Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.107 – 114

REVIEW ARTICLE

Speech Feature Extraction Techniques: A Review

Shreya Narang[1], Ms. Divya Gupta[2]

[1]CSE Dept., Amity University, Noida, India

[2]CSE Dept., Amity University, Noida, India

[1] shreyan93@gmail.com; [2] dgupta1@amity.edu

ABSTRACT: This paper provides a survey on speech recognition and discusses the techniques and system that enables computers to accept speech as input. This paper shows the major developments in the field of speech recognition. This paper highlights the speech recognition techniques and provides a brief description about the four stages in which the speech recognition techniques are classified. In addition, this paper gives a description of four feature extraction techniques: Linear Predictive Coding (LPC), Mel-frequency cepstrum (MFFCs), RASTA filtering and Probabilistic Linear Discriminate Analysis (PLDA). The objective of this paper is to summarize the feature extraction techniques used in speech recognition system.

Keywords: Automatic Speech Recognition (ASR), Feature Extraction, LPC, MFCCs, RASTA filtering, PLDA

I. INTRODUCTION

Automatic Speech Recognition (ASR) also known as computer speech recognition is a process in which speech signal is converted into a sequence of words, other linguistic units by making use of an algorithm which is implemented as a computer program. The major objective with which ASR works is the development of the techniques and a system that enables the computers to recognize speech as input. [3] In a speech recognition system we convert speech into text in which the text is the output of the speech recognition system which is equivalent to the recognized speech. Speech recognition applications have evolved over the past few years. These applications include voice search, call routing command and control, appliance control by voice, voice dialling, computer aided language learning, robotics and many more. [4] The modern speech recognition systems are based on the HMMs that are the Hidden Markov Models. The main reason why HMM is widely used is that HMM has parameters that can be automatically learned or trained and the techniques that are used for learning are easy and are computationally feasible to use. [1] Many advances have been made is the automatic recognition of speech by machine but we are still unable to develop a machine that understands the human speech by numerous speakers in any kind of environment.

In section II we discuss about the major systems developed in speech recognition over the years. In section III we describe the speech recognition techniques. And in section IV we give the final conclusion and the future scope in speech recognition.

II. DEVELOPMENTS OVER THE YEARS

Speech recognition has been an area of research for about fifty decades. Many developments have been made in this field, but we have still not been able to develop a system which is completely accurate. This section gives the major systems that have been developed by far, over the years. The following table gives the description of the same [5].

Table I: Major Developments over years

Year	Major Developments
2000	CMU Sphinx – Includes a number of speech recognizers that are Sphinx 2, Sphinx 3 and Sphinx 4. They make use of Hidden Markov Model.
2001	RWTH ASR – also known as RASR contains a decoder for speech recognition. Also, this system contains the tools required for acoustic models development that can be used in recognition systems. [2]
2003	Dragon NaturallySpeaking 7 Medical – by ScanSoft which lowered the cost of Health-care by accurate speech recognition.[5]
2008	Siri Inc was founded and Google Voice Search in iPhone was developed [5]
2010	Google Voice Search was launched in android. [5]
2012	Siri launched in Apple iPhone 4s. [5]

III. SPEECH RECOGNITION TECHIQUES

The main objective of a speech recognition system is to have capacity to listen, understand and then after act on the spoken information. A speech recognition system includes four main stages which are further classified as shown in the figure below.

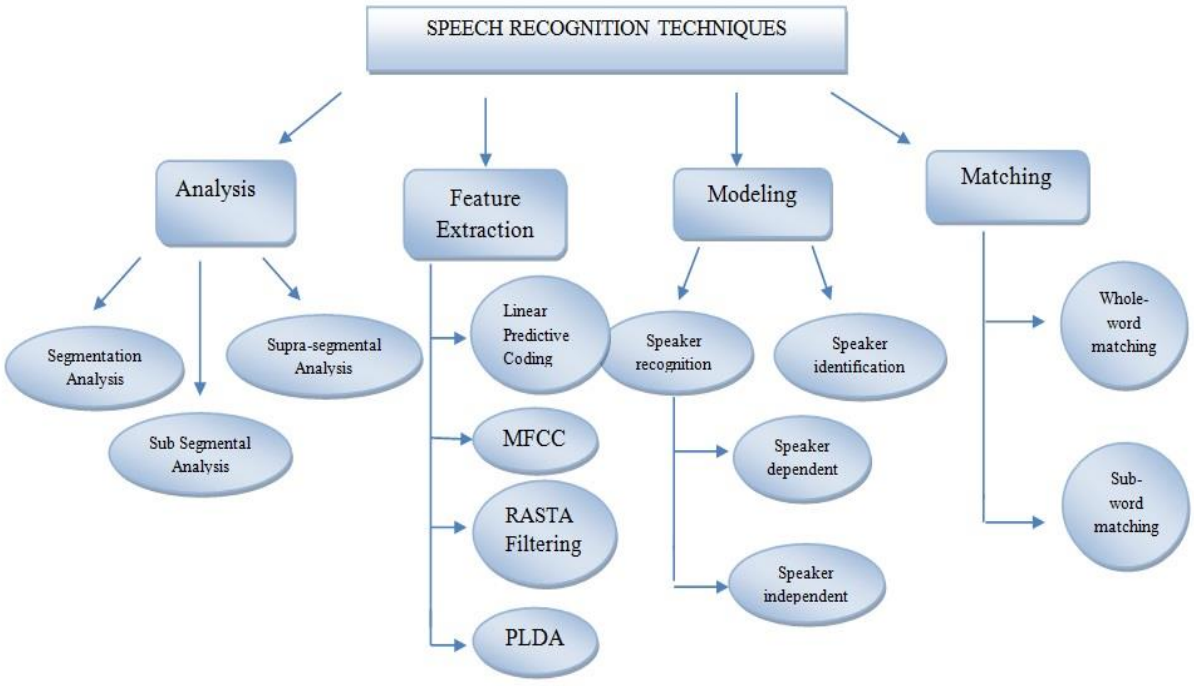

Figure 1: Speech Recognition Technique [1]

1. **Analysis:** The first stage is analysis. When the speaker speaks, the speech includes different types of information that help to identify a speaker. The information is different because of the vocal tract, the source of excitation as well as the behaviour feature. As shown in the figure above, the speech analysis stage can be further classified into three analyses:
    a. *Segmentation Analysis*: In segmentation analysis, the testing to extort the information of speaker is done by utilizing the frame size as well as the shift which is in between 10 to 30 milliseconds (ms) [Range].
    b. *Sub-segmental Analysis:* In this analysis technique, the testing to extract the information of speaker is done by utilizing the frame size as well as the shift which is in between 3 to 5 milliseconds (ms) [Range]. The features of excitation state are analyzed and extracted by using this technique.
    c. *Supra-segmental Analysis:* In Supra-segmental analysis, the analysis to extract the behaviour features of the speaker is done by utilizing the frame size as well as the shift size that ranges in between 50 to 200 milliseconds. [1]
2. **Feature Extraction Technique**: Feature extraction is the main part of the speech recognition system. It is considered as the heart of the system. The work of this is to extract those features from the input speech (signal) that help the system in identifying the speaker. Feature extraction compresses the magnitude of the input signal (vector) without causing any harm to the power of speech signal. There are many feature extraction techniques.

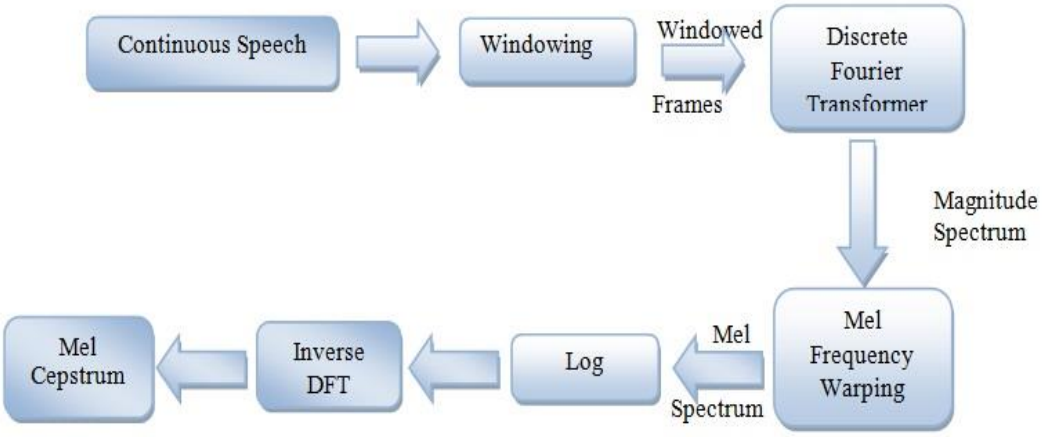

Figure 2: Feature Extraction Diagram [4]

The above figure is the feature extraction diagram. In this, from one side we input the continuous speech signals for the process of windowing. In the process of windowing the disruptions which are present at the start as well as at the end of the frame are minimized. After this process, the continuous speech signal is converted into windowed frames. These windowed frames are passed into the discrete Fourier transformer which converts the windowed frames into magnitude spectrum. Now in the next step, spectral analysis is done with a fixed resolution along a subjective frequency scale that is the Mel-frequency scale which produces a Mel-spectrum. This spectrum is then passed to Log and then to inverse of discrete Fourier transform which produces the final result as Mel-Cepstrum. The Mel-Cepstrum consists of the features that are required for speaker identification. A few feature extraction techniques include:

a. *Linear Predictive coding*: LPC is a tool which is used for speech processing. LPC is based on an assumption: In a series of speech samples, we can make a prediction of the n^{th} sample which can be represented by summing up the target signal's previous samples (k). The production of an inverse filter should be done so

that is corresponds to the formant regions of the speech samples. Thus the application of these filters into the samples is the LPC process.[7] The following table briefly describes this technique:

Table II: Advantages and Disadvantages of Linear Predictive Coding:

Technique	Characteristics	Advantages	Disadvantages
LINEAR PREDICTIVE CODING	• Provides auto-regression based speech features.[8] • Is a formant estimation technique • A static technique.[1] • The residual sound is very close to the vocal tract input signal.[7]	• Is a reliable, accurate and robust technique for providing parameters which describe the time-varying linear system which represent the vocal tract. [9] • Computation speed of LPC is good and provides with accurate parameters of speech. • Useful for encoding speech at low bit rate.	• Is not able to distinguish the words with similar vowel sounds [10]. • Cannot represent speech because of the assumption that signals are stationary and hence is not able to analyze the local events accurately. • LPC generates residual error as output that means some amount of important speech gets left in the residue resulting in poor speech quality.

b. *Mel-frequency cepstrum (MFFCs):* Mel Frequency Cepstral Coefficients are based on the known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. The main purpose of the MFCC processor is to copy the behaviour of human ears. The derivation of MFCCs is done by the following steps[11]:

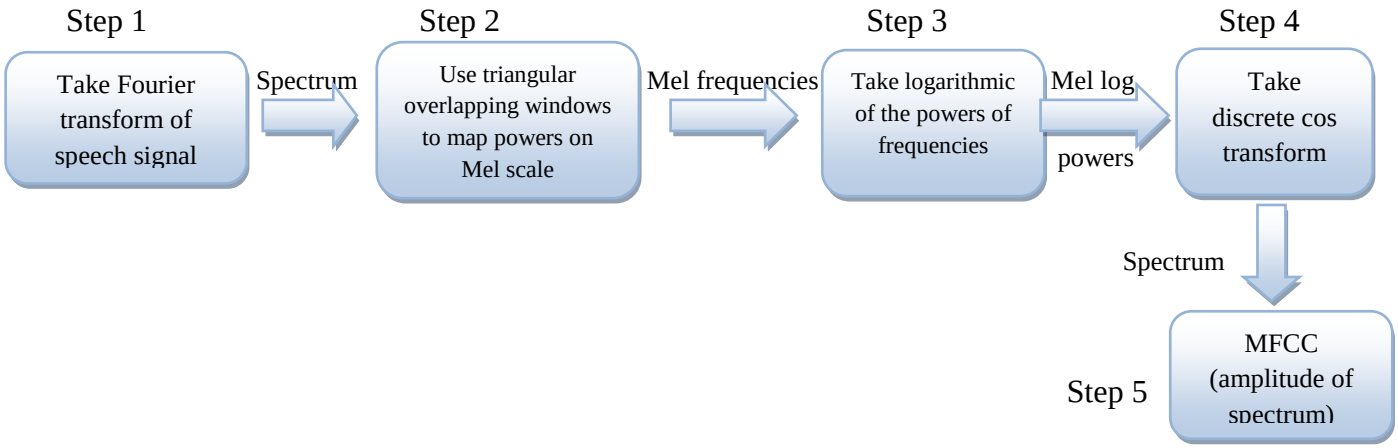

Figure 3: MFCCs Derivation [11]

Table III: Advantages and Disadvantages of Mel-Frequency Cepstrum:

Technique	Characteristics	Advantages	Disadvantages
MEL – FREQUENCY CEPSTRUM (MFCC)	• Used for speech processing tasks. • Mimics the human auditory system • Mel frequency scale: linear frequency spacing below 1000Hz & a log spacing above 1000Hz.	• The recognition accuracy is high. That means the performance rate of MFCC is high. • MFCC captures main characteristics of phones in speech. • Low Complexity.	• In background noise MFCC does not give accurate results. [10] • The filter bandwidth is not an independent design parameter • Performance might be affected by the number of filters.[12]

c. *RASTA filtering*: RASTA is short for RelAtive SpecTrAl. It is a technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the representations of the speech signals are band pass filtered in RASTA. Initially, it was just used to lessen the impact of noise in speech signal but know it is also used to directly enhance the signal [13]. The following figure shows the process of RASTA technique. The main thought here is to subdue the constant factors. [15]

The following table briefly describes this technique:

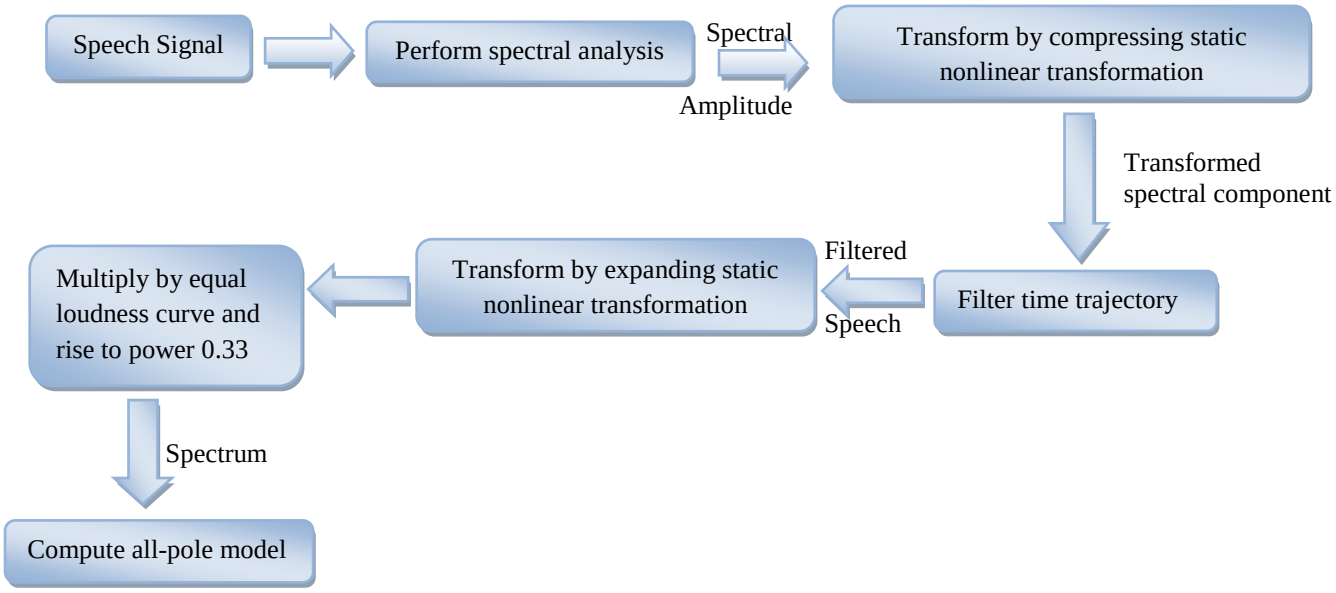

Figure 4: Process of RASTA Technique [15]

Table IV: Advantages and Disadvantages of Relative Spectral:

Technique	Characteristics	Advantages	Disadvantages
RelAtive SpecTrAl (RASTA Filtering)	• Is a band pass filtering technique. • Designed to lessen impact of noise as well as enhance speech. That is, it is a technique which is widely used for the speech signals that have background noise or simply noisy speech.	• Removes the slow varying environmental variations as well as the fast variations in artefacts.[14] • This technique does not depend on the choice of microphone or the position of the microphone to the mouth, hence it is robust. [15] • Captures frequencies with low modulations that correspond to speech. [16]	• This technique causes a minor deprivation in performance for the clean information but it also slashes the error in half for the filtered case. [15] RASTA combined with PLP gives a better performance ratio. [16]

d. *Probabilistic Linear Discriminate Analysis (PLDA):* This technique is an extension for linear probabilistic analysis (LDA). Initially this technique was used for face recognition but now it is used for speech recognition. The following table briefly describes this technique[17]:

Table V: Advantages and Disadvantages of Probabilistic Linear Discriminate Analysis:

Technique	Characteristics	Advantages	Disadvantages
Probabilistic Linear Discriminate Analysis (PLDA)	• Based on i-vector extraction. The i-vector is one which is full of information and is a low dimensional vector having fixed length. • This technique uses the state dependent variables of HMM. • PLDA is formulated by a generative model.	• Is a flexible acoustic model which makes use of variable number of interrelated input frames without any need of covariance modelling.[17] • High recognition accuracy	• The Gaussian assumption which are on the class conditional distributions. This is just an assumption and is not true actually. • The generative model is also a disadvantage. The objective was to fit the date which takes class discrimination into account.[18]

3. **Modeling Techniques**: The goal of the modeling techniques is to produce speaker models by making use of the features extracted (feature vector). As shown in the figure the modeling techniques are further categorized into speaker recognition & identification. Speaker recognition can be further classified into speaker dependent and speaker independent. Speaker identification is a process in which the system is able to identify who the speaker is on the basis of the extracted information from the speech signal. In speech recognition process we can use the following modeling approaches:
    a. *Acoustic-Phonetic approach*: The basic principle that this approach follows is identifying the speech signals and then providing these speech signals with apt labels to these signals. Thus the acoustic phonetic approach postulates that there exists finite number of phonemes of a language which can be commonly described by acoustic properties.
    b. *Pattern recognition approach*: It involves two steps: Pattern Comparison and Pattern Training. It is further classified into Template Based and Stochastic approach. This approach makes use of robust mathematical formulas and develops speech pattern representations.
    c. *Dynamic Time Warping (DTW):* DTW is an algorithm which measures whether two of the sequences are similar that vary in time or even in speed. A good ASR system should be able to handle the different speeds of different speakers and the DTW algorithm helps with that. It helps in finding similarities in two given data keeping in mind the various constraints involved.
    d. *Artificial Intelligence Approach (AI):* In this approach, the procedure of recognition is developed in the same way as a person thinks, evaluates (or analyzes) and thereafter makes a decision on the basis of uniform acoustic features. This approach is the combination of acoustic phonetic approach and pattern approach. [1]
4. **Matching Techniques**: The word that has been detected is used by the engine of speech recognizer to a word that is already known by making use of one of the following techniques:
    a. *Sub word matching*: Phonemes are looked up by the search engine on which the system later performs pattern recognition. These phonemes are the sub words thus the name sub word matching. The storage that is required by this technique is in the range 5 to 20 bytes per word which is much less in comparison to whole word matching but it takes a large amount of processing.
    b. *Whole word matching:* In this matching technique there exists a pre-recorded template of a particular word according to which the search engine matches the input signal. The processing that this technique takes is less in comparison to sub word matching. A disadvantage that this technique has is that we need to record each and every word that is to be recognized beforehand in order for the system to recognize it and thus it can only be used when we know the vocabulary of recognition beforehand. Also these templates need storage that ranges from 50 bytes to 512 bytes per word which very large as compared to sub word matching technique. [1]

IV. CONCLUSION AND FUTURE SCOPE

There has been a lot of research in the field of speech recognition but still the speech recognition systems till date are not a hundred percent accurate. The systems developed so far have limitations: there are a limited number of vocabularies in the current systems and we need to work towards expanding this vocabulary, there exists a problem of overlapping speech that is

the systems cannot identify speech from multiple users, the user needs to be in a place which is background noise free for an accurate recognition, there occurs a problem with the accent and the pronunciation of the user or speaker. In the future the speech recognition systems need to be free of these limitations to give hundred percent results. In this paper we firstly attempt to show the major systems developed under speech recognition over the years. We then give a brief description of speech recognition techniques. A speech recognition system should include the four stages: Analysis, Feature Extraction, Modelling and Matching techniques as described in the paper. Also, through this paper we show four techniques used in feature extraction: Linear Predictive Coding, Mel-frequency cepstrum, Relative Spectral and Probabilistic Linear Discriminate Analysis. By studying each of these techniques we conclude that they have their own advantages and disadvantages and all of them are being used for different purposes. Through research we conclude the Mel frequency cepstrum is a feature extraction technique that is used widely for many speech recognition systems as it is able to mimic the human auditory system and it gives a better performance rate.

REFERENCES

[1] Santosh K.Gaikwad and Pravin Yannawar, A Review, International Journal of Computer Applications ***A*** Review on Speech Recognition Technique Volume 10– No.3, November 2010

[2] Rybach, D.; C. Gollan; G. Heigold; B. Hoffmeister; J. Lööf; R. Schlüter; H. Ney (September 2009). "The RWTH Aachen University Open Source Speech Recognition System". Interspeech-2009: 2111–2114.

[3] Sanjivani S. Bhabad Gajanan K. Kharate International Journal of Advanced Research in Computer Science and Software Engineering **,** An Overview of Technical Progress in Speech Recognition Volume 3, Issue 3, March 2013

[4] Wiqas Ghai and Navdeep Singh International Journal of Computer Applications (0975 – 8887) a Literature Review on Automatic Speech Recognition, Volume 41– No.8, March 2012.

[5] Melanie Pinola (2011-11-02). "Speech Recognition Through the Decades: How We Ended Up With Siri". www.techhive.com.

[6] Anil K.Jain, et.al., Statistical Pattern Recognition: A Review , IEEE Transactions on Pattern Analysis and Machine Intelligence ,Vol.22, No.1, January 2000.

[7] Celso Auguiar, in CCRMA - Center for Computer Research in Music and Acoustics. Stanford University on Modelling the Excitation Function to Improve Quality in LPC's Resynthesis**.**

[8] Tomyslav Sledevic,˘ Artu¯ras Serackis, Gintautas Tamulevici˘ us, Dalius Navakauskas, International Journal of Electrical, Computer, Electronics and Communication on Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System Vol:7 No:12, 2013

[9] Shanthi Therese Chelpa Lingam, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) a Review of Feature Extraction Techniques in Automatic Speech Recognition, Volume No.2, Issue No.6, pp : 479-484 1 June 2013

[10] Navnath S Nehel and Raghunath S Holambe Journal on Audio, Speech, and Music Processing, on DWT and LPC based feature extraction methods for isolated word recognition, 2012

[11] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication 54 (4): 543–565. doi:10.1016/j.specom.2011.11.004.

[12] Nilu Singh, R.A Khan and Raj Shree, International Journal of Computer Applications (0975 – 8887) , A Comparative Study on MFCC and Prosodic Feature Extraction Techniques:, Volume 54– No.1, September 2012

[13] Hynek Hermansky , Eric A. Wan, and Carlos Avendano, Oregon Graduate Institute of Science & Technology Department of Electrical Engineering and Applied Physics, Speech enhancement based on temporal processing.

[14] Chia-Ping Chen Jeff Bilmes and Daniel P. W. Ellis, Department of Electrical Engineering University of Washington Seattle, WA on Speech Feature Smoothing for Robust ASR

[15] H. Hermansky and N. Morgan, Rasta processing of speech, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 578{589, 1994.

[16] Yuxuan Wang, Kun Han, and DeLiang Wang, Fellow, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, Exploring Monaural Features for Classification-Based Speech Segregation, 2012.

[17] Liang Lu, Member, IEEE and Steve Renals, Fellow, IEEE, IEEE SIGNAL PROCESSING LETTERS, Probabilistic Linear Discriminant Analysis for Acoustic Modelling, VOL. X, NO. X, 2014

[18] Jun Wang, Dong Wang, Ziwei Zhu, Thomas Fang Zheng and Frank Soong, at Center for Speaker and Language Technologies (CSLT), on I-vectors, a Discriminative Scoring for Speaker Recognition Based, 2014

