International Journal of Engineering Research and General Science Volume 2, Issue 3, April-May 2014
ISSN 2091-2730

A Comparative Study on Feature Extraction Techniques for Language Identification

Varsha Singh[1], Vinay Kumar Jain[2], Dr. Neeta Tripathi[3]

[1]Research Scholar, Department of Electronics & Telecommunication, CSVTU University

[2] Associate Professor, Department of Electronics & Telecommunication, CSVTU University

[3]Principal, SSITM, CSVTU Univeristy, FET, SSGI, SSTC jumwani Bhilai, C. G, India

E-mail- varshasingh.40@gmail.com

ABSTRACT— This paper presents a brief survey of feature extraction techniques used in language identification (LID) system. The objective of the language identification system is to automatically identify the specific language from a spoken utterance. Also the LID system must perform quickly and accurately. To fulfill this criteria the extraction of the features of acoustic signals is an important task because LID mainly depends on the language-specific characteristics. The efficiency of this feature extraction phase is important since it strongly affects the performance and quality of the system. There are different features which are used in LID are cepstral coefficients, MFCC, PLP, RASTA-PLP, etc.

Keywords— LID (Language Identification), feature extraction, LPC, Cepstral analysis, MFCC, PLP, RASTA-PLP.

INTRODUCTION

The Speech is an important and natural form of communication with others. Over the past three decades there is the tremendous development in the area of speech processing. Applications of speech processing include speech/ speaker recognition, language identification etc. The objective of the automatic speaker recognition system is to extract, characterize and recognize the information about speaker identity [1]. Language identification system automatically identifies the specific language from a spoken utterance. Automatic language identification is therefore an essential component of, and usually the first gateway in, a multi-lingual speech communication/interaction scenario. There are many potential applications of LID. In the area of telephone-based information services, including customer service, phone banking, phone ordering, information hotline and other call-centre/Interactive Voice Response (IVR) based services; LID systems would be able to automatically transfer the incoming call to the corresponding agent, recorded message, or speech recognition system. LID system can be made efficient by extracting the language-specific characteristics. In this paper we mainly focus on the language specific characteristics for language identification system. Spectral features are those features that characterize the short-time spectrum and based on the time-varying properties of the speech signal. Temporal features are assumed constant over a short period and its characteristics are short-time stationary.

LITERATURE REVIEW

Feature Extraction is a process of reducing data while retaining speaker discriminative information. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data [2]. We can define requirement that should be taken into account during selection of the appropriate speech signal characteristics of features [3, 4]:

- large between-speaker and small within-speaker variability
- not change over time or be affected by the speaker's health
- be difficult to impersonate/mimic
- not be affected by background noise nor depend on the specific transmission medium
- Occur naturally and frequently in speech.

It is not possible that a single feature would meet all the criteria listed above. Thus, a large number of features can be extracted and combined to improve the accuracy of the system.

The pitch and formant features of speech signal are extracted and used to detect the three different emotional states of a person [5]. Pitch is originates from the vocal cords. When air flow from glottal through the vocal cords, the vibration of vocal cords/folds produces pitch harmonics. The rate at which the vocal folds vibrate is the frequency of the pitch. So, when the vocal folds oscillate at 300 times per second, they are said to be producing a pitch of 300 Hz. Pitch is useful to differentiate speaker genres. In males, the average pitch falls between 60 and 120 Hz, and the range of a female's pitch can be found between 120 and 200 Hz [2]. The Cepstral analysis method is used for pitch extraction and LPC analysis method is used to extract the formant frequencies**.** Formants are defined as the spectral peaks of sound spectrum, of the voice, of a person. In speech science and phonetics, formant frequencies refer to the acoustic resonance of the human vocal tract. They are often measured as amplitude peaks in the frequency spectrum of the sound wave. Formant frequencies are very much important in the analysis of the emotional state of a person. The Linear predictive coding technique (LPC) has been used for estimation of the formant frequencies [5].

LPC is one of the feature extraction method based on the source-filter model of speech production. B.S. Atal in 1976 [3] uses linear prediction model for parametric representation of speech derived features. The predictor coefficients and other speech parameters derived from them, such as the impulse response function, the auto-correlation function, the area function, and the cepstrum function were used as input to an automatic speaker recognition system, and found the cepstrum to provide the best results for speaker recognition.

Reynolds in 1994 [6] compared different -features useful for speaker recognition, such as Mel frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), LPCC (linear predictive cepstral coefficients) and perceptual linear prediction cepstral coefficients (PLPCCs). From the experiments conducted, he had concluded that, of these features, MFCCs and LPCCs give better performance than the other features. Revised perceptual linear prediction was proposed by Kumar et al. [7], Ming et al. [8] for the purpose of identifying the spoken language; Revised Perceptual Linear Prediction Coefficients (RPLP) was obtained from combination of MFCC and PLP.

Of all the various spectral features, MFCC, LPCC and PLP are the most recommended features which carry information about the resonance properties of vocal tract [9].

METHODOLOGY

In this section a comprehensive review of several methods for feature extraction are presented for language identification.

LPC: It is one of the important method for speech analysis because it can provide an estimate of the poles (hence the formant frequency- produced by vocal tract) of the vocal tract transfer function. LPC (Linear Predictive Coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering and the remaining signal is called the residue [1]. The basic idea behind LPC coding is that each sample can be approximated as a linear combination of a few past samples. The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters. The computation involved in LPC processing is considerably less than cepstrum analysis.

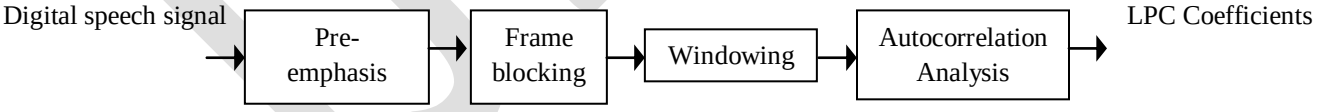

Fig. 1 Block diagram of LPC algorithm

Cepstral Analysis: This analysis is a very convenient way to model spectral energy distribution. Cepstral analysis operates in a domain in which the glottal frequency is separated from the vocal tract resonances. The low order coefficients of the cepstrum contain information about the vocal tract, while the higher order coefficients contain primarily information about the excitation. (Actually, the higher order coefficients contain both types of information, but the frequency of periodicity dominates). The word cepstrum was derived by reversing the first syllable in the word spectrum. The cepstrum exists in a domain referred to as quefrency (reversal of the first syllable in frequency) which has units of time. The cepstrum is defined as the inverse Fourier transform of the logarithm of the power spectrum. The Cepstrum is the Forward Fourier Transform of a spectrum. It is thus the spectrum of a spectrum, and has certain properties that make it useful in many types of signal analysis [10]. Cepstrum coefficients are calculated in short frames over time. Only the first M cepstrum coefficients are used as features (all coefficients model the precise spectrum, coarse spectral shape is

modeled by the first coefficients, precision is selected by the number of coefficients taken, and the first coefficient (energy) is usually discarded). The cepstrum is calculated in two ways: LPC cepstrum and FFT cepstrum. LPC cepstrum is obtained from the LPC coefficients and FFT cepstrum is obtained from a FFT. The most widely parametric representation for speech recognition is the FFT cepstrum derived based on a Mel scale [11]. A drawback of the cepstral coefficients: linear frequency scale. Perceptually, the frequency ranges 100–200Hz and 10 kHz 20 kHz should be approximately equally important. The standard cepstral coefficients do not take this into account. Logarithmic frequency scale would be better. Mimic perception is necessary because typically we want to classify sounds according to perceptual dissimilarity or similarity; perceptually relevant features often lead to robust classification, too. It is desirable that small change in feature vector leads to small perceptual change (and vice versa). The Mel-frequency cepstral coefficients fulfill this criterion.

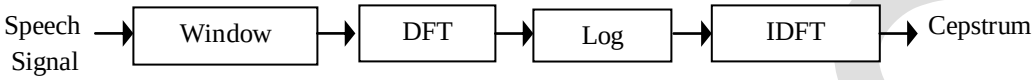

Fig. 2 Cepstral analysis

MFCC: This technique is considered as one of the standard method for feature extraction and is accepted as the baseline. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale (the Mel scale was used by Mermelstein and Davis [11] to extract features from the speech signal for improving the recognition performance). MFCC are the results of the short-term energy spectrum expressed on a Mel-frequency scale [1]. The MFCCs are proved more efficient better anti-noise ability than other vocal tract parameters, such as LPC. Various steps to calculate MFCC are shown in the figure below:

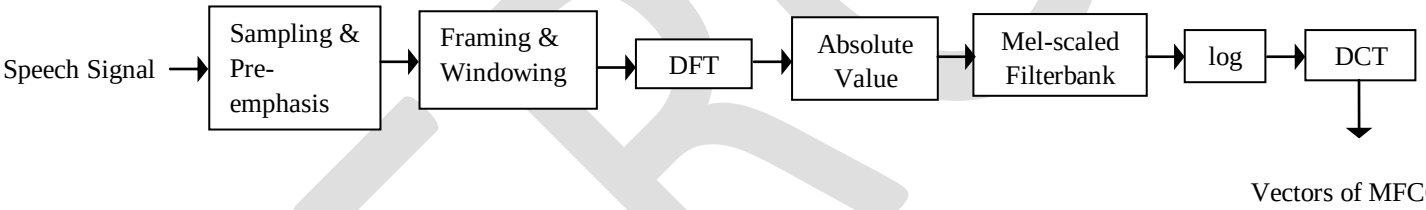

Fig. 3 Block diagram of MFCC processor

LFCC speech features (LFCC-FB40): The methodology of LFCC [11] is same as MFCC. The only difference is that the Mel-frequency filter bank is replaced by linear-frequency filter bank.. Thus, the desired frequency range is implemented by a filter-bank of 40 equal-width and equal-height linearly spaced filters. The bandwidth of each filter is 164 Hz, and the whole filter-bank covers the frequency range [133, 6857] Hz.

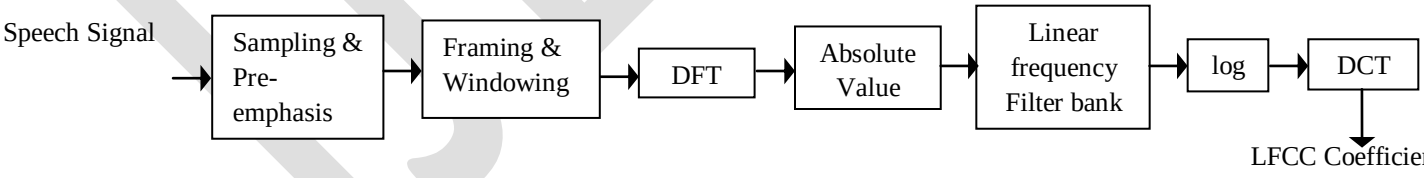

Fig. 4 LFCC Implementation

HFCC-E of Skowronsky & Harris: Skowronski & Harris [12] introduced the Human Factor Cepstral Coefficients (HFCC-E). In the HFCC-E scheme the filter bandwidth is decoupled from the filter spacing. This is in contrast to the earlier MFCC implementations, where these were dependent variables. Another difference to the MFCC is that in HFCC-E the filter bandwidth is derived from the equivalent rectangular bandwidth (ERB), which is based on critical bands concept of Moore and Glasberg's expression rather than on the Mel scale [11]. Still, the centre frequency of the individual filters is computed by utilizing the Mel scale. Furthermore, in HFCC-E scheme the filter bandwidth is further scaled by a constant, which Skowronski and Harris labelled as E-factor. Larger values of the E-factor E= {4, 5, 6} were reported [12] to contribute for improved noise robustness.

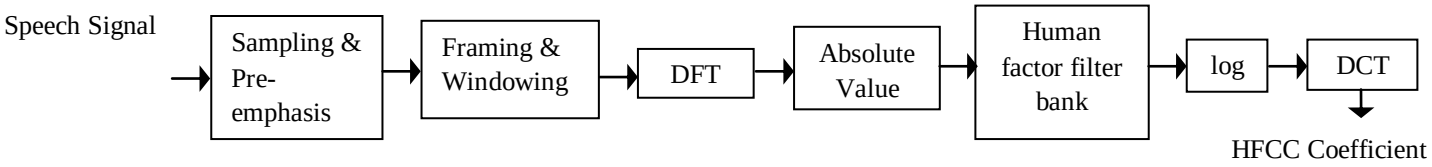

Fig. 5 HFCC Implementation

PLP: The Perceptual Linear Predictive (PLP) speech analysis technique is based on the short-term spectrum of speech. PLP is a popular representation in speech recognition, and it is designed to find smooth spectra consisting of resonant peaks [13]. PLP parameters are the coefficients that result from standard all-pole modeling [14] which is effective in suppressing speaker-specific details of the spectrum. In addition, the PLP order is smaller than is typically needed by LPC-based speech recognition systems. PLP models the human speech based on the concept of psychophysics of hearing [13]. In PLP the speech spectrum is modified by a set of transformations that are based on models of the human auditory system. The PLP computation steps are critical-band spectral-resolution, the equal-loudness hearing curve and the intensity-loudness power law of hearing. Once the auditory-like spectrum is estimated, it is converted to autocorrelation values by doing a Fourier transform. The resulting autocorrelations are used as input to a standard linear predictive analysis routine, and its output is perceptually-based linear prediction coefficients. Typically, these coefficients are then converted to cepstral coefficients via a standard recursion [14].

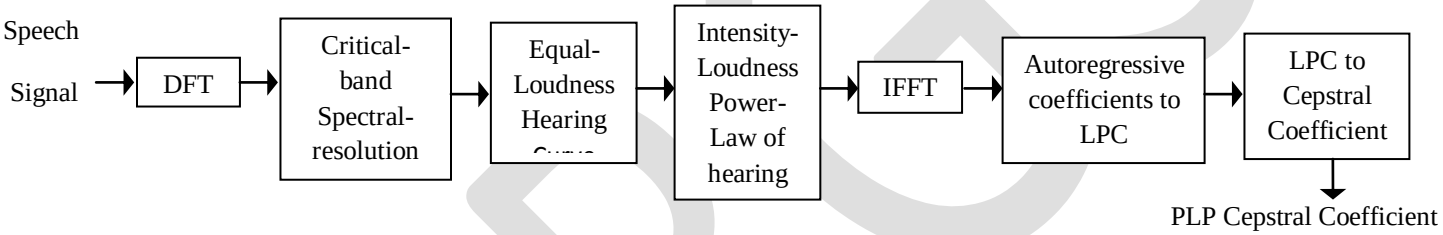

Fig. 6 PLP Implementation

RASTA-PLP: A popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform – Perceptual Linear Prediction. PLP was originally proposed by H. Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information [13]. The term RASTA comes from the words RelAtive SpecTrA. RASTA filtering is often coupled with PLP for robust speech recognition. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency sub band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g. from a telephone line [15]. In essence, RASTA filtering serves as a modulation-frequency band pass filter, which emphasizes the modulation frequency range most relevant to speech while discarding lower or higher modulation frequencies.

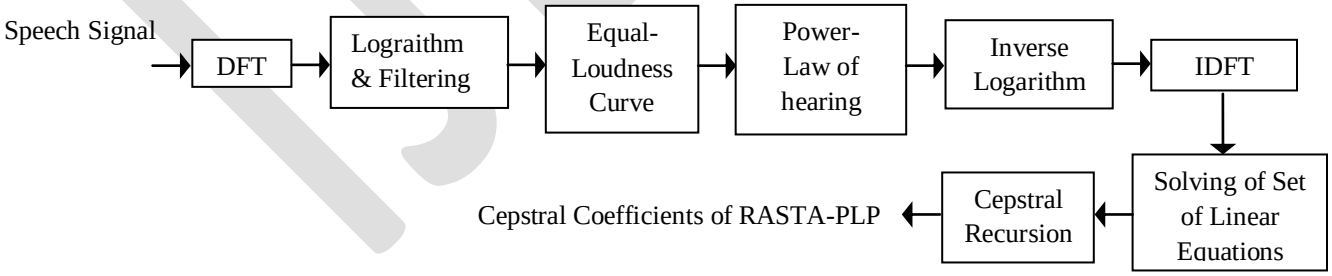

Fig. 7 RASTA-PLP Model

CONCLUSION

MFCC, PLP and LPC are the most proposed acoustic features used in language identification. The accuracy and speed of LID system is enhanced by combining more features of speech signal. In the following table some important conclusion has been made of above discussed feature extraction technique.

Table No.1 Showing the concluding highlights of the different types of feature extraction methods

S. No.	Method	Property	Comments
1.	Linear Predictive Coding	Static feature extraction method, 10 to 16 lower order coefficient.	The LP algorithm is a practical way to estimate formant of the speech signal especially at high frequencies. It is used for feature extraction at lower order.
2.	Cepstral Analysis	Static feature extraction method, power spectrum.	The Cepstrum is a practical way to extract the fundamental frequency of the speech signal. The Cepstral algorithm shows some limitations in the localization of formants especially at high frequencies.
3.	Mel Frequency Cepstral Coefficients	It is the result of short-term energy spectrum and expressed on Mel-scale which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.	The MFCC reduces the frequency information of the speech signal into a small number of coefficients. It is easy and relatively fast to compute.
4.	Linear Frequency Cepstral Coefficients	Uses a bank of equal bandwidth filters with linear spacing of the central frequencies.	The equal bandwidth of all filters renders unnecessary the effort for normalization of the area under each filter.
5.	Human Factor Cepstral Coefficients	Uses Moore and Glasberg's expression for critical bandwidth (ERB), a function only of center frequency, to determine filter bandwidth.	Larger values of the E-factor contribute for improved noise robustness.
6.	Perceptual Linear Predictive Analysis	Short term spectrum is modified based on psychophysically based transformation.	Lower order analysis results in better estimates of recognition parameters for a given amount of training data.
7.	RASTA-PLP	Applies a band pass filter to each spectral component in the critical-band spectrum estimate.	These features are best used when there is a mismatch in the analog input channel between the development and fielded systems.

REFERENCES:

[1] Vibha Tiwari, "MFCC and its applications in speaker recognition", International Journal on Emerging Technologies 1(1): 19-22(2010).
[2] Premakanthan P. and Mikhad W. B., "Speaker Verification/Recognition and the Importance of Selective Feature Extraction: Review", MWSCAS. Vol. 1, 57-61, 2001.
[3] B. S. Atal, "Automatic Recognition of Speakers from their Voices", Proceedings of the IEEE, vol. 64, 1976, pp 460 – 475.
[4] Douglas A. Reynolds and Richard Rose, "Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models", IEEE transaction on Speech and Audio Processing, Vol.3, No.1, January 1995.
[5] Bageshree V. Sathe-Pathak, Ashish R. Panat, "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person", International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.
[6] D.A. Reynolds, "Experimental evaluation of features for robust speaker identification", IEEE Trans. Speech Audio Process. , vol. 2(4), pp. 639-43, Oct. 1994.

[7] Kumar, P., A.N. Astik Biswas and M. Chandra, "Spoken Language identification using hybrid feature extraction methods", J. Telecomm., 1: 11-5, 2010.
[8] Ming, J., T. Hazen, J. Glass and D. Reynolds, "Robust speaker recognition in noisy conditions", IEEE Trans. Audio Speech Language Proc., 15:1711-1723, DOI: 10.1109/TASL.2007.899278, 2007.
[9] Hassan Euaidi and Jean Rouaf, "Pitch and MFCC dependent GMM models for speaker identification systems", CCECE IEEE, 2004.
[10] Childers, D.G., Skinner, D.P., Kemerait, R.C., "The cepstrum: A guide to processing" Proceedings of the IEEE Volume 65, Issue 10, Oct. 1977 Page(s):1428 – 1443.
[11] Merlmestein P. and Davis S., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. On ASSP, Aug, 1980. pp. 357-366.
[12] Skowronski, M.D., Harris, J.G., "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition", J. Acoustic Soc. Am., 116(3):1774–1780, 2004.
[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech", J. Acoustic Soc. Am., pp. 1738-1752, 1990.
[14] L. Rabiner and R. Schafer, "Digital Processing of Speech Signals", Prentice Hall, Englewood Cliffs, NJ, 1978.
[15] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Trans. On Speech and Audio Processing, Vol. 2, 578-589, Oct. 1994.