AIVP ADVANCES IN IMAGE AND VIDEO PROCESSING
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 4, ISSUE 2
ISSN 2054-7412

Visual Interface to Speech-Cue Representation Coding

[1]Ibrahim Patel, [2]Raghavendra Kulkarni, and [3]Y Srinivasa Rao
[1]Dept. of ECE B.V. Raju Inst. of Tech., Narsapur Medak,(T. S) India;
[2]Dept. of ECE K.M.C.E&T, Devarkonda under JNT University, Hyderabad, T. S
[3]Department of Instrument Technology, Andhra University, Visakhapatnam, A. P,
ptlibrahim@gmail.com; srinniwasarau@gmail.com; raghavendrakulkarni444@gmail.com

ABSTRACT

There have being great efforts made in the development of automated Instrumentation system for speech recognition (AISR) to provide a two-way communication between deaf and vocal people. This system performance achievable with the output of current real-time speech recognition systems would be extremely poor relative to normal speech reception. An alternate application of AISR technology to aid the hearing impaired would derive cues from the acoustical speech signal that could be used to supplement speechreading. We propose a study of highly trained receivers of speech signal that indicates that nearly perfect reception of everyday connected speech materials can be achieved at near normal speaking rates. To understand the accuracy that might be achieved with automatically generated cue symbols for visual representation. The system uses (HMM) for recognition of voiced data & Euclidian distance approach for sign language. 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. A New communication Paradigm: "Action-To-Speech"

Keywords: AISR, Speech recognition, HMM, vocally disabled, communication gap, speech-processing, cue-symbol.

1 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 the rest is the gift of "Speech". Through speech everyone can very convincingly transfer their thoughts and understands 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". There are approximately 10 million (8.487%) deaf people in India and nearly 1.25 billion persons with hearing impairments and close to a million who are functionally deaf in the United States. Without Assistive Technologies, there is no possibility for the hearing impaired to recognize sounds efficiently. Medical or surgical solutions such as cochlear implants may not always be possible. Using sign language they are limited to their own world. This limitation prevents them from interacting with the outer world to share their feelings, creative ideas and Potentials.

Another problem is that very few people who are not themselves deaf ever learn to Sign language. This further increases the isolation of deaf and dumb people from the common society. Technology is one way to remove this hindrance and benefit these people. Several researchers have explored these possibilities and have successfully achieved finger spelling recognition with high levels of accuracy. But progress in the recognition of sign language, as a whole has various limitations in today's applications.

DOI: 10.14738/aivp.42.2010
Publication Date: 25rd April, 2016
URL: http://dx.doi.org/10.14738/aivp.42.2010

Various systems and algorithms were proposed for the recognition of sign language. A system called "Boltay Haath" is developed to recognize "Pakistan Sign Language" (PSL) for vocally disabled peoples 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 able to eliminate a major communication gap between the vocally disable with common community.

2 State-of-The-Art

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 the rest is the gift of "Speech". Through speech everyone can very convincingly transfer their thoughts and understands 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. This limitation prevents them from interacting with the outer world to share their feelings, creative ideas and Potentials. Another problem is that very few people who are not themselves deaf ever learn to Sign language. This further increases the isolation of deaf and dumb people from the common society. Technology is one way to remove this hindrance and benefit these people. Several researchers have explored these possibilities and have successfully achieved finger spelling recognition with high levels of accuracy. But progress in the recognition of sign language, as a whole has various limitations in today's applications. Various systems and algorithms were proposed for the recognition of sign language. A system called "Boltay Haath" [1] is developed to recognize "Pakistan Sign Language"(PSL) for vocally disabled peoples 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 able to eliminate a major communication gap between the vocally disable 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 signer 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. A similar system is proposed by Kodous and Waleed [2] where they propose a Recognition system for Australian sign language using Instrumented gloves. This proposal also gives the same limitations as seen with Boltay Haath. Don Pearson in his paper "Visual Communication Systems for the Deaf" [6] 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 paper attention is given to the requirements of picture communication systems, which enable the deaf to communicate over distances using telephone lines. Extensions of such systems using the public switched telephone network may be possible if the images can be coded into low data rates [13].

3 Methodology

Speech recognition is motivated by the need to improve the performance of voice communications systems in noisy conditions. The applications range from front-ends for speech recognition systems, to enhancement of telecommunications in aviation, military, teleconferencing, cellular and biomedical applications. The goal is either to improve the perceived quality of the speech, or to increase its intelligibility. Speech enhancement is concerned with the processing of noisy and corrupted speech to improve the quality or intelligibility of the signal. Improving quality can be important for reducing listener accuracy in high stress and high noise environments. The precision for a speech recognition system can be measured in terms of speech recognition performance. Various rang of application were found for speech recognition in which one major application is the speech reading.

4 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. Fig. 1 shows the proposed architecture for automated recognition system.

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. In this work the design of the overall system will be implemented. The system will be operating in close to real-time 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

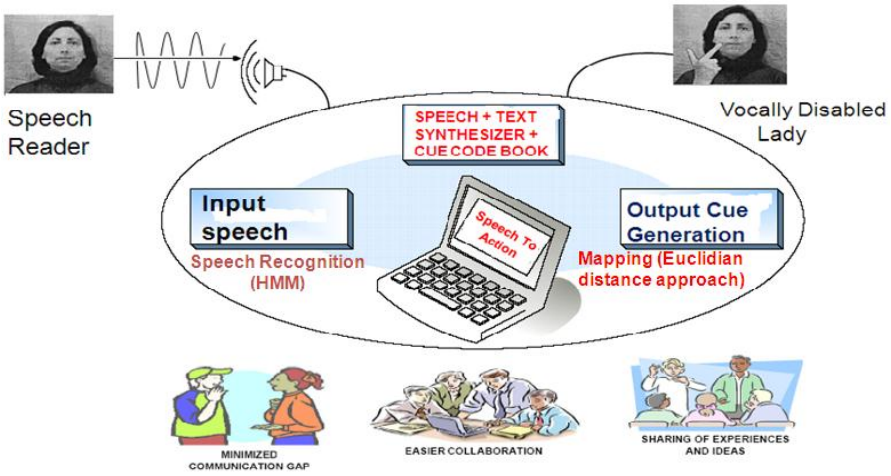


Figure 1: Proposed Automated Instrumentation Speech Recognition System

5 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. Finger spelling synthesis will be implemented. The system is given in Figure 2

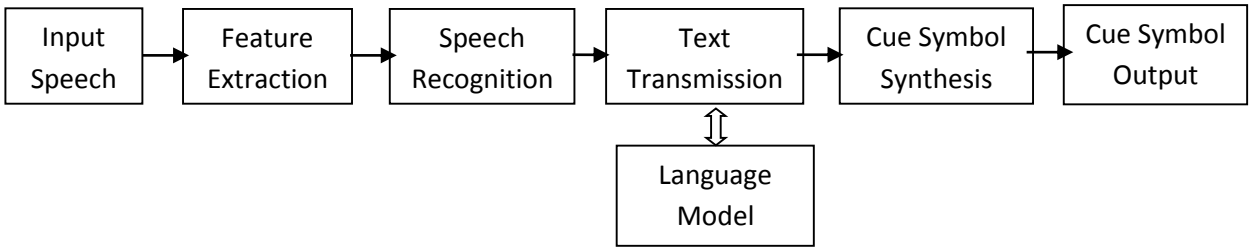


Figure.2: over system implementation

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.3 shows the cue symbols passed to the system.

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.

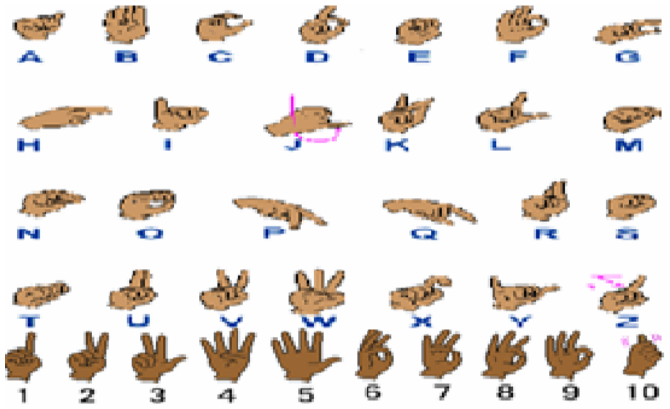


Figure 3 Equivalent English cue symbols for database. The symbols passed are the equivalent English characteristics.

5.1 Mel Spectrum Approach

A block diagram of the structure of an MFCC processor is given in Figure 4 the speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.

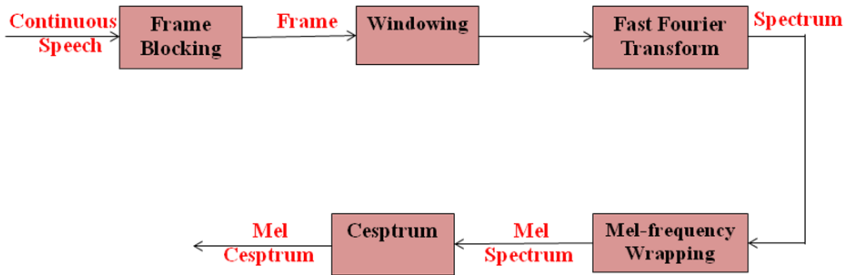


Figure 4: Block diagram of the MFCC Processor.

5.2 Hidden Markow Model Operation (HMM)

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 shown in fig 4. This observed sequence of observation vectors O can be denoted by

$$O = o(1), o(2), \ldots\ldots o(T) \tag{1}$$

where each observation of (t) is an m-dimensional vector, extracted at time t with

$$O(t) = [O_1(t), O_2(t), \ldots\ldots\ldots\ldots O_m(t)]^T \tag{2}$$

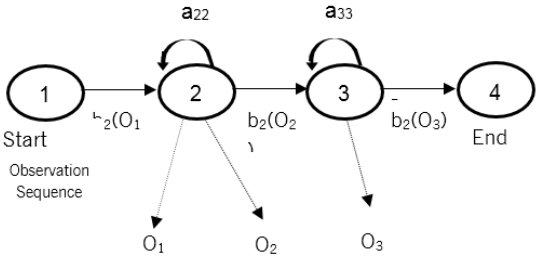


Figure.5 A typical left-right HMM (a_{ij} is the station transition probability from state i to state j; $O(t)$ is the observation vector at time t and bi $O(t)$ is the probability that $O(t)$ is generated by state i).

An HMM could be very complicated, but in general they can all be characterized by the following parameters:

a) N, the number of the states in the model. The state is hidden, however, each state within a process usually has some physical significance, like in the case of speech recognition, and each state could represent a basic speech unit. The state were denoted as $S = (s_1, s_2 \ldots\ldots s_N)$ and the state at time t as q_t.

b) M, the number of the Gaussian mixture components per state, i.e., the discrete alphabet size. The individual symbols are denoted as $V = \{v_1, v_2, \ldots\ldots v_M\}$

c) A, the state transition probability distribution $A = \{a_{ij}\}$ where the probability of being in state s_j at time $t+1$ given that we were in state s_j at time t and

$$a_{ij} = p[q_{t+1} = s_j, q_t = s_i], 1 < i,\ j < N \tag{3}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad 1 < j < N, \tag{4}$$

There are many types of HMMs. For the special case such as ergodic model where all states can be reached by any other states, $a_{ij} > O$ for all $i,\ j,$

d) B, for continuous HMMs, it is the matrix of observation probability distribution over all the state and all the observations. $B = \{b_j(k)\}$, where

$$b_j(k) = p[o_t = v_k \ \ q_t = s_i], \qquad 1 < j < N \qquad 1 < j < T \tag{5}$$

$$V = \{v_1, v_2, \ldots\ldots v_M\} \quad \text{and}$$

$$\sum_{t=1}^{T} b_j(t) = 1 \qquad 1 < j < N \tag{6}$$

e) $\prod$, the initial state distribution $\prod = \{\pi_i\}$, in which

$$\pi_i = p[q_1 = s_j] \qquad 1 < i < N \tag{7}$$

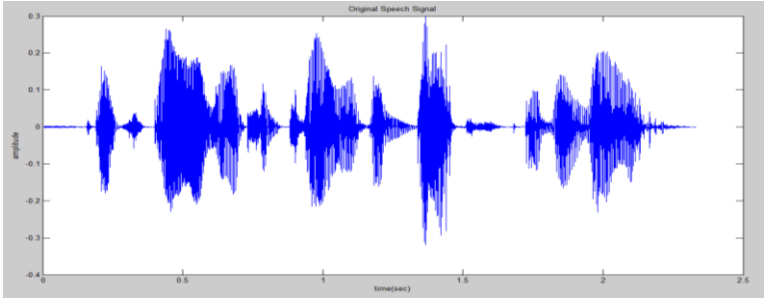
A complete specification of a HMM requires specification of two model parameters, N and M, specification of the observation symbols, and the specification of three sets of probability measures $A,\ B,\ \pi_i$ so an HMM can also be defined as a compact form $\lambda = \{A,\ B,\ \pi\}$.

5.3 Analyzer

The system evaluates the parameter of recognition system for various noises considering MFCC & MFCC with sub band as feature extraction technique. The analyzer model reads the parameter such as computation time the learning rate accuracy level, qualification rate & with respect to time to analyzing efficiency implemented system.

For the training of HMM network for the recognition of speech a vocabulary consist of collection words are maintained. The vocabulary consists of words given as, “DISCRETE”, “FOURIER”, “TRANSFORM”, “WISY”, “EASY”, “TELL”, “FELL”, “THE”, “DEPTH”, “WELL”, “CELL”, “FIVE”, each word in the vocabulary is stored in correspondence to a feature define as a knowledge to each speech word during training of HMM network. The features are extracted on only voice sample for the corresponding word. Test speech utterance: “it’s easy to tell the depth of a well”, taken at 16 KHz shown in figure 6 (a) (b) and (c) and 7. The speech signal are decomposed into a set of sub-bands with a hierarchical coding of speech signal using set of high and low pass filters. The obtained bands are then processed with the mel-frequency (MFCC) estimation, where mel frequencies are extracted for each of the band. This results in extraction of mel feature coefficient at a finer spectral level.

Test sample S1:
File Name: S1.wav
Sentence: “It easy to tell the depth of well”
Duration: 0.04 sec



(a)

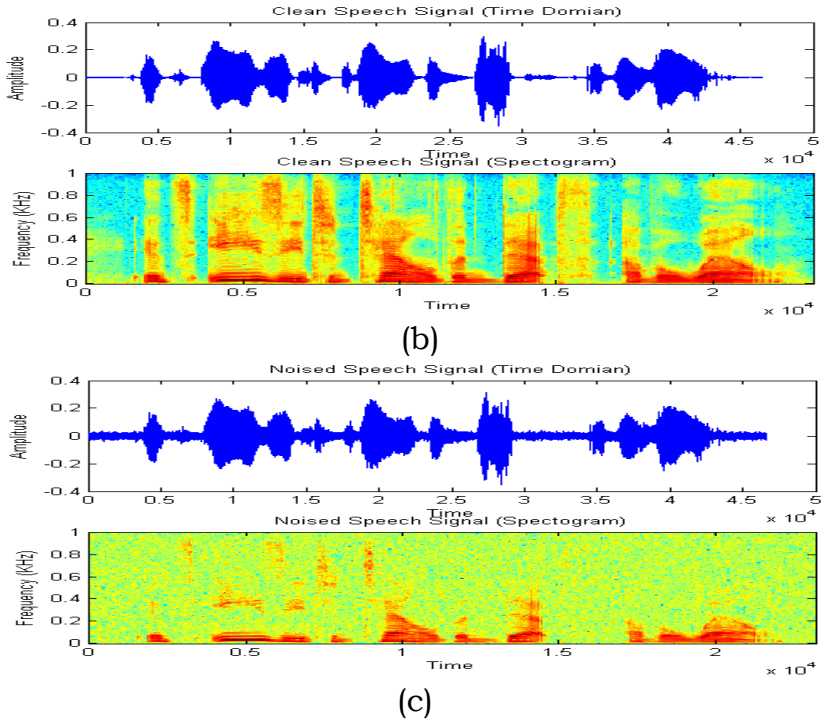


Figure 6 (a): Original Test sample (S1), (b) Spectral plot for clean speech, (c) Noise affected signal, with its spectral plot

Figure 7: Computation Iteration for the developed methods

$$Retrieval\ Accuracy(\%) = \left(\frac{No.of\ truely\ recognition\ Words}{Total\ No of\ Words} \right) \times 100$$

6 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:

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

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 starts with the specified word. This process continues letter by word until the last word. The system recognizes if the sign exists in the database or not. 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.

6.1 Simulation Observation

For the simulation of the suggested approach various speech samples are been trained and tested. Speech samples from 'A' to 'Z' were recorded and their corresponding cue symbols are stored onto a database. The training database features are as tabulated,

Training Character	Energy Level, $E = (\sum_{i=1}^{m}\sum_{j=1}^{n} x(i,j))$
A	75
B	120
C	39
D	65
E	46
F	52
G	73
H	78
I	42
J	53
K	71
L	45
M	106
N	110
O	70
P	108
Q	98
R	77
S	57
T	56
U	87
V	59
W	91
X	78
Y	54
Z	61
A	95
B	97
C	62
D	93
E	68
F	70
G	78
H	106
I	50
J	56
K	89
L	62
M	120
N	71
O	84
P	76
Q	94
R	97
S	69

T	60
U	65
V	59
W	104
X	73
Y	61
Z	100

If the word is found in the Dictionary, then the cue clip related to the word is displayed filling the entire page as shown in Figure 8 to 11. However, if the word is found not to be in the database, the window is divided according to the number if words so that the entire word is displayed in the window as clearly as possible as shown in Figure 11. For the training of HMM network for the recognition of speech a vocabulary consist of collection words are maintained. The vocabulary consists of words given as, “BOOK”, “BANK”, “FINISH”, “AND”, “DAWN”, “DUSK”, “FLIES”, “BUGS”, “DEPTH”, “WELL”, “CELL”, “FIVIE”, each word in the vocabulary is stored in correspondence to a feature define as a knowledge to each speech word during training of HMM network. The features are extracted on only speech sample for the corresponding word. Test speech utterance: “it’s easy to tell the depth of a well”, taken at 16 KHz. The recognized of the speech words is processed for first 6 words and the recognized character and there symbol is as shown below

1) Test sample: ‘BOOK’, Obtained cue symbol is,

Figure 8: Obtained cue symbol for speech Sample ‘BOOK’

2) Test sample: ‘AND’ , Obtained cue symbol is,

Figure 9: Obtained cue symbol for speech sample ‘AND’,

2) Test sample: ‘FINISH’, Obtained cue symbol is,

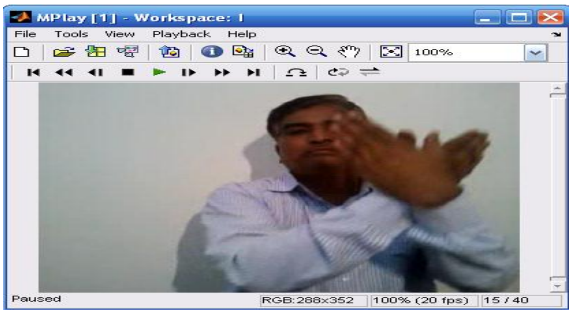


Figure 10: Obtained cue symbol for speech sample ‘FINISH’

4) Test sample: ‘BANK’, Obtained cue symbol is,

Figure.11. Obtained cue symbol for speech sample ‘BANK’

7 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 “Action-to-Speech” for a complete communication between the common person and the vocally disable people.

REFERENCES

[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 Disabilities 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,March18-23,2005Page(s):609-612