Real Time Speech to Text Converter for Mobile Users

Ms. Anuja Jadhav
M.Tech. Student Y.C.C.E. Nagpur India

Prof. Arvind Patil
Y.C.C.E., Nagpur, India

ABSTRACT

Mobile phone usage in World is spreading rapidly and has gone through great changes due to new developments and innovations in mobile phone technology. This project based on evaluating voice versus keypad as a means for entry and editing of texts. In other words, messages can be voice/speech typed. The project will make use of a dictating-machine prototype for the English language, which recognizes in real time natural-language sentences built from a 2000 word vocabulary. A speech to text converter is developed to send SMS .It is found that large-vocabulary speech recognition can offer a very competitive alternative to traditional text entry.

Keywords

Short Message Service (SMS); speech acquisition; Hidden Markov Model (HMM); HMM-based recognition.

1.INTRODUCTION

There is no doubt that more and more Mobile phone users are using short message service (SMS) instead of making voice calls. In order to satisfy the needs and demands of users, mobile phone manufacturers are constantly adapting and innovating to ensure that they can survive in this competitive market. An important innovation in SMS technology lately is the speech recognition technology that can convert voice messages into text messages. In other words, messages can be voice/speech typed. Currently, voice messages can only be converted into text messages in the form of normal/standard text using fully spelled words.

The cell phones are very important part of modern life. Many of us need to make a call or massage at anytime from anywhere. Many of them needs their cell phones when they can't do so e.g. At the time of driving, cooking accidents may occur because of this activity an speech to text converter for mobile design for this purpose so to avoid accidents. The study of speech to text conversion is from 1970s where the first experiment of phoneme- to-grapheme conversion, this conversion consists of segmentation of phoneme string into word. This work is again extended to stenotype-to-grapheme conversion. Voice massaging is slowly and gradually reducing the importance of text massaging because it is safer to massage at the time of cooking and driving. This paper introduces an idea about the speech-to-text conversion for SMS application. This software enable user to send the SMS without using keypad with fully spelled word.

FUNCTION DESCRIPTION

This project converts speech into text. At the run time speech data is quire from microphone and converted into text speech frames the speech frames are then pass for preprocessing and after preprocessing of the sample frames HMM-based training is applied on speech frame. Functionally the project divided into three modules

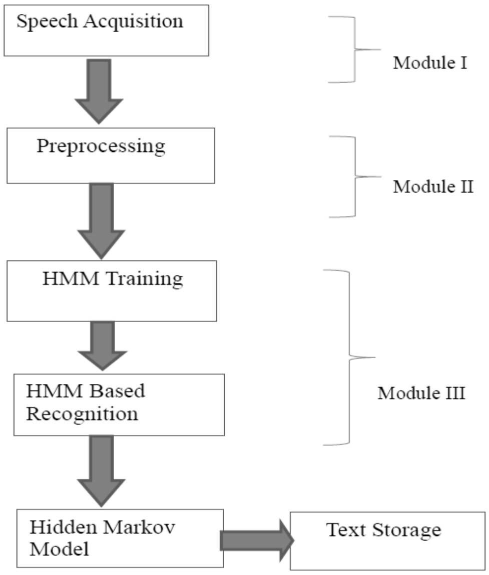

Fig:-1 Speech to text conversion system it is divided into three modules.

2.1 Speech Acquisition

In this phase speech samples are obtained from speaker at real time and stored for preprocessing. Speech acquisition require microphone to receive voice speech signals, Speech acquisition can be easily done by the microphone present in the mobile phone, In the acquisition phase the different M/C is depends upon the its own configuration, hence there is need to store the sample of different users to make system more compatible to any type of voice. To recognize the speech HMM-based automatic recognition was conducted. For continuous phoneme recognition, an 86% phoneme correct was achieved for the normal-hearing.

To achieve speech preprocessing sphinx frame work is used this is the best tool found to acquiesce speech signals. Sphinx is design with high flexibility modularity. Figure2 shows the overall architecture of sphinx. There are three modules in the sphinx frame work frontend, decoder, Linguist Front end takes the input speech signals and parameterized it into sequence of features. The Linguist translates any type of standard language model, along with pronunciation information from the Dictionary and structural information from one or more sets of AcousticModels, into a SearchGraph. The SearchManager in the Decoder uses the Features from the FrontEnd and the SearchGraph from the Linguist to perform the actual decoding, generating Results. At any time prior to or during the recognition process, the application can issue Controls to each of the modules, effectively becoming a partner in the recognition process.

2.2 Speech Preprocessing

The speech signals consist of background noise that need to be removed. The preprocessing reduces the amount of efforts in next stages. Input to the speech preprocessing is speech signals which then converted into speech frames and gives unique sample

Steps:

1. The system must identify useful or significant samples from the speech signal. To accomplish this goal, the system divides the speech samples into overlapped frames.
2. The system performs checks for the voice activity using endpoint detection and energy threshold calculations.
3. The speech samples are then passed through a pre-emphasis filter.
4. The frames with voice activity are passed through a Hamming window. The system performs autocorrelation analysis on each frame.
6. The system finds linear predictive coding (LPC) coefficients using the Levinson and Durbin algorithm.
7. From the LPC coefficients, the system determines the cepstral coefficients and weighs them using a tapered window. The cepstral coefficients serve as feature vectors.

i) Pre-Emphasis

The signals after digitization i.e. s(n) is put through a low-order LPF to make it less susceptible to finite precision effects later in the signal processing.

ii) Frame Blocking

The speech frames are then form with the duration of 56.25ms if we consider N=450 sample length and it then overlap of 18.75 Ms if M=18.75of 18.75 ms (M = 150 sample length) between adjacent frames. The overlapping ensures that the resulting LPC spectral estimates are correlated from frame to frame and are quite smooth.

iii) Windowing

We apply a Hamming window to each frame to minimize signal discontinuities at the beginning and end of the frame.

iv) Detection of Voice Activity

The system uses the endpoint detection algorithm to find the start and end points of the speech. The speech is sliced into frames that are 450 samples long. Next, the system finds the energy and number of zero crossings of each frame. The threshold energy and zero crossing value are determined based on the computed values and only frames crossing the threshold are considered, removing most background noise.

2.3 HMM Training

An important part of speech-to-text conversion using pattern recognition is training. Training involves creating a pattern representative of the features of a class using one or more test patterns that correspond to speech sounds of the same class. A model commonly used for speech recognition is the HMM, which is a statistical model used for modeling an unknown system using an observed output sequence. The system trains the HMM for each digit in the vocabulary using the Baum-Welch algorithm. The codebook index created during preprocessing is the observation vector for the HMM model.

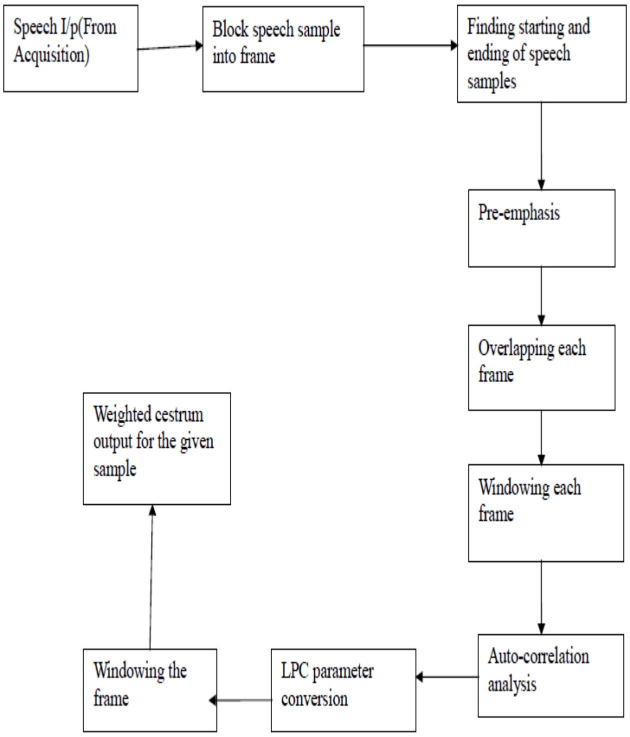
Speech I/p(From Acquisition)
Block speech sample into frame
Finding starting and ending of speech samples
Pre-emphasis
Overlapping each frame
Windowing each frame
Auto-correlation analysis
LPC parameter conversion
Windowing the frame
Weighted cestrum output for the given sample

Fig: 2 the diagram showing Preprocessing of samples before conversion of speech into text.

i)Temporal Cepstral Derivative

We can obtain improved feature vectors for the speech frames using temporal cepstral derivatives. We use them with the cepstral derivative if the cepstral coefficients do not have acceptable recognition accuracy.

ii) Vector Quantization

A codebook of size 128 is obtained by vector quantizing the weighted cepstral coefficients of all reference digits generated by all users. The advantages of vector quantization are:

- Reduced storage for spectral analysis information.
- Reduced computation for determining the similarity of spectral analysis vectors.
- Discrete representation of speech sounds. By associating phonetic label(s) with each codebook

Vector, choosing the best codebook vector to represent a given spectral vector is the same as assigning a phonetic label to each spectral speech frame, making the recognition process more efficient. One obvious disadvantage of vector quantization is the reduced resolution in recognition. Assigning a codebook index to an input speech vector amounts to quantizing it, which results in quantization errors?

Errors increase as the codebook size decreases. Two com algorithms are commonly used for vector

Quantization: the K-means algorithm and the binary split algorithm. In the K-means algorithm, a set of L training vectors can be clustered into M (<L) codebook vectors, as follows:

- Initialization— arbitrarily chooses M vectors as the initial set of codewords in the codebook.
- Nearest neighbor search— for each training vector, find the codeword in the current codebook that is

closest and assign that vector to the corresponding cell.

- Centroid update— Update the codeword in each cell using the centroid of the training vectors assigned to that cell.
- Iteration— Repeat the above two steps until the average distance falls below a preset threshold.

Our implementation uses the binary split algorithm, which is more efficient than the K-means algorithm because it builds the codebook in stages as described in the following steps:

1. Design a 1-vector codebook, which is the centroid of the entire training set and hence needs no iteration.

2. Double the codebook by splitting each current codebook yn according to the rule:

yn+ = yn(1 + e)

yn- = yn(1 - e)

Where n varies from 1 to the codebook size and e is the splitting parameter.

3. Use the K-means iterative algorithm to obtain the best set of centroids for the split codebook.

4. Iterate the above two steps until the required codebook size is obtained.

2.4 HMM- Recognition

Recognition or pattern classification is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity (distance) between the test pattern and each reference pattern. The digit is recognized using a maximum likelihood estimate, such as the Viterbi decoding algorithm, which implies that the digit whose model has the maximum probability is the spoken digit. Preprocessing, feature vector extraction, and codebook generation are same as in HMM training. The input speech sample is preprocessed and the feature vector is extracted. Then, the index of the nearest codebook vector for each frame is sent to all digit models. The model with the maximum probability is chosen as the recognized digit.

PERFORMANCE PARAMETERS

1) Recognition accuracy— the most important parameter in any recognition system is its accuracy. A recognition accuracy of 100% for all digits, independent of the speaker, is the goal.

2) Recognition speed— if the system takes a long time to recognize the speech, users would become restless and the system loses its significance. A recognition time of less than 1 second is required for the project.

3) Recognition Accuracy

Case 1—Used samples from speakers 1, 2, and 3 that were also used for training.

Case 2—Used samples from speakers 1, 2, and 3 that were not used for training.

Case 3—used samples of 3 untrained speakers (4, 5, and 6) whose voices were not used for training.

Case 4—Used samples from speakers 1, 2, 3, 4, 5, and 6 to obtain the overall recognition performance.

4.CONCLUSION

In this paper, isolated speech recognition and continuous text generation for users, Hidden Markov Model is used here to perform HMM-based automatic recognition

With this software the mobile phone usage and communication among the mobile users increases. Call routers become easier for users, since they don't need to know how to spell a name in order to say it. It becomes easier for users who are driving or otherwise incapable of looking at keypads to interact with a system.

5. REFERENCES

[1] Andreas Stolcke, , Barry Chen, Horacio Franco, Venkata Ramana Rao Gadde, Martin Graciarena, , Mei-Yuh Hwang, Katrin Kirchhoff, , Arindam Mandal, Nelson Morgan, , Xin Lei, Tim Ng, Mari Ostendorf, Kemal Sönmez, Anand Venkataraman, Dimitra Vergyri, and Qifeng Zhu, "Recent Innovations in Speech-to-Text Transcription at Sri-icsi-uw" IEEE Transactions On Audio, Speech, And Language Processing, vol. 14, no. 5, september 2006, pp 1729-1744

[2] Brandon Ballinger, Cyril Allauzen, Alexander Gruenstein, Johan Schalkwyk, "On-Demand Language Model Interpolation for Mobile Speech Input", INTERSPEECH 2010, 26-30 September 2010, Makuhari, Chiba, Japan, pp 1812-1815

[3]Ryuichi Nisimura, Jumpei Miyake, Hideki Kawahara and Toshio Irino, "Speech-To-Text Input Method For Web System Using Javascript", IEEE SLT 2008 pp 209-212

[4]M. Tomalin, F. Diehl, M.J.F. Gales, J. Park & P.C. Woodland , "Recent Improvements To The Cambridge Arabic Speech-To-Text Systems", ICASSP 2010 pp 4382-4385

[5] Janet See, Umi Kalsom Yusof, Amin Kianpisheh, "User Acceptance towards a Personalised Handsfree Messaging Application (iSay-SMS)", CSSR 2010 Initial Submission December 5-7,2010 pp 1165-1170

[6] Panikos Heracleous, Hiroshi Ishiguro and Norihiro Hagita, "Visual-speech to text conversion applicable to telephone communication for deaf individuals" 18th International Conference on Telecommunication 2011. pp 130-133

[7] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1524–1538, Sep. 2006.

[8] M.J.F Gales, F. Diehl, C.K. Raut, M. Tomalin, P.C. Woodland, and K. Yu, "Development of a phonetic system for large vocabulary arabic speech recognition," in Proc. of ASRU, 2007.

[9]L. Nguyen, T. Ng, K. Nguyen, R. Zbib, and J. Makhoul, "Lexical and phonetic modeling for arabic automatic speech recognition,"in Proc. of Interspeech, 2009.

[10] C.C. Wong, "Enabling Ecosystem for Mobile Advertising in an Emerging Economy," Monash University Doctoral Colloquium. Langkawi, Kedah, Malaysia, 14-16 December 2009.

[11]G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," in Proceedings of the IEEE, vol. 91, Issue 9, pp. 1306–1326, 2003.

[12] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," in Proceedings of ICASSP 2002, 2002.

[13] Garg, Mohit. Linear Prediction Algorithms. Indian Institute of Technology, Bombay, India, Apr 2003.

[14] Li, Gongjun and Taiyi Huang. An Improved Training Algorithm in Hmm-Based Speech Recognition. National Laboratory of Pattern Recognition. Chinese Academy of Sciences, Beijing.

[15] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara,"Voice conversion through vector quantization," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 655-658, IEEE, April 1988

[16] M.J.F Gales, F. Diehl, C.K. Raut, M. Tomalin, P.C. Woodland, and K. Yu,"Development of a phonetic system for large vocabulary arabic speech recognition," in Proc. of ASRU, 2007.

[17] L. Nguyen, T. Ng, K. Nguyen, R. Zbib, and J. Makhoul, "Lexical and phonetic modeling for arabic automatic speech recognition," in Proc. of Interspeech, 2009.