AUTOMATED SPEECH RECOGNITION APPROACH TO CONTINUOUS CUE-SYMBOLS GENERATION

Ibrahim Patel[1], Dr. Y. Srinivasa Rao[2]

[1] Assoc. Prof., Department of Electronic & Communication, [2] Assoc. Prof., Department of Instrument Technology
[1,2]Padmasri Dr.B.V.Raju Institute of Technology, Narsapur, Medak Dist., A.P. India
Andhra University, Visakhapatnam, A. P. India
[1]ptlibrahim@gmail.com [2]srinniwasarau@gmail.com

ABSTRACT

The work described in this paper is with an aim of developing a system to aid deaf-dumb people which translates the voice into sign language. This system translates speech signal to American Sign Language. Words that correspond to signs from the American sign language dictionary calls a prerecorded American sign language (ASL) showing the sign that is played on the monitor of a portable computer. If the word does not have a corresponding sign in the sign language dictionary, it is finger spelled. This is done in real life by deaf for words that do not have specific signs like for proper names. Hidden Markov Model (HMM) is used for recognition of speech signal from the user and translated to cue symbols for vocally disabled people. The proposed task is a complementary work to the ongoing research work for recognizing the finger movement of a vocally disabled person, to speech signal called "Boltay Haath". The proposed AISR system integrated with Boltay Haath system could eliminate the communication gap between the common man and vocally disabled people and extend in both ways.

Index Terms— Speech recognition, HMM, vocally disabled, two-way communication, speech-signal, American Sign Language (ASL),

I. INTRODUCTION

Humans know each other by conveying their ideas, thoughts, and experiences to the people around them. There are numerous ways to achieve this and the best one among all is the gift of "Speech". Through speech everyone can very convincingly transfer their thoughts and understand each other. It will be injustice if we ignore those who are deprived of this invaluable gift. The only means of communication available to the vocally disabled is the use of "Sign Language**"**. Using sign language they are limited to their own world. Deaf-Dumb people need to communicate with normal people for their daily routine. There are some difficulties when they come across in certain areas like banking, hospital etc. To overcome their problem a proper sign language is needed other than their existing communication method like lip reading, writing down word and finger spelling. Sign language is the main technique for deaf-dumb communication. This language cannot be recognized by most of the normal people and blind people. They face difficulties in their way of communication. To facilitate their communication, system that translates spoken language into sign language could be helpful. The developed system is the first step towards the final goal of translating spoken language to sign language via speech recognition. This system may be installed on a portable computer that acquires the speech and translates it immediately to sign language displaying it on the portable computer. The current stage of the progress focuses on translating speech signal to American Sign Language.

II. SURVEY

Various systems were proposed for the automatic recognition of sign language Don Pearson in his approach "Visual Communication Systems for the Deaf" [1] presented a two way communication approach, where he proposed the practicality of switched television for both deaf-to-hearing and deaf-to-deaf communication. In his approach, attention is given to the requirements of picture communication systems, which enable the deaf to communicate over distances using telephone lines. This section discusses some research done on translating other text and spoken languages to Sign language. Cox et. al, [2] presented a system that translates the English speech to the British Sign Language (BSL) using a specially developed avatar. However, the system is constrained for post office operations. The system uses a phrase lookup approach due to the highly constraint in the Macintosh operating system. The authors divided the task into three different problems:

- Automatic speech to text conversion
- Automatic translation of arbitrary English text into suitable representation of Indian sign language
- Display of this representation as a sequence of sign using computer graphics techniques

Developed system achieved accuracy of identification of the signed phrases of 61% for complete phrases & 81% for sign units. However, the feedback of deaf users and post office clerks were very encouraging for further development. A group of 21 researchers at DePaul University [3,4] participated in developing an automated American sign Language Synthesizer. Suszczanska. et.[5] developed a system to translate texts written in Polish Language into Polish Sign Language. They used Avatar as well with a dictionary of 600 signs. Scarlatos. et.al., [6] introduced a system to translate speech into video clip of the American Sign Language (ASL). The system displays the ASL clips along with the written words. They used a built-in speech recognition engine in the Macintosh operating system.This added a limitation as this engine can only recognize words from a pre-defined set. They plan to extend the system to recognize more words and later for phrases. San-Segundo, and others, [7] developed a system to translate speech into Spanish Sign Language. Their system is made up of four modules: speech recognizer, semantic analysis, gesture sequence generation and gesture playing. For the speech recognizer, they used modules developed by IBM. For the semantic analysis they used modules developed by the University of Colorado. For gesture sequence generation they used the semantic concepts associated to several Spanish Sign Language gestures. For gesture animation they developed an animated character and a strategy for reducing the effort in gesture generation. The strategy consists of making the system generate automatically all agent positions necessary for the gesture animation

Towards the development of automated speech recognition for vocally disabled people a system called "Boltay Haath" [8] is developed to recognize "Pakistan Sign Language"(PSL) at Sir Syed university of Engineering and Technology. The Boltay Haath project aims to produce sound matching the accent and pronunciation of the people from the sign symbol passed. A wearing Data Glove for vocally disabled is designed, to transform the signed symbols to audible speech signals using gesture recognition. They use the movements of the hand and fingers with sensors to interface with the computer.

The system is able to eliminate a major communication gap between the vocally disabled with common community. But Boltay Haath has the limitation of reading only the hand or finger movements neglecting the body action, which is also used to convey message. This gives a limitation to only transform the finger and palm movements for speech transformation. The other limitation that can be seen with Boltay Haath system is, the finger could be able to communicate with a normal person but the vice versa is not possible with it. This gives the limitation of one-way communication between the listeners and vocally disabled. This paper opens the door for research of translating the speech signal to American Sign Language

III. SYSTEM APPROACH

An automated speech recognition system is proposed for the recognition of speech signal and transforms it to a cue symbol recognizable by vocally disabled people. Figure 1 shows the proposed architecture for automated recognition system.

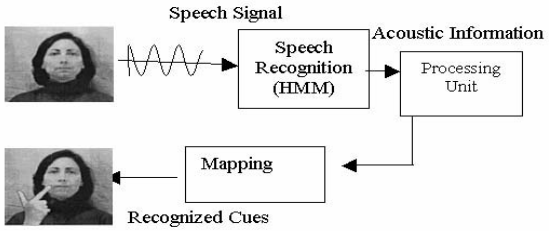


Figure 1 proposed AISR systems

The system implements a speech recognition system based on the speech reading and the cue samples passed to the processing unit. The processing system consists of a speech recognition unit with cue symbol generator, which determines the speech signal and produces an equivalent coded symbol for the recognized speech signal using HMM process; the system performed four principal functions:

1) Capture and parameterization of the acoustic speech input.
2) Signal identification via speech recognition.
3) Presentation of the identified signals to the processing unit.
4) Mapping the corresponding speech information into equivalent Cue symbols using Euclidian distance approach

(A) WORKING PRINCIPLE

The proposed system perform three principle functions

1) Capture and parameterization of the acoustic speech input.
2) Signal identification via speech recognition and generates an equivalent symbol.

3) Generate an equivalent cue symbol based on the coded symbol obtained from the speech recognition unit.

The recognition is performed using Hidden Markov Model (HMM), training the recognition system with speech features. A speech vocabulary for commonly spoken speech signal is maintained and its features are passed to the recognition system. On the recognition of the speech sentence the system generates and equivalent coded symbol in the processing unit. The symbols are then passed to the cue symbol generator unit, where an appropriate cue symbol is generated using the LMSE algorithm. For the generation of cue symbol a cue data base consisting of all the cue symbols are passed to the cue symbol generator. Figure 2 shows the cue symbols passed to the system.

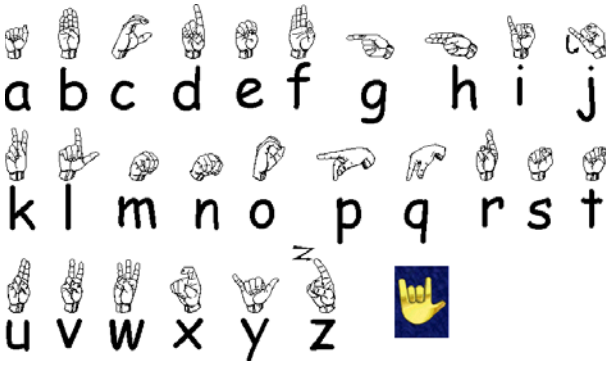
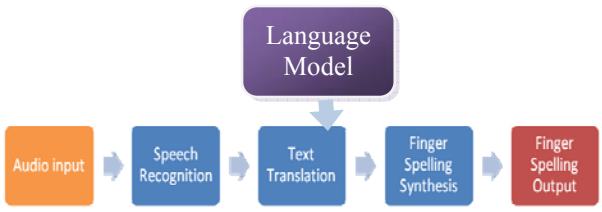


Figure 2 Equivalent English cue symbols for Database the symbols passed are the equivalent English characters.

(B). DESIGN OF THE OVERALL SYSTEM

In this work the design of the overall system will be implemented. The system will be operating in close to realtime and will take the speech input from the microphone and will convert it to synthesized speech or finger spelling. Speech recognition will be implemented for the considered languages. Language models will be used to solve ambiguities. Finger spelling synthesis will be implemented. The system is given in Figure 3



The system is given in Figure 3

IV. HIDDEN MARKOW MODEL OPERATION (HMM)

The project implements a speech recognition system based on the speech reading and the cue samples passed to the processing unit. The processing system consists of a speech recognition unit with cue reader, which determines the speech signal and reproduces the extracted speech signal using HMM process.

The operational functionality of the HMM modeling is made as; A Hidden Markov Model is a statistical model for an ordered sequence of variables, which can be well characterized as a parametric random process. It is assumed that the speech signal can be well characterized as a parametric random process and the parameters of the stochastic process can be determined in a precise, well-defined manner. Therefore, signal characteristics of a word will change to another basic speech unit as time increase, and it indicates a transition to another state with certain transition probability as defined by HMM. This observed sequence of observation vectors O can be denoted by

$$O = (o(1), o(2),\text{----}o(T))$$

Where each observation of ('t') is an m-dimensional vector, extracted at time't' with

$$O(t)=[0_1(t),0_2(t),\ldots..,o_m(t)]^T$$

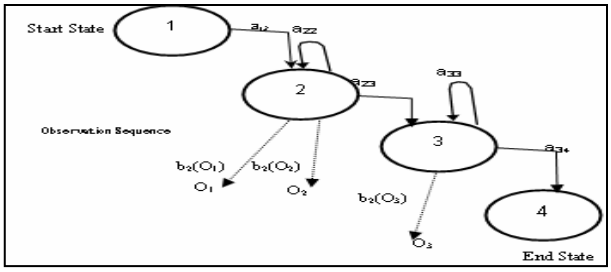


Figure.4 A typical left-right HMM (a_{ij} is the station transition probability from state i to state j; O_t is the observation vector at time t and $bi(O_t)$ is the probability that O_t is generated by state i).

The HMMs is used to solve three main problems. These problems are described as following:

1: Given the model $\lambda = \{A,B, \prod\}$ and the observation sequence, how to efficiently compute $P(O|\lambda)$, the probability of occurrence of the observation sequence in the given model.

Problem 2: Given the model $\lambda=\{A,B, \prod\}$ and the observation sequence, how to choose a optimal corresponding state sequence.

Problem 3: How to adjust the model parameters $\lambda = \{A,B,\prod\}$ so that $P(O \mid \lambda)$ is maximized.

Problem 1 and Problem 2 are analysis problems while problem 3 is a synthesis or model-training problem. To solve these three problems, some basic assumptions are being made in HMM.

a) The output independence assumption: The observation vectors are conditionally independent of the previously observed vectors.

b) The stationary assumption: It is assumed that state transition probabilities are independent of the actual time at which the transition takes place. It can be formulated mathematically as,

$$P[q_{t1+1} = j \mid q_{t1} = i] = P[q_{t2+1} = j \mid q_{t2} = i] \text{ for any } t_1 \text{ and } t_2.$$

The determination of the optimal set Ω of parameters in correspondence to a given utterance can be undertaken by relying on a simple property of the quantities to be maximized in both the two cases (MLE, MAP). Both the quantity to be maximized and the parameters we are looking for are probabilities, i.e. nonnegative quantities smaller than 1. Their variations during the optimization process from the starting values to the final optimized ones are very small. As a consequence, all these variations can be considered as differentials. If Q is the quantity to be maximized and its starting and final value, after maximization, is respectively Q_{start} and $Q_{opt,}$ we can write:

$$Q_{opt} - Q_{start} = Dq$$

Similarly, the variations of the parameters of the model, from the starting values to the final optimized ones, can be considered as differentials:

$$d, \pi_{i,} da_{ij,} db_i(Y_t), i = 1,\ldots,N, J=1,\ldots N, t = 1,\ldots, T.$$

q being a parameter, q' denoting its optimal value and q_{start} the initial value from which we start the maximization. Consequently, the determination of the optimal values of e can be simply undertaken by maximizing above equation with respect to Ω' and therefore neglecting in above equation the initial values Ω_{start}. The coefficients multiplying logarithms of the parameters are determined on the basis of Y_T and Ω_{start}. The maximization procedure initially requires modeling densities $b_i(y)$. Among the several possible solutions, the most used is based on mixtures of Gaussian functions and the $b_i(y)$ is themselves constrained to

$$\int b_i(Y)\, dy = 1 \; ; \; \int b_{ik}(Y)\, dy = 1$$

The above model is reasonable on the basis of the regularization theory applied to the approximation of unknown mappings, as is the case in the present situation. The consequence of this model on function a (Ω, Ω') is that of modifying the term where the output probabilities appear. The developed system performed two principal functions:

A. PARAMETERIZATION

The first two functions were related with each other where the recognition algorithm determining the type of speech preprocessing. The speech waveform was captured using a standard omni-directional microphone, sampled at 10-kHz, high-frequency pre processed by a first-order filter with a cutoff frequency of 150 Hz (to flatten the spectrum.) For each frame a vector of 25 parameters was derived from the samples. Differences were computed over a four-frame span so that the difference parameters of the frame were computed from the static parameters of frames n+2 and n-2.

B. RECOGNITION

The second subsystem recognized the phones corresponding to the acoustic input and converted them to a time-marked stream of signal codes, which was sent to the display subsystem Static and dynamic parameters formed distinct "streams" with different probability densities, under the implicit assumption that static and dynamic parameters were statistically independent

IV. MAPPING

Mapping of corresponding speech information into equivalent Cue symbols is done using Euclidian distance approach. The classification of the query is carried out using Euclidean distance. The Euclidean distance function measures the query & knowledge distance. The formula for this distance between a point ***X*** (***X1, X2,*** etc.) and a point ***Y*** (***Y1, Y2,*** etc.) is:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Deriving the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values. The system automatically starts searching the database for the words that start with the specified word. This process continues word by word until the last word is compared. The system recognizes the presence of sign in the database. If it exists, it is called and shown on the monitor of the portable computer; otherwise the sign is finger-spelled just like what deaf people do in their daily life. If the word is found in the Dictionary, then the corresponding cue symbol related to the word is displayed filling the entire page as shown in Figures 5 to 8. However, if the word is not found to be in the database, then all the words in the entire

database are displayed in the window as clearly as possible as shown in Figure 9.

V. RESULT OBSERVATION

For the training of HMM network for the recognition of speech, a vocabulary consisting of collection of words are maintained. The vocabulary consists of words given as, "ANTS", "BOAT", "BONFIRE", "CAMP", "DAWN", "DUSK", "FLIES", and "BUGS". Each word in the vocabulary is stored in correspondence to a feature defined as knowledge to each speech word during training of HMM network. The features are extracted on only speech sample for the corresponding word. The recognized speech word is processed for first 6 words and their symbol as shown below.

1) Test sample: 'BONFIRE'
Obtained cue symbol is,
'BONFIRE'

Figure 5 Obtained cue symbol for speech Sample

2) Test sample: 'BUGS'
Obtained cue symbol is,

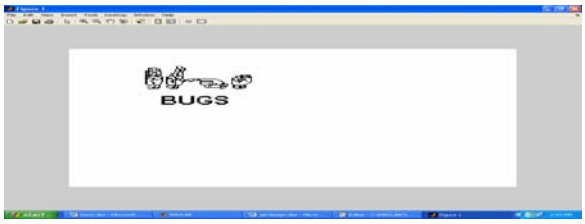


Figure 6: Obtained cue symbol for speech sample' BUGS'

3) Test sample: 'DAWN'
Obtained cue symbol is,

Figure 7: Obtained cue symbol for speech sample 'DAWN'

5) Test sample: 'FLIES'
Obtained cue symbol is,

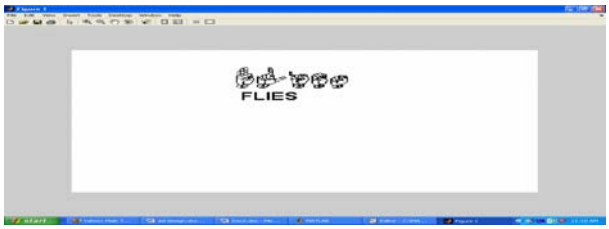


Figure 8: Obtained cue symbol for speech sample 'FLIES'

6) Test sample: 'WATER'
Does not exist cue symbol is,

Figure 9.The word that does not exist in the Dictionary is finger-spelled

VI. CONCLUSION

This paper presents an approach towards automated recognition of speech signal for vocally disabled people. The system proposed could efficiently recognize the speech signal using HMM and generate an equivalent cue symbol. The proposed AISR system find its application for the vocally disable peoples for providing a communication link between normal and disabled people. The system could be integrated with finger spelling recognition system such as "Boltay Haath" for a complete communication between the common person and the vocally disable people. For the suggested work converting speech sample to characters and then cue is to be further extended for generating video samples of continuous sentences. To integrate speech interfacing device with visual system for real time speech cue transformation.

VII. REFERENCE

[1] DONPEARSON "Visual Communication Systems for the Deaf" IEEE transactions on communications, vol. com-29, no. 12, December 1981

[2] Alison Wary, Stephen Cox, Mike Lincoln and Judy Tryggvason "A formulaic Approach to Translation at the Post Office: Reading the Signs", The Journal of Language & Communication, No. 24, pp. 59-75, 2004.

[3] Glenn Lancaster, Karen Alkoby, Jeff Campen,Roymieco Carter, Mary Jo Davidson, Dan Ethridge, Jacob Furst, Damien Hinkle, Bret Kroll, Ryan Layesa,

BarbaraLoeding, John McDonald, Nedjla Ougouag, Jerry Schnepp, Lori Smallwood, Prabhakar Srinivasan, Jorge Toro, Rosalee Wolfe, “Voice Activated Display of American Sign Language for Airport Security”. Technology and Persons with Dsabilities Conference 2003. California State University at Northridge, Los Angeles, CA March 17-22, 2003

[4] Eric Sedgwick, Karen Alkoby, Mary Jo Davidson, Roymieco Carter, Juliet Christopher, Brock Craft, Jacob Furst, Damien Hinkle, Brian Konie, Glenn Lancaster,Steve Luecking, Ashley Morris, John McDonald, Noriko Tomuro, Jorge Toro,Rosalee Wolf, “Toward the Effective Animation of American Sign Language”**.**Proceedings of the 9th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media. Plyen, Czech Republic,February 6 - 9, 2001. 375-378.

[5] Suszczanska, N., Szmal, P., and Francik, J., “Translating Polish Texts into Sign language in the TGT System”, the 20th IASTED International Multi-Conference on Allpied Informatics, Innsbruck, Austria, pp. 282-287, 2002.

[6] Scarlatos, T., Scarlatos, L., Gallarotti, F., “iSIGN: Making The Benefits of ReadingAloud Accessible to Families with Deaf Children”. The 6th IASTED International Conference on Computers, Graphics, and Imaging CGIM 2003, Hawaii, USA, August 13-15, 2003.

[7] San-Segundo, R., Montero, J.M., Macias-Guarasa, J., Cordoba, R., Ferreiros, J., and Pardo, J.M., “Generating Gestures from Speech”, Proc. of the InternationalConference on Spoken Language Processing (ICSLP'2004). Isla Jeju (corea).October 4-8, 2004.

[8] Aleem khalid ,Ali M, M. Usman, S. Mumtaz, Yousuf “Bolthay Haath – Paskistan sign Language Recgnition” CSIDC 2005

[9] Kadous, Waleed “GRASP: Recognition of Australian sign language using Instrumented gloves”, Australia, October 1995,pp. 1-2,4-8.

[10] D. E. Pearson and J. P. Sumner, “An experimental visual telephone system for the deaf,” J . Roy. Television Society vol. 16, no. 2. pp. 6-10, 1976.

[11] Guitarte Perez, J.F.; Frangi, A.F.; Lleida Solano, E.; Lukas, K. “Lip Reading for Robust Speech Recognition on Embedded Devices” Volume 1, March 18-23, 2005 PP473 – 476

[12] SantoshKumar,S.A.; Ramasubramanian, V.” Automatic Language Identification Using Ergodic HMM” Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP'05).IEEE International Conference Vol1, March18-23, 2005 Page(s) : 609-612