ISSN:0975-9646

Efficient Methods and Implementation of Automatic Speech Recognition System

Himali Junghare[#1], Prof. Prashant Borkar[*2]

[#]Department of computer science,
G.H. Raisoni College of Engineering,
Nagpur University, Nagpur, India

***Abstract* – An automatic speech recognition (ASR) system is very popular and efficient process to converts sound signals into text. It is widely used in various applications in market such as speech-to-text, entries of data etc. Accuracy of ASR is one of the big challenge for the researchers because many voice inputs have noise and declined channel, it results mismatched condition from the point of view of competency and reliability. In sound recognition, it has three most important approaches as Artificial intelligence, second is Pattern recognition and the last one is Acoustic phonetics. The goal of this paper is to clear an idea about basics of voice recognition by differentiating methods and summarizing a theory of what is ASR? And how does it works? in each and every stages of voice recognition system. It discusses the progress and development of voice recognition in last five decades that has been performed and describes a literature review of a system. This paper presents innovative and distinctive features of a system.**

***Keywords*–Automatic speech recognition, Pattern recognition, Language model, Hidden markov model, Linear discriminate analysis, Mel frequency cepstral coefficient**

I. Introduction

An ASR is the process which converts speech signal into the text message or word sequence, it is also called as speech-to-text system. Speaking is very essential and vital means of conversation in the midst of the people as communication which is basically, the uttermost lenient form to deal with sharing an information in the humans. Speech is transmitted in its original form in the ordinary speech communication system without knowing its properties. ASR required to compressed an input speech words into small set of data to classify correctly as phonemes and involves to create a words one by one sequencely with foremost matches to the given input signal of speech waveform. It is complicated to convert speech into word sequence without compression of input data. An average rate of uttered sounds is approximately 12 per sec.

There are Many applications of speaker recognition such as data entry, speech-to-text, voice dialing, accessing the database services, telephone banking, telephone shopping by speaker dialing, information services, Forensic Purpose are in existence today. The goal of speech recognition is recognizing the voice in spoken words, also to analyse the speaker by extraction of features, simulating and enacting the information which consist in the input voice signal[1]. The accuracy of an ASR system are prevail by many parameters such as Independence or dependence from speaker, diverse word detection, consecutive word detection, thesaurus and discriminating of available large trained data or particular vocabulary in dictionary, environment like nature of noise, ratio of signal and noise, working status, transducers such as amplitude of band, microphone or telephone, distortion or repetition in channel conditions, also age, gender and physical state of speaker, speech style such normal, quite, shouted voice tone and different pronunciation of each word.

In voice recognition system, we have only sound input signal to build a model for word or phonetic structure to improve prediction by using often statisticaly. In ASR system, it should be identify and address the differences between spoken language and written language. The methods of a voice recognition is to be divided according to dependency of the text i.e. text dependent and text independent. If speech is recognized by speaker with his or her pin number, password or some particular phrases then it is called as text dependent class and if speech do not required to recognized by the speaker, what speaker is saying is called as text independent class. Speech recognition is again splits into close set and open set. Based on the Euclidian distance matching technique the speaker is not matched to any other speaker from database comes under open set speech and in close set speech any speaker is matched to the other speaker in the system database. [2]Element wise difference between the reference vector and unknown vector is found, squared and summed in the Euclidian distance procedure. If the Euclidian distance is small then better the match.

II. Types Of Speech

A voice recognition system is differentiate into the kinds of parameters in speech. Which depends on competency to recognize text and list of words in speech as below-

A. Isolated:

An isolated word generally accepts a single word or single utterances at a time that only involves a pause between the utterances and it have sample windows. Isolated utterance is a better name according to work. There are two states in this system Listen or Non-listen states where speaker have to wait for completion of each and every utterences.

B. Spontaneous:

Basically, it would be concluding a system as spontaneous speech ability that having fluency in a communication, also not trained well which should be able to dependency of the and "ahs", and even slight stutters. Also to grasp a different kind of natural features of voice like running of words simultaneously[3].

C. Continuous:

Continuous speech is described in which its grant the speaker to speak like natural conversation. In this type of speech, words are connected together in consecutive manner, while the computer will examine the content it is also known as computer dictation. The boundaries of the phonetic units as utterance and the number of problems existence, there are special methods to utilized.

D. Connected:

Connected speech is congruent to isolated as defined above but it grant to split a phonemes for running in sync with least pause between them. The connected state of the words or phonemes is extract from vocabulary which would be moderate size or small size dictionary such as sequence of digits, concatenated alphanumeric words, queue of spelled letters. In the Connected speech is a class of fluent speech words such like isolated word recognition it has a quality that the prime voice recognition unit is the hypothesis of the phrases or words with high intensity.

III. Related Issues Of ASR Model

There are various issues of ASR model on which accuracy depends as below:

TABLE I: RELATED ISSUES OF ASR MODEL

Speakers	Speaker independent; speaker dependent; Gender; Age etc
Word dictionary	Generic, Specific vocabulary; characteristics of trained data
Background	Noise ratio; noise signal and its type
Input	Telephone; Microphone
Speech type	Spontaneous speech; Isolated speech; Continuous speech;
Tone or Speed of speech	Normal; slow; fast; shouted; quite

IV. Various Techniques And Approaches Of Speech Recognition

The objective of recognition is to analysing, feature extracting, characterizing on the basis of voice input signal. For determining the speech characteristics various approaches and techniques are defined and it has some approaches. As given in the below figure, these are the several methods for analyzing and recognition of voice[26] as:

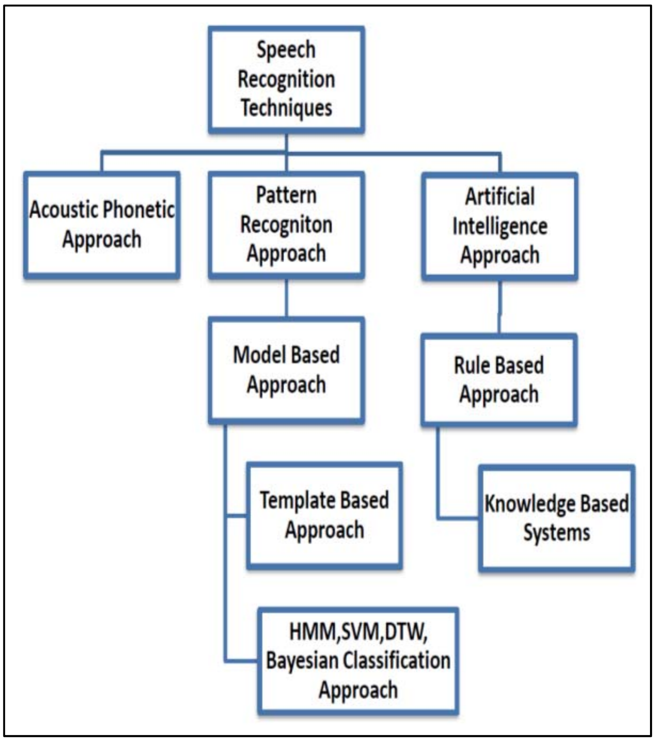
Speech Recognition Techniques
Acoustic Phonetic Approach
Pattern Recogniton Approach
Artificial Intelligence Approach
Model Based Approach
Rule Based Approach
Template Based Approach
Knowledge Based Systems
HMM,SVM,DTW, Bayesian Classification Approach

Fig.1: Speech recognition techniques classification

1. *Acoustic Phonetic:*

This approach differentiate phonetic units in speech that characterized by a set of acoustics properties which varies with who is speaking also with observing environmental condition like noise, it means following laws the variability in sound is straightforward that could be learned before by system. In this approach the first step is analysis of speech spectral, second step is extraction of feature that performs conversion of spectral analysis into collection of features which observes the acoustic condition of separated phonetic units.[4][5] The speech signal is segmented into isolated section in next stage that is segmentation and labelling step that will proceed by combining phonetic units to each segmental frame. Valid word is to be find in last step from the phonetic label sequences. This approach has not been used widely. Finding the sound and providing specific labels to that sound has been done before recognition of speech. A set of acoustic properties broadly characterized there different phonetic units in voice, this is the basis of the acoustic phonetic approach. If system requires to determine language then support vector machine is useful.

2. *Pattern Recognition:*

This approach is very useful for selecting patterns on the basis of some terms and to differentiate classes. There are number of stages as pattern classification. On the input signal a consecutive sequel of measurements to describe pattern. [6][7]One or more test patterns corresponding to voice creates reference pattern, it could be in the form HMM and it would be implement to a voice input, a phrase. Flow of pattern recognition approach as shown below: [26]

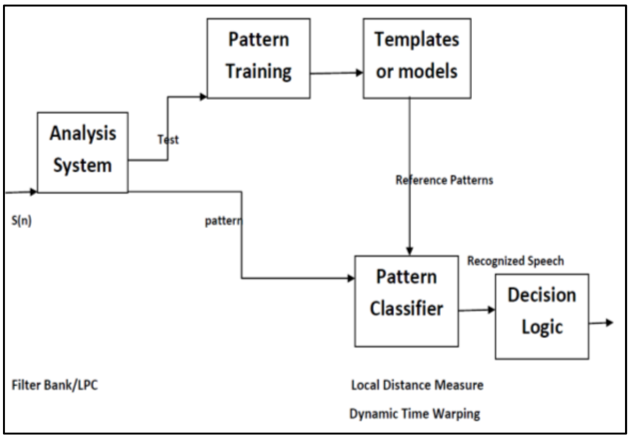

Fig.2: Pattern Recognition Approach

Matching Integrity of patterns is determined by the decision logic stage. In the last five decades this approach has become the best method,it has been developed for speech recognition [8].There are exist two methods i.e. stochastic model method and template method as below:

2.1 Stochastic method

To deal with incomplete information, stochastic methods characterise the use of probabilistic models. Incompleteness arises from several sources in speech recognition e.g. speaker variability, contextual effect etc.[9]Hidden Markov modeling is the most popular stochastic approach now a days.

2.2 Template Method

In this method, comparison is takes place between the template set and input audio according to find most similar match. Since, last six decades, a number of methods and techniques are broadly provided for this method to sound recognition. A large vocabulary is listed and stored in a prototypical sound patterns for the reference[10]. Using many templates per word that variations in speech can model has the disadvantage which eventually becomes unfeasible.

3. Artificial intelligence:

This approach is the combination of above two approaches that is pattern and the acoustic recognition[11]. To design a varieties of recognition systems these methods are very useful.

In the artificial intelligence approach it correlatively recognised process in order to whom it applies.

4. Knowledge based approach

In this approach, speech voice is handcoded from knowledge of experts as speech varies that guidance should be taken. It could be failure as it is complicated to obtain explicit model but it gives advantage of it[12]. In this approach, it required an information about spectrogram, phonetic and linguistic. All codebooks consider the test speech. In the point of view for speech coders for advantage generally VQ is used in system.

5. Learning based approach

This approach has overcome some disadvantages of hidden markov model for machine learning which introduced natural language and genetic algorithm. But by using evolutionary and the emulations process it can be done automatically.

V. Modules Of ASR System

Voice recognition is the conversion of acoustic signals into a digital signals which should be same as the input spoken words. These modules will help to understand the ASR system according to its phases. There are three most important phases are feature extraction, acoustic model and language model. Again there is a matching of words and matching of sequence also execute by the system. It is possible to allow a system which can perform as a stenographer, read the newspaper of user own choice and so on[13]. ASR system has following modules as shown in the figure[26].

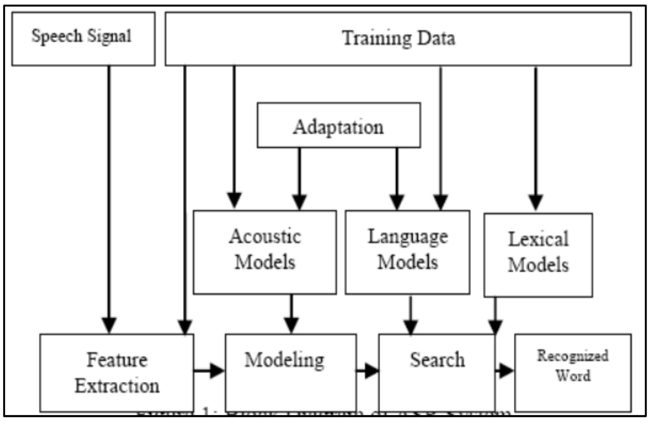

Fig.3: Block Diagram of ASR system

In the speech signal, it access the input signal convert into modified digital signal. It filters the noise in the signal. This is the preprocessing of the ASR system[14]. After preprocessed the signal next step is the feature extraction which analyse the input signal and generate power spectrum of each phoneme, also create a waveform of each phoneme and computes a feature vector. There are numerous techniques to analyse input signal such as LDA, PCA, MFCC [15][16]etc.

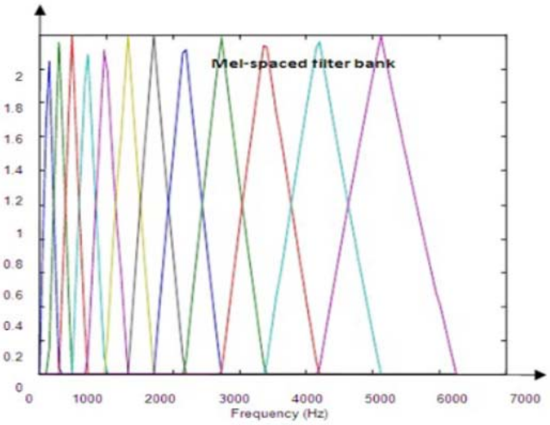

Fig.4: Mel Frequency Cepstral Coefficients(MFCC)

Mel-frequency cepstral coefficient is broadly used for analyzing the inout signal and for power spectrum. To perform a better recognition it combines all observed features to result a valid word by creating maps among the phonetics or syllables it comes in decoding[17]. To connect a feature vector sequence markov modeling is widely used. Viterbi algorithmcan used for the mapping because it is the fast computation for shortest paths[18]. Steps of the viterbi algorithm as shown below:-

1. **Initialization:**

   $\delta_1 (i) = \boldsymbol{\pi}_i b_i(x_1), 1 \leq i \leq N$

2. **Recursion:**

   $\delta t(j) = max_{1 \leq i \leq N}(\delta_{t-1}(i) a_{ij}) b_j(x_t), 2 \leq t \leq T \leq, 1 \leq j \leq N$

3. **Termination:**

   $P^* = max_{i:qt€Qe} \delta_T(i)$, Q_e *is the set of final states of* λ

In Viterbi algorithm there is a every phoneme has numbering to create a map. In the language model, on the basis of sequence of word co-occurrence arrange all words sequencely and make a hypothesis of sentences or phrases[19][20]. It is the post processing of the recognition system. Language and lexical model gives the final output of recognition system from audio to text. Also improves the performance of the system. System is trained for matching the words and arrange it in a sequence.

VI. Proposed Work

For recognition experiment is carried out for input audio.

Audio are taken as source from the internet, the British National Corpus and some are from American corpus. British National Corpus has numerous range of source from which defense driving technique is used from corpus. This audio clip has conversation of 7 hours. There are five steps of the system. First is the input which having audio corpus. Second is the preprocessing in that feature extraction has been done. Third is the acoustic model in that input signal is going to verified for the required terms. Fourth is the language model by which word are sequencely arranged and make a hypothesis of phrases and sentences. That is in the form of readable to the user, it is the last step of text data converted from the speech. Following is the flow of proposed work.

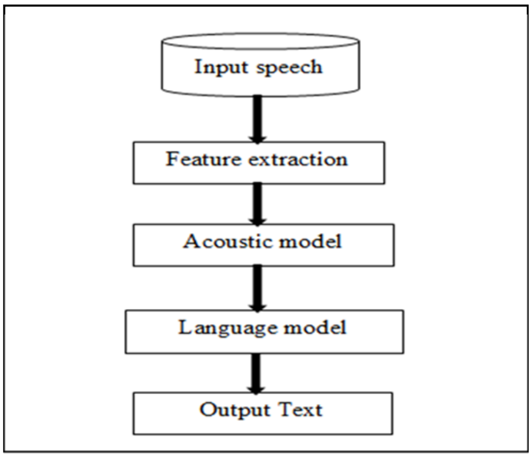

Fig.5: Proposed plan

From the web different topics are taken for experimental analysis and having different size of each audio. All speech audio are in .wav format. Also all audio are in English language of British accent.

Following are the stages for conversion of audio to the text file:

First of all Acoustic data is to be entered for processing. There are three main phases of the speech recognition system as feature extraction, language module and acoustic data. After entering an acoustic data it executes to the first stage as feature extraction in which input signal analyses and for extracting the features of the input signal. For this Mel-frequency cepstral coefficient is used for analyze the input speech waveform. As the MFCC is very efficient technique has the ability to perform on power spectrum which is calculated by using fourier analysis.By analyzing it shows the best resolution of time than the fourier transform also it allows high frequency. The most important phase is acoustic model and language model after analyse the input audio signal it required to create a mapping between the phoneme to generates a valid word of hypothesis phrases. For the fast computing of shortest path to create a map Viterbi algorithm is used in HMM. The result of 10 audio clip and ASR accuracy is shown below:

VII. Experimental Results

As shown in the following table ASR accuracy is given according to each recorded speech. It is the comparison between the original text and output.

TABLE II: ASR ACCURACY OF AUDIO SAMPLES

Sr.no.	Audio	ASR Accuracy
1	Defense Driving techniques	70%
2	Buiseness letters	62%
3	Backbone	69%
4	Cortec	73%
5	British primary school	78%
6	Applied behavior analysis	59%
7	United Kingdom	67%
8	Hong kong	74%
9	York university career service	80%
10	Bioenergetics lecture	77%

As shown in above table, topics are specified with accuracy given in percentage. Accuracy is to be calculated by matching and comparing words of the output of speech recognition system and the original text.

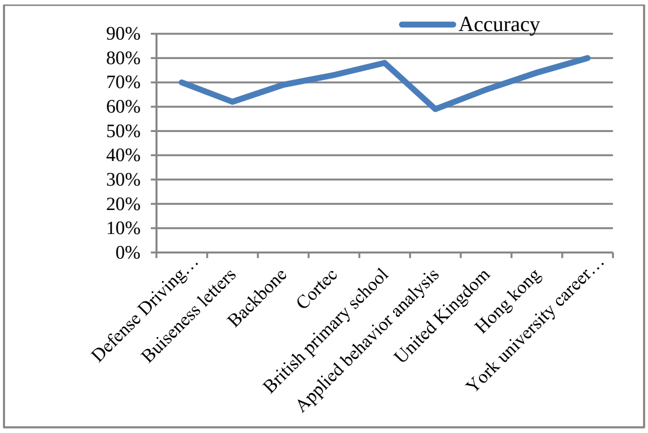

Fig.6: Accuracy of ASR system

As shown in the above fig no.6 it shows graphical structure of accuracy of ten speech in ASR system. Percentage of accuracy is given from 0% to 90% is labeled on y-axis and on x-axis there is a speech topic is given, also blue colored line indicates the percent wise accuracy which was compared with the original text.

VIII. Conclusion

Speech is the most prominent communication between human beings. Automatic speech recognition system converts audio input into the text file. In this paper, the concept of ASR and various techniques are used for the system. Also focuses on the developing and creating system that should be robust with issues with variability in speaker characteristics, language characteristics, background noise, lack of vocabulary. In this paper, there are several properties and methods are define for the feature exraction such as LPC, MFCC,LDA,PCA. In the feature extraction MFCC is used in several applications. HMM is the best technique for the language modeling. And Viterbi algorithm is used to mapping for better result and reliable system. For the desirable result to make a system robust different methods have done. In our experiment accuracy is improved of system. By using Viterbi algorithm it results to speed up a system.

To make a system more reliable vocabulary can be increase of various languages by that system will support more than one language at a time.

References

[1] Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, “A Review on Speech Recognition Technique,” International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[2] Sadaoki Furui, November 2005, 50 years of Progress in speech and Speaker Recognition Research , ECTI Transactions on Computer and Information Technology,Vol.1. No.2.

[3] Vivek Sharma Meenakshi Sharma,R.I.E.I.T Railmajra Punjab India R.I.E.I.T Railmajra Punjab India,” A quantitative study of the automatic speech recognition technique,”International Journal of Advances in Science and Technology (IJAST) Vol I Issue I,December 2013.

[4] Cheong Soo Yee and abdul Manan ahmad, Malay Language Text Independent Speaker Verification using NN-MLP classifier with MFCC, 2008 international Conference on Electronic Design.

[5] B.H.Juang and S.Furui, 2000, Automatic speech recognition and understanding: A first step toward natural human machine communication , Proc.IEEE,88,8,pp.1142-1165.

[6] Giuseppe Riccardi, July 2005,“Active Learning: Theory and Applications to Automatic Speech Recognition”, IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 4.

[7] M.A.Anusuya, S.K.Katti, 2009, “Speech Recognition by Machine: A Review”, International Journal of Computer Science and Information Security, vol. 6, No. 3.

[8] Preeti Saini, Parneet Kaur” Automatic Speech Recognition: A Review “,International Journal of Engineering Trends and Technology- Volume4Issue2- 2013.

[9] F. Sha and L. K. Saul, “Large margin Gaussian mixture modelling for automatic speech recognition,” in Advances in Neural Information Processing Systems,pp. 1249–1256, 2007.

[10] S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarisation systems,” IEEE Transactions Speech and Audio Processing, vol. 14, no. 5,September 2006.

[11] L. K. Saul and M. G. Rahim, “Maximum likelihood and minimum classification error factor analysis for automatic speech recognition,” IEEE Transactions on Speech and Audio Processing, vol. 8, pp. 115–125, 2000.

[12] L. R. Neumeyer, A. Sankar, and V. V. Digalakis, “A comparative study of speaker adaptation techniques,” in Proceedings of Eurospeech, pp. 1127–1130,Madrid, 1995.

[13] E. McDermott, T. J. Hazen, J. Le Roux, A. Nakamura, and K. Katagiri,“Discriminative training for large-vocabulary speech recognition using minimum classification error,” IEEE Transactions on Audio Speech and Language Processing, vol. 15, no. 1, pp. 203–223, 2007.

[14] H. Liao, Uncertainty Decoding For Noise Robust Speech Recognition. PhD thesis, Cambridge University, 2007.

[15] H. Jiang, X. Li, and X. Liu, “Large margin hidden Markov models for speech recognition,” IEEE Transactions on Audio, Speech and Language Processing,vol. 14, no. 5, pp. 1584–1595, September 2006.

[16] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” Journal of Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination,” IEEE Transactions on Speech and Audio Processing,vol. 4, no. 5, pp. 352–359, 1996.

[18] M. J. F. Gales, “Cluster adaptive training of hidden Markov models,” IEEE Transactions on Speech and Audio Processing, vol. 8, pp. 417–428, 2000.

[19] S. S. Chen and R. A. Gopinath, “Model selection in acoustic modelling,” in Proceedings of Eurospeech, pp. 1087–1090, Rhodes, Greece, 1997.

[20] Rabiner L R, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition” Proc. IEEE, vol. 77, 1989, pp. 257-286.

[21] Young, S.: HMMs and Related Speech Recognition Technologies. In: Benesty, J., Sondhi, M.M., Huang, Y. (eds.): Springer Handbook of Speech Processing. Springer-Verlag, Heidelberg, Berlin (2008) 539-583

[22] Leonard, R.: A database for speaker-independent digit recognition. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84., Vol. 9 (1984) 328-331

[23] Bridle, J., Brown, M., Chamberlain, R.: An algorithm for connected word recognition. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82., Vol. 7 (1982) 899-902.

[24] Gauvain, J.L., Chin-Hui, L.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. Speech and Audio Processing, IEEE Transactions on 2 (1994) 291-298

[25] Shipra J. Arora,Rishi Pal Singh.:Automatic Speech Recognition:A Review, International Journal of Computer Applications (0975 – 8887) Volume 60– No.9, December 2012.

[26] M. J. F. Gales, D. Y. Kim, P. C. Woodland, D. Mrva, R. Sinha, and S. E.Tranter, “Progress in the CU-HTK broadcast news transcription system,”IEEE Transactions on Speech and Audio Processing, vol. 14, no. 5, September 2006.