we have taken isolated digits from poojam to ombadu (0 to 9).

Children have spoken the Tamil digits 20 times and their corresponding speeches were recorded using high quality microphone. The specifications were, Frequency Response: 100-15,000Hz, maximum sound pressure: approximately 110dB S.P.L, sensitivity: 100 dB (at1 KHz, 0.5Vrms), directivity: Omni-directional. 15 speeches of 10 speakers were considered for training and remaining 5 speeches of 10 speakers were considered for testing. Initially the words were recorded from each HI child using wave surfer and labels are given accordingly by manually hearing the sound. The labels were be assigned carefully to the words uttered by the children. Silence was introduced artificially for pause between the words. Then the label files for isolated words were created and features were extracted.

ANALYSIS OF SPEECH BASED ON PITCH AND FORMANTS WITH 10 NORMAL AND 10 HEARING IMPAIRED CHILDREN

One of the problems encountered in analyzing the speech of the HI was the large variability between speakers. Differences between speeches of HI speakers were substantially greater than differences between that of normal speakers. Hence, more data was needed to distinguish between talkers (Harry Levitt 1971).The language skills of these children were on the average, severely retarded, their speech production and their speech reception were of limited use, their vocabulary, grammar, and reading show great deficiencies compared to normal children (Pickett 1969).

In general, HI speakers cannot have clarity for the mid frequencies from 250 to 2000Hz which is called the speech area, since all voiced phonemes will come under this frequency. Figure 2a and 2b show the speech waveform of the normal and HI speaker when uttering the word *poojam*.

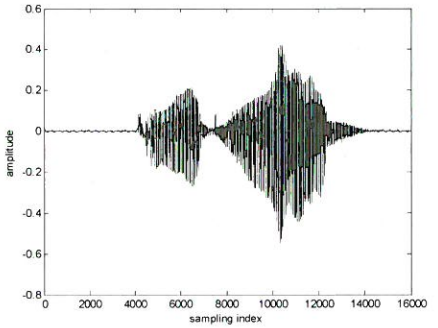

Fig.2a. Speech waveform of a normal speaker for the word *poojam*.

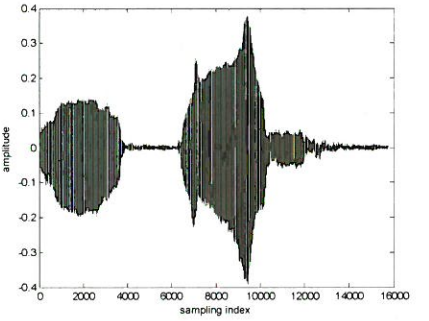

Fig.2b. Speech waveform of a HI for the word *poojam*.

From the Figure 2a and 2b it is clear that the voiced, unvoiced and silence regions of the two speakers are different for the same word. After the initial silence, the normal speaker suddenly utter the word but in the case of the hearing impaired speaker, silence was produced in between the voiced regions. The pronunciation of the word "*poojam*" taken for analysis is relatively difficult even for the normal speakers. Since HI children are speaking after observing the facial expressions and feeling the throat vibrations of persons, they are finding it difficult to pronounce these difficult words. So, they introduced the pauses between the phonemes of these difficult words naturally.

An alternative way of characterizing the speech signal and representing the information associated with the sound is through spectral representation. Mostly to represent this type, sound spectrogram in which a three-dimensional representation of the speech intensity, in different frequency bands, over time is depicted. Figure 3a and 3b indicate the spectrogram of normal and hearing impaired speaker. The spectral intensity at each point in time is indicated by the intensity (darkness) of the plot. The voiced sections are resolved and are seen as vertical striations in the spectrogram.

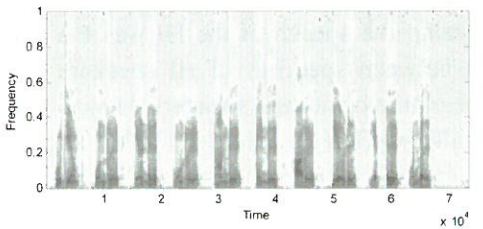

Fig.3a. Spectrogram of a normal speaker from poojam to *ombadu*.

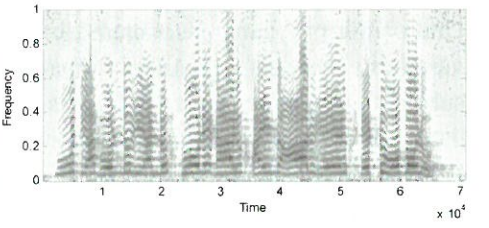

Fig.3b. Spectrogram of a HI speaker from poojam to *ombadu*.

The Figure 3a and 3b clearly show how the words are clearly identified for a normal speaker; whereas, for a HI speaker it is totally overlapped and it is difficult to

segment each word. Due to this reason, recognition of the words is difficult in the case of the hearing impaired speaker. We can also analyze the speech signal using short time Fourier transform which is also called windowed Fourier transform. Figure 4a and 4b illustrate the spectrum plot of normal and hearing impaired speakers.

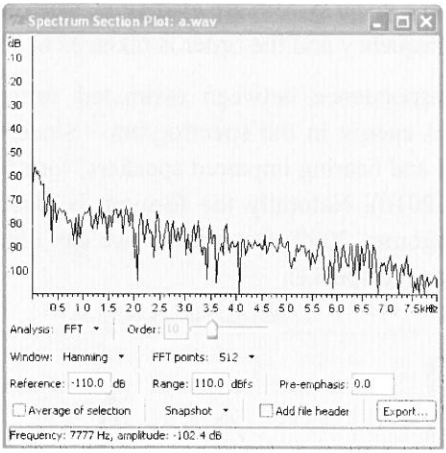

Fig.4a. Spectrum plot of a normal speaker for the utterance *poojam*.

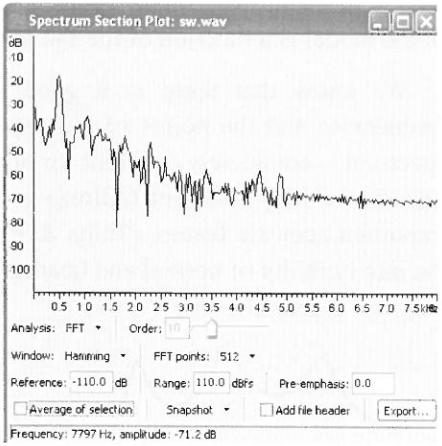

Fig.4b. Spectrum plot of a HI speaker for the utterance *poojam*.

Figure 4a and 4b clearly indicate variations in frequency distribution for a normal and hearing impaired speaker. All the frequency components of the normal speaker are clearly seen and the magnitude response at the highest frequency is -102.4db. But in the case of hearing impaired speaker there is no voiced sounds for the frequencies from 5kHz to 7.5kHz and it looks like unvoiced sounds. The amplitude also reduces from -110 to -71.2 db. There is little low frequency energy and it is high at 1.5 and 2.25kHz and it ceases to be minimum at high frequency in the case of HI. Whereas, in the case of normal speaker the energy is high at 2, 3.5, 4kHz so it has high second and third order formants.

Another way of representing the time-varying signal characteristics of speech is based on the model of speech production. Because the human vocal tract is essentially a tube, the transfer of energy from the excitation source to the output can be described in terms of natural frequencies or resonances of the tube. Such resonances are called formants of the speech, and they represent the frequencies that pass the most acoustic energy from the source to the output. Typically there are about three resonances of significance for a human vocal tract, below about 3500Hz. But the major problem is, difficulty in estimating the formants for low level voiced sound and for unvoiced or silence regions. Figure 5a and 5b show the formant frequencies for the utterance *poojam*. The power spectral density is calculated using covariance algorithm for normal and hearing impaired. The covariance method estimates the spectral density

by fitting an Auto Regressive prediction model of a given order to the signal. A model which depends only on the previous outputs of the system is called an Auto Regressive (AR) model, which is the simplest model for the vocal tract. We are using the AR model to determine the characteristics of the vocal system and from this system model, formants or resonant frequencies of the vocal system are evaluated. The order of the model is a function of the sampling frequency and the order is taken as 6.

We know that there is a good correspondence between estimated formant frequencies and the points of high spectral energy in the spectrogram. Since the spectrum is completely different for normal and hearing impaired speakers, formants are also entirely different (Alireza *et al.*, 2010). Naturally the formant is also an important acoustic feature (Tolba & El Torgoman 2009) from which we can obtain the intelligibility of normal and hearing impaired speaker.

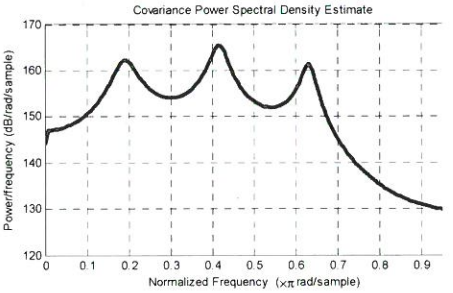

Fig.5a. Formant frequency representation of the utterance *poojam* for a normal speaker.

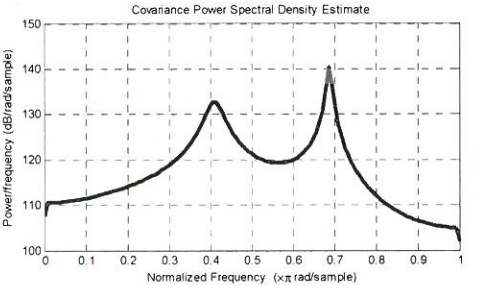

Fig.5b. Formant frequency representation of the utterance *poojam* for a HI speaker.

It is often measured as an amplitude peak in the frequency spectrum of the sound, using a spectrogram. Formant values can vary widely from person to person, and all voiced phonemes have formants even if they are not as easy to recognize. Voiceless sounds are not usually having formants; instead, the plosives should be visualized as a great burst. Formant frequencies are observed from the above figures with reference to maximum signal frequency of 8 kHz and it is given in Table 1.

Table 1. Formant Frequency of Hearing impaired and Normal speaker.

Formant	Frequency in Hz	
	Normal Speaker	Hearing impaired
1 st	1520	3280
2 nd	3280	5520
3 rd	5040	-----

From Table 1 it is understood that the third formant frequency is not clearly seen for the speech uttered by the hearing impaired speaker. Also, the 1st and 2nd formants are higher than normal speaker.

We can also do the analysis based on fundamental frequency of the vocal cords which is called pitch. The fundamental frequency F_0 of voiced sounds is determined physiologically by the vocal fold vibration rate. Control of F_0 is used to communicate prosodic features of speech such as stressing and intonation. Production of prosodic features is an essential part of the normal human communication process. Hence, it is essential that F_0 be measured accurately in assessing and in rehabilitating people with hearing impaired speech (Prashant S. Dikshit, *et al*, 1993). Several investigators have reported the problems of profoundly deaf speakers with pitch control. The characteristic difficulties include abnormally high average pitch and unnatural intonation patterns. These problems make hearing impaired speech sound unnatural and even unintelligible. So poor pitch control decreases the intelligibility of hearing impaired speech (Thomas & Francis 1972). By studying the individual subjects, we could not find evidence for a clear distinction between the HI and normal hearing subjects by means of Fo. It can be concluded that the HI subjects showed more variation in their phonation than the normal (Chris *et al*., 1996). The Figure 6a and 6b illustrate the pitch contour along with the spectrogram of speech uttered by normal and hearing impaired, and from the figure it is evident that, if the spectral energy is high the pitch frequency is also high for both the speakers.

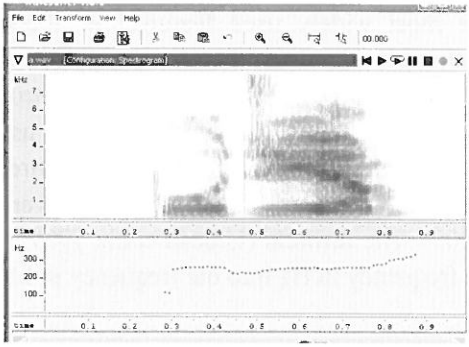

Fig.6a. Pitch contour of normal speaker for the utterance *poojam*.

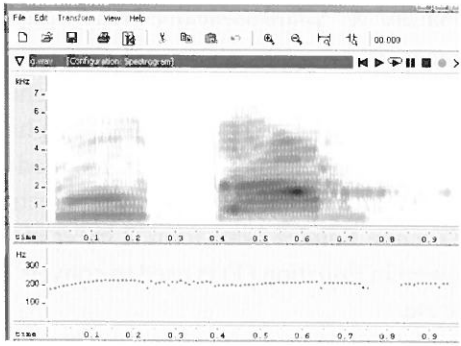

Fig.6b. Pitch contour of HI speaker for the utterance *poojam*.

The pitch detection algorithm estimates pitch through spectrum shifting on logarithmic frequency scale using Subharmonic-to-Harmonic Ratio (SHR) principle. This algorithm performs considerably better than other pitch detection algorithms and can also be applied to voice quality analysis (Xuejing 2002). First the isolated words uttered by the normal speakers are sampled at 16KHz with 16-bit resolution. Here, the frame length is taken as 40ms with 20ms overlap, 50Hz-200Hz for Fo range

and upper bound of the frequencies that are used for estimating pitch is taken as 1250Hz with SHRP threshold is taken as 0.2. Then pitch values are estimated using SHR algorithm (Xuejing 2002). Similarly pitch extraction is done using SHR for the isolated words of deaf and hard of hearing children and it is depicted in Figure 6c.

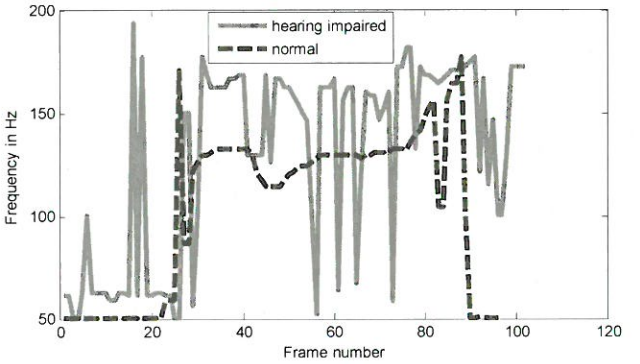

Fig.6c. Pitch contour of normal speaker and HI for the utterance *poojam*.

MFCC EXTRACTION

The important part of any speech recognition system is to extract features from the speech which should not change for the same words and should not change with time or be unaffected by the speaker's health. Mel Frequency Cepstral Coefficients (Murty & Yegnanarayana 2006), are the most widely used features for speech recognition applications. MFCCs are estimated based on human perception of frequencies. Psychophysical studies show that human perception of sound spectrum does not follow a linear scale (Shaughnessy 2003). Mel scale cepstral analysis uses cepstral smoothing to smooth the modified power spectrum. This is done by direct transformation of the log power spectrum to the cepstral domain using an inverse Discrete Fourier Transform (Lim *et al.*, 2009). The formula (Umesh et al., 1999) as given in Equation (1) is used to convert the frequency in Hz into the frequency in Mel scale.

$$mel(f) = 2595*\log(1+f/700) \tag{1}$$

The modified power spectrum thus consists of power coefficients Sk , and the mel frequency cepstrum is calculated as given in Equation (2).

$$c_n = \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\frac{\pi}{K}] \tag{2}$$

where S_k is the output power of the k^{th} filter of the filter bank, and n is from 1 to 13. Spectral transition plays an important role in human speech perception. So it is desired

to add information about time difference or delta coefficients and also acceleration coefficients (second derivative) with the feature vectors to get more relevant information about the speech.

Hidden Markov Model Tool Kit (HTK) is a free software by Cambridge university for researchers. HCopy is a built in function used to find the features from the source speech file. This program will copy one or more data files to a designated output file, optionally converting the data into a parameterized form. The steps to produce the features are shown in Figure 7. While the source files can be in any supported format, the output format is always HTK (Steve *et al.*, 2002) and each source data file has an associated label file, and a target label file is created. The name of the target label file is the root name of the target data file with the extension .lab. This new label file will contain the appropriately concatenated labels to correspond with the target data file and all start and end boundaries are recalculated if necessary.

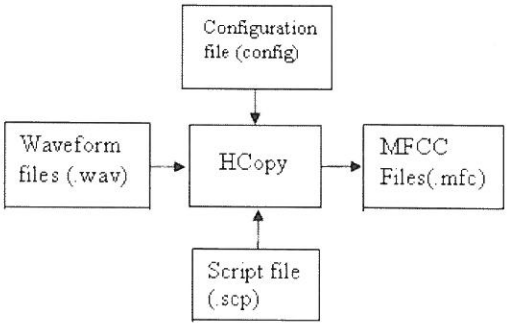

Fig.7. Steps to produce MFCC parameter vectors

The function HList will list the contents of one or more data sources in any HTK supported format. It can thus read data from a waveform file, from a parameter file and directly from an audio source. HList is used for examining the contents of speech data files and it is used for checking that input conversions are being performed properly. A configuration designed for a recognition system can be used with HList to make sure that the translation from the source data into the required observation structure is exactly as intended.

Initially, the Tamil digits from *poojam to ombadu* are recorded from 10 deaf speakers and label files are created from input wave files. MFCC coefficients 13, Delta coefficients 13, Acceleration coefficients 13 are generated from these label files, i.e. the modeled acoustic feature vector was composed of a 12 dimension base feature concatenated with a logarithmic energy coefficient. This was then concatenated with delta and acceleration coefficients to produce a final 39-dimensional feature vector.

DEVELOPMENT OF TRAINING MODELS

Hidden Markov models are widely used for automatic speech recognition (Rabiner 1989) because they have a powerful algorithm in estimating the model parameters and achieve a high performance. Once a structure of the model is given, the model parameters are obtained automatically by applying training data. The types of signal models are deterministic and statistical models. In statistical model (Picone 1993) one tries to characterize the statistical properties of the signal. In HMM for each state, there is an output probability distribution of an acoustic vector, and each iteration is associated with a state-transition probability. These probabilities are called the model parameters and can be estimated effectively by using Baum Welch algorithm (Juang & Rabiner 1991). A HMM structure (Pujol, *et al.*, 2005) can be expressed in a matrix form $A=[a_{ij}]$ when $a_{ij}=1$, there exists a transition from state i to state j and when $a_{ij}=0$, the transition does not exist. An HMM is a finite-state machine that changes state once every time unit. For every time unit t, a state j is entered, acoustic speech vector y_t is generated with probability density $b_j(y_t)$. The transition from state i to state j is governed by the probability a_{ij}. If we take the word *ondru* it has 4 phonemes (sounds oh,in,ir,uw) and it has 4 states. So the size of the transition matrix to be chosen as 4x4. Here we have used HTK toolkit for building Hidden Markov Models (HMMs). However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers.

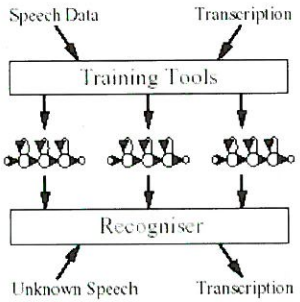

Fig.8. HTK structure

Figure 8 shows the structure of HTK in which the input training speech and its corresponding transcription are given to the training block. Here the models are created for each word. During testing whenever test speech is given, its corresponding transcription is obtained. There are two major processing stages involved. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools. In HTK, HInit is used to provide initial estimates for the parameters of a single HMM using a set of observation sequences. It works by repeatedly using Viterbi alignment to segment the training observations, and then re-computing the parameters by pooling the vectors in each segment. For

mixture Gaussians, each vector in each segment is aligned with the component with the highest likelihood. HRest is intended to operate on HMMs with initial parameter values estimated by HInit. HRest performs basic Baum-Welch re-estimation of the parameters of a single HMM using a set of observation sequences.

HRest can be used for normal isolated word training in which the observation sequences are realisations of the corresponding vocabulary word. This causes the parameters of the given HMM to be re-estimated repeatedly using the data in train Files until either a maximum iteration limit is reached or the re-estimation converges. HMM model definition read from a file is called "hmm" and the list of train files is stored in a script file. HMM definition consists of the number of states in the model inclusive of the non emitting entry and exit states. The information for each state is then given in turn, followed by the parameters of the transition matrix and the model duration parameters. HInit supports multiple mixtures, multiple streams, parameter tying within a single model, full or diagonal covariance matrices, tied-mixture models and discrete models. The output of HInit is typically input to HRest.

From the feature files, models are generated for Tamil digits from p*oojam to ombadu*. All the hearing impaired speakers uttered each digit twenty times and totally 1500 utterances were taken from them for training and 500 utterances are taken for testing. In general the number of states should be equal to number of phonemes in a word. In our work, number of mixtures are taken as 7 and number of states are taken between 2 and 6. Initially the proto type models are generated for 2s-7m (2states & 7mixture), 3s-7m, 4s-7m, 5s-7m, 6s-7m and then models are re estimated. Proto type models are initialized with number of mixtures and number of states and models are trained for the input training speeches and the model parameters such as mean vector and covariance matrix are re-estimated using Baum Welch re-estimation algorithm(Juang & Rabiner 1991).

Now the training models are developed for all the digits from *poojam to ombadu* for normal speaker. Initially the Tamil digits from *poojam to ombadu* for 10 normal speakers are recorded and models are generated for these words. The number, of mixture is taken as 7 and the states are taken as 2 to 6.

In the testing phase the test speech .wav signals are converted into series of acoustical vectors, i.e. feature vectors using HCopy tool, in the same way as in training. These vectors are compared with the models already developed for each word from *poojam* to *ombadu* using HVite tool. It compares the speech file against the HMM networks and produces a transcription for it.

EXPERIMENTAL EVALUATION

Digit recognition in Tamil from poojam to ombadu has been done using HTK tool kit and comparison was done for normal speakers and HI. Evaluation was done

by considering 10 speakers with 20 utterances for each digit and no. of utterances considered for training and testing each digit are 150 and 50 respectively. Recognition was done for 3 cases such as Normal Vs Normal, HI Vs HI and HI Vs Normal. For Normal Vs Normal, training and testing were done using the speeches of normal speaker. For HI Vs HI, training and testing are done using the speeches of HI. For HI Vs normal, training was done with the speeches of normal speakers and testing is done with speeches of HI. The results are discussed in the next section.

Hearing Impaired Vs Hearing Impaired

For speaker independent hearing impaired speech, out of 500 utterances only 462 were correctly recognized. It is also shown in Figure 9.

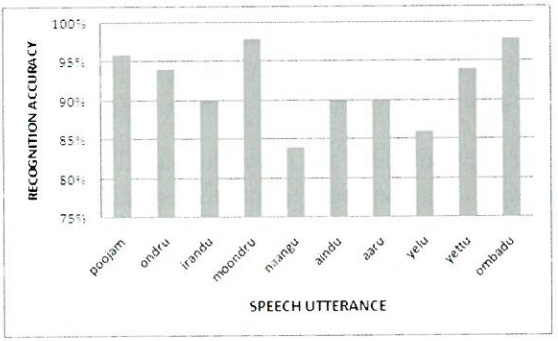

Fig.9. Recognition performance of HI speaker

The reason behind the relatively less performance for the hearing impaired case is probably the lack of clear pronunciation of words by them. But some speakers performed well like normal speakers. Figure 9 clearly shows that the accuracy is more than 75% for all the words. The reason is that, their impairment is identified earlier and their individual intelligence in grasping what is being said has improved during speech training. In addition to this, cooperation from the parents in continuing the training, for their children at home is also an important factor. So, the student can learn to speak and pronounce the words, somehow understandable by others. This is the reason why the system gives a good performance for their speech utterances. The overall accuracy was 92.4%. So this system, developed to understand the speeches of HI gives good result which is the main research of this paper.

Hearing Impaired Vs Normal Speaker

In order to compare the performance of recognition with the normal speaker, first the training was done for the normal speech and then it was tested with HI speech.

Testing was done by applying test speeches uttered by HI people to the training models developed for normal speakers. Recognition performance is indicated in the following Figure 10.

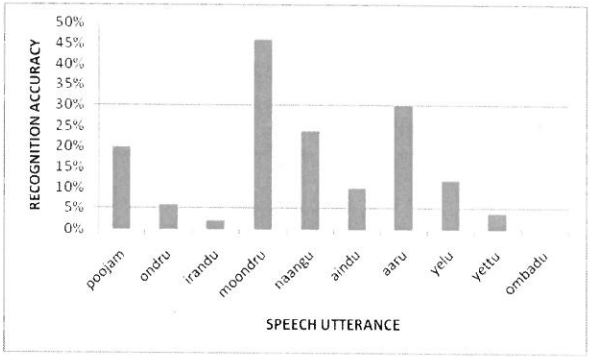

Fig.10. Recognition performance of HI speaker with normal speech models.

Normal training and testing with deaf is concerned, out of 500 utterances only 77 were correctly recognized. This shows that training models developed for normal speakers cannot be used for testing the speeches of hearing impaired. The reason is vocalizations with regard to hearing impaired children vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume and speed. Due to this high variation between the speeches pronounced by the normal speaker and HI, the overall accuracy is less.

Normal Vs Normal speaker

In the case of normal speakers, out of 500 utterances 492 were correctly recognized and overall accuracy obtained was 98.4%. It is shown in Figure 11.

As a whole by comparing the recognition performance of normal speaker and HI, even though the speech utterances do not have clarity, the system produced good results for HI speech. The overall recognition accuracy is depicted in Table 2.

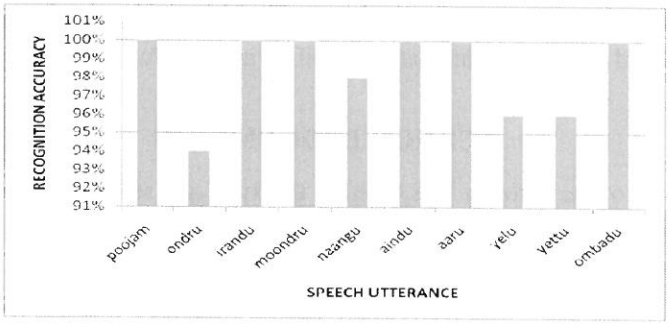

Fig.11. Recognition performance of normal speaker.

Table 2. Recognition accuracy

Input condition	Average recognition accuracy
Normal Vs Normal	98.4%
HI Vs HI	92.4%
HI Vs Normal	15.4%

CONCLUSION

In this paper we have presented speech recognition system for the hearing impaired and compared their performance with the normal speaker. Speeches of normal and HI speakers are analysed in terms of parameters such as, pitch, formants and spectrogram. The work required the creation of data base to train and test the system. The database was created by considering 10 speakers and among them seven were profoundly deaf and three were hard of hearing. Initially they were trained to utter the words with the help of their teachers and speeches were recorded for using them in our work. Since standard database (like TIMIT data base for normal speakers) was not available for HI, the speeches were collected from them by using tactile method. First we had to write the words in the blackboard, and then they were asked to observe our facial expressions and feel the throat vibrations by touching our throat with their hands, when we pronounced a particular word. This process was continued till they spoke the word to be understandable. Likewise each word was collected 20 times from individual speaker.

Though their pronunciation was very poor and difficult to understand, labels were given to the input speech by manually hearing the sound. With MFCC features, this word based speech recognition system using HTK tool kit for recognition produced 98.4% accuracy for the normal speaker and 92.4 % for the HI. The primary contribution of this paper is, with the available method we can achieve good performance for the proposed recognition system. Accuracy can be improved by taking more number of samples for training. The results of our work also demonstrate that speech analysis of hearing impaired in different language is still an area for further investigation.

ACKNOWLEDGEMENTS

Our thanks to the Director Mrs.Geetha, the staff and students of Maharisi vidya mandir service centre for the hearing impaired, Tiruchirappalli, who helped us in creating the database for the hearing impaired.

REFERENCES

Alireza A. Dibazar, Hyung O, Park, and Theodore W. Berger. 2010. Nonlinear dynamic modeling of impaired voice, 32nd Annual International Conference of the IEEE EMBS, 2770 – 2773.

B.H. Juang, L.R.Rabiner,1991. Hidden Markov Models for speech recognition, Techno metrics, 251-272.

Chris J. Clement, Florien J. Koopmans-van Beinum & Louis C. W. Pols, 1996. Acoustical characteristics of sound production of deaf and normally hearing infants, Fourth international conference on spoken language, 1549-1552.

Craig w. newman, Sharon a. sandridge, 2004. Hearing loss is often undiscovered but screening is easy, Cleveland clinic Journal of Medicine, 71:3.

Dr. Colin Brooks, 2000. Speech to text system for deaf, deafened and hard of hearing people, The IEEE Seminar on Speech and Language Processing for Disabled and Elderly People, 5:1-4.

Harry Levitt, Member, 1971. Acoustic Analysis of Deaf Speech Using Digital Processing Techniques, IEEE Fall Electronics Conference, Chicago.

Harry Levitt, 1973. Speech Processing Aids for the Deaf: an overview, IEEE Transactions on audio and Electro acoustics, 21(3): 269-273.

J. M. Pickett, 1969. Some Applications of Speech Analysis to Communication Aids for the Deaf, IEEE Transactions on Audio and Electro acoustics, 17(4):283-289.

Lim Sin Chee, Ooi Chia Ai, M.Hariharan & Sazali Yaacob, 2009. MFCC based recognition of repetitions and prolongations in stuttered speech using K-NN and LDA, Proceedings of the IEEE student Conference on Research and Developement, 146-149.

L.R.Rabiner, 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 77(2): 257-286.

Murty, K.S.R., & Yegnanarayana. B,2006. Combining evidence from residual phase and MFCC features for speaker recognition, IEEE Signal Processing Letters, 13(1): 52–55.

Picone J, 1993. Signal modelling techniques in speech recognition, Proceedings of the IEEE, 81(9): 1215–1247.

Prashant S. Dikshit, Edward L. Goshorn, and Ronald L. Seaman, 1993. Differences in fundamental frequency of deaf speech using FFT and Electroglottograph, Proceedings of the Twelfth Southern IEEE Biomedical Engineering Conference, 111 - 113.

Pujol. P, Pol. S, Nadeu. C, Hagen. A, Bourlard. H, 2005. Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system, IEEE Transactions on Speech and Audio processing, 13(1): 14-22.

Rabiner, L. R., and B. H. Juang, 1993. Fundamentals of Speech Recognition, Prentice Hall, New Jersey.

Shaughnessy D.O., 2003. Speech communication: human and machine, Addison-Wesley.

Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, 2001. The HTK Book, Cambridge University Engineering department.

Tolba.H, El Torgoman.A.S, 2009. Towards the improvement of automatic recognition of dysarthric speech, IEEE International Conference on Computer Science and Information Technology, 277-281.

Thomas R. Willemine, Francis F. Lee, Fellow IEEE, 1972. Tactile Pitch Displays for the Deaf, IEEE Transaction on Audio and Electro acoustics, 20(1):9-16.

Umesh S. & Cohen L. and Nelson D, 1999. Fitting the Mel scale, Proceedings of IEEE ICASSP, 217–220.

Valerie Henderson-Summet1, Rebecca E. Grinter1, Jennie Carroll & Thad Starner, 2007. Electronic Communication: Themes from a Case Study of the Deaf Community, IFIP International Federation for Information Processing, 347–360.

Xuejing, Sun, 2002. Pitch determination and voice quality analysis using sub harmonic-to-harmonic ratio, Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing, 333 - 336.

Submitted: 27/05/2013
Revised: 13/01/2014
Accepted: 08/02/2014