International Journal Of Scientific Research And Education
||Volume||2||Issue||11||Pages-2313-2321||November-2014|| ISSN (e): 2321-7545
Website: http://ijsae.in

RASTA-PLP For Speech Recognition of Articulatory Handicapped People

Author
Prof. Sanjivani S. Bhabad[1] , Kamaraj Naidu[2]
[1]Associate Professor, Department of E & TC Engineering K. K. Wagh Institute of Engg. Education and Research, Nasik, India
[2]PG Student, Department of E & TC Engineering, K. K. Wagh Institute of Engg. Education and Research, Nasik, India
E.mail-ssb.eltx@gmail.com, ksnaidu23@gmail.com, kamaraj.naidu@cgglobal.com

ABSTRACT
This paper describes an approach of isolated word recognition for normal people and articulatory handicapped people using relative spectral and cepstral perceptual linear predictive (RASTA-PLP) feature extraction techniques. Recognition is carried out using a minimum distance classifier. The results of RASTA-PLP Cepstral coefficients and RASTA-PLP spectral coefficients are compared. The result for normal persons shows that the recognition accuracy is 75.11% from features of RASTA-PLP Cepstral coefficients as compared to 62.11% calculated from RASTA-PLP spectral coefficients. The result for articulatory handicapped persons shows that the recognition accuracy is 45.60% from features of RASTA-PLP Cepstral coefficients as compared to 38.50% calculated from RASTA-PLP spectral coefficients.

Index Terms— MATLAB, classifier, Relative Spectral Perceptual Linear Predictive (RASTA-PLP)

INTRODUCTION
SPEECH recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. One of the most promising aspects of automatic speech recognition is that the potential for hands-free interaction with machines. Feature extraction is the most basic and important part any automatic word recognition process. There are different types of feature extraction methods like Linear predictive coding (LPC), Mel frequency cepstral coefficients (MFCC), Relative spectral analysis (RASTA), Perceptual linear predictive (PLP). Every feature extraction technique has got the advantage and disadvantages associated with it. In this novel technique we shall combine the RASTA and PLP feature extraction method to create and hybrid model RASTA-PLP.

PLP parameters are the coefficients that result from standard all-pole modeling, or linear predictive analysis, of a specially modified, short-term speech spectrum. In PLP the speech spectrum is modified by a set of transformations that are based on models of the human auditory system. The spectral resolution of human hearing is roughly linear up to 800 or 1000 Hz, but it decreases with increasing frequency above this linear

range. PLP incorporates critical-band spectral-resolution into its spectrum estimate by remapping the frequency axis to the Bark scale and integrating the energy in the critical bands to produce a critical-band spectrum approximation. At conversational speech levels, human hearing is more sensitive to the middle frequency range of the audible spectrum. PLP incorporates the effect of this phenomenon by multiplying the critical-band spectrum by an equal loudness curve that uppresses both the low- and high-frequency regions relative to the midrange from 400 Hz to 1200 Hz.

The term RASTA comes from the words RelAtive SpecTrA. The RASTA technique applies a bandpass filter to each spectral component in the critical band spectrum estimate. This filtering emphasizes frame-to-frame spectral changes that occur between the rates of 1 to 10 Hz. Before applying the bandpass filter, log-RASTA takes the natural logarithm of each spectral component. This logarithm converts multiplicative distortions in the frequency domain into an additive distortion, which can be filtered. Conversion to the

This paper describes the hybrid model combining the RASTA and PLP technique, to create a new model RASTA-PLP with spectral & Cepstral coefficients for feature extraction. The paper mainly consist of three main sections namely section two, section three and section four.

The section two of the paper gives the details of feature extraction using RASTA-PLP hybrid model. It also describes the basics of minimum distance classifier, which is used to recognize word.

The section three illustrates the creation of database in a noiseless room. The section four presents results& conclusion. The result shows that the recognition accuracy is more when RASTA-PLP Cepstral features are used as compared to RASTA-PLP spectral features.

feature extraction techniques

Feature Extraction

RASTA-PLP:

When we hear a sound, it is perceived by our human ear. This perceptual property of human ear is captured in this technique. The power spectrum of the speech signal is converted to Bark scale which is similar to human ear's perceptual model. Perceptual linear prediction (PLP) and MFCC approaches are quiet similar. PLP is combination of DFT and LP techniques. A block schematic for the calculation of PLP is shown in Fig. 1.

The function of each block can be described as follows:

Estimation of power spectrum: The system first computes a power spectrum estimate for the analysis window. Here, the signal is passed though the hamming window and squared magnitude of the FFT is taken.

Critical band analysis: Here, the power spectrum is integrated within the critical band filter response. In the case of PLP, trapezoidal-shaped filters are applied at roughly one-Bark intervals where the bark axis is derived from the formula shown below. The Mel scale filters are triangular filters whereas the Bark scale filters are trapezoidal in shape.

$\Omega(\omega) = 6\ln\left((\omega/1200\pi) + ((\omega/1200\pi)^2 + 1)^{0.5}\right)$

Where Ω represents the angular frequency in Bark scale, and ω represents angular linear frequency=2πf.

This reduces the frequency sensitivity over the original spectrum estimate at high frequencies in particular as the bandwidth is high at high frequencies. The high frequencies are also somewhat emphasized. This simulates the frequency resolution of the ear, which is approximately constant on the bark scale.

Equal loudness pre-emphasis: We need to compensate for the non-equal perception of loudness at different frequencies. Pre-emphasis is executed using an equal-loudness curve given by

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)}{(\omega^2 + 6.3 \times 10^6)(\omega^2 + 0.38 \times 10^9)(\omega^6 + 9.58 \times 10^{26})}$$

Where ω represents angular linear frequency=2πf,

and E (ω) is the energy at frequency ω.

Intensity-loudness power law conversion: In the case of PLP, cube root is taken in the place of log as perceived loudness is approximately the cube root of the intensity. This step can be considered as a reasonable approximation for the speech. This compresses the power spectrum.

Autoregressive Modeling: Apply the IDFT, for PLP, as the log is not computed, the results are more like autocorrelation coefficients. We are taking cube root of the power spectrum instead of the log. Hence, when we compute IDFT, we will not enter the Cepstral domain but we will enter autocorrelation domain. We know that the DFT of autocorrelation is the power spectrum so if we take IDFT of the power spectrum we will get the autocorrelation. An autoregressive model is used to smooth the compressed spectrum. The autoregressive coefficients can be converted to Cepstral variables.

The additional step not show in block diagram can be liftering. In liftering the Cepstral parameters are often multiplied by some simple function such as $n^{\alpha.}$. The value of α is less than 1. Feature vectors computed from PLP provide smooth estimates of the power spectrum.. Many researchers have used the derivative of smooth spectrum. The most common form of measure is the delta spectrum. This is typically implemented as a least-squares approximation to the local slope and is expressed as shown below.

$$\Delta Ci(n) = \frac{\sum_{k=-N}^{N} k\, Ci\,(n+k)}{\sum_{k=-N}^{N} k^2}$$

Where ΔCi represents the delta cepstrum at position n, that is, the first derivative; Ci is cepstrum; and k is the offset. The second derivative commonly referred to as delta-delta cepstrum corresponds to the similar correlation but with parabolic function.

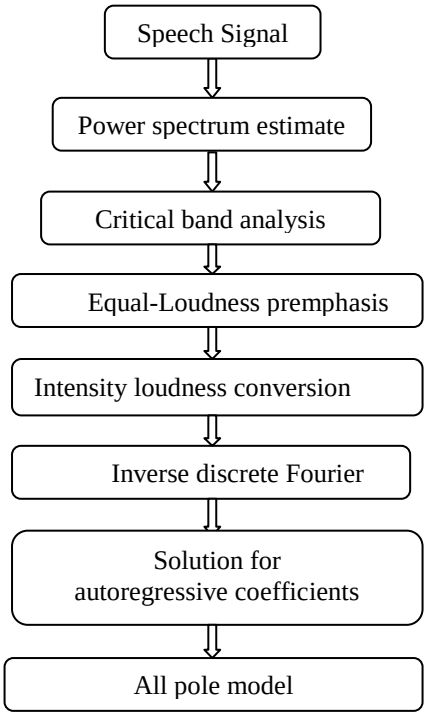

Fig 1: Perceptual Linear prediction (PLP) Block diagram

RelActive SpecTrAI Perceptual Linear prediction (Rasta-PLP):

Let us consider a short-time spectrum S(ω,t) and let it be processed by an LTI filter with transfer function H(ω,t) resulting in a filtered short-time spectrum X(ω,t). The relation can be written as

$$X(\omega,t) = S(\omega,t)\, H(\omega,t)$$

The corresponding log power spectrum can be written as

$$\text{Log}|\, X\,(\omega,t)\,| = \log|\, S(\omega,t)\,|2\, \log|\, H(\omega,t)|2$$

Thus convolution in time domain corresponds to multiplication in the frequency domain and addition in the log power domain.

The spectral analysis is done using the first step of PLP. The log of each critical band trajectory is filtered using band pass filter (BPF)..

The Rasta filter can be used either in the log spectral or Cepstral domain. In effect, the RASTA filter band passes each feature coefficient. Linear channel distortions appear as an additive constant in both the log spectral and Cepstral domain. This can be easily filtered using LTI filters as they are additive.

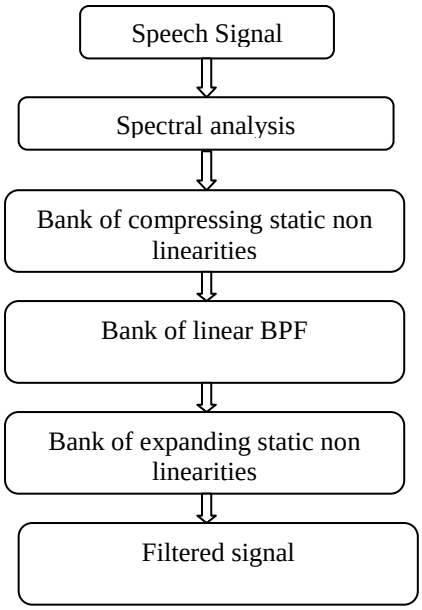

Fig 2: Block diagram of RASTA filtering.

Distance Measure:

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vector {x1, x2 ….xi), and then it is compared with the codebooks from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance[6].The formula used to calculate the Euclidean distance can be defined as following:

The Euclidean distance between two points P = (p1,
p2…pn) and Q = (q1, q2...qn).

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

Database The recording was done in multiple sessions after obtaining due informed consent from the subjects. They were asked to read aloud a set of 11 digits, in a normal manner, repeatedly for 10 times in a sound proof room. The recording was done using microphone type. The RODE NT1 MIC microphone and NUENDO 4 software was used for recording the words. This arrangement was maintained throughout multi-session recording process. The recorded speech was loaded into the computer through a memory card. The data individually stored as Wav files. Database consists of speech samples recorded in a noise proof room. The group consists of 6 normal speakers (3 male and 3 female). From each of them we take speech sample for digits from zero to ten with each digit taken repeated for ten times. Hence making the overall samples to be 660 Nos. For speaker dependent speech recognition we use cross validation process that is we use the

first five samples of each digit from each user to train the system so in all we use 330 samples to train the system. The remaining 330 samples are used as test samples. For the speaker independent system we recorded speech of 10 Articulatory Handicapped persons. They were made to utter words from zero to 10, each word repeated for 10 times.

RESULTS

For speaker dependent system for normal persons, five test samples of each digit from zero to ten from each user is used as test samples. There were total 6 speakers, three male and 3 female speakers. Here Female 1-F1, Female 2-F2, Female 3-F3, Male 1-M1, Male 2-M2, Male 3-M3.

Table I Recognition Results For Male And Female Users Using Rasta-Plp Spectral Method For Normal Speakers

	Correctly recognized words out of 5 samples					
Digit	F1	F2	F3	M1	M2	M3
0	4	4	5	3	4	3
1	2	3	3	4	1	3
2	5	4	2	2	2	3
3	4	4	4	2	2	0
4	4	4	3	3	0	4
5	0	2	2	5	4	3
6	4	5	1	5	0	5
7	4	4	4	2	2	4
8	4	2	4	3	3	3
9	4	1	3	3	2	5
10	3	3	4	3	3	5
Accuracy %	69.09	65.45	63.63	63.63	41.81	69.09

TABLE II Recognition Results For Male And Female Users Using Rasta-Plp Cepstral Method For Normal Speakers

	Correctly recognized words out of 5 samples					
Digit	F1	F2	F3	M1	M2	M3
0	4	4	4	4	3	3
1	3	4	5	4	5	4
2	5	4	3	5	2	4
3	4	4	3	4	3	3
4	5	3	4	5	4	5
5	0	2	4	5	3	5
6	4	4	4	5	2	4
7	5	5	4	3	4	4
8	4	2	4	5	3	3
9	4	3	2	4	5	5
10	3	4	4	3	3	5
Accuracy %	74.54	70.90	74.54	85.45	62.27	81.81

Table III Male Female Recognition Accuracy Results For Normal Speakers

Features RASTA-PLP	Recognition Accuracy In Percentage					
	F 1	F 2	F 3	M 1	M 2	M 3
Spectral	69.09	65.45	63.63	63.63	41.81	69.09
Cepstral	74.54	70.90	74.54	85.45	62.27	81.81

Table III Overall accuracy for normal speakers. Using spectral & cepstral features

Features RASTA-PLP	Recognition accuracy in percentage		Overall accuracy %
	Female speaker	Male speaker	
Spectral	66.05	58.17	62.11
Cepstral	73.72	76.51	75.11

For speaker independent system for Articulatory Handicapped persons 10 test samples of each digit from zero to ten from each user is used as test samples. There were total 10 speakers namely User1-U1, User2-U2, User3-U3, User4-U4, User5-U5, User6-U6, User6-U6, User7-U7, User8-U8, User9-U9, User10-U10.

Table IV Recognition Results For Articulatory Handicapped Speakers Using Rasta Plp Spectral Method

	Correctly recognized words out of 10 samples									
Digit	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
0	8	2	10	0	6	9	2	6	5	8
1	9	9	5	5	3	8	1	5	7	6
2	8	8	10	5	3	9	9	10	5	0
3	4	3	4	2	0	1	6	0	1	4
4	9	8	10	3	1	9	8	9	6	4
5	6	6	10	2	0	3	2	2	0	5
6	0	0	0	0	0	0	0	0	0	4
7	8	2	0	0	0	0	1	0	0	0
8	9	1	2	1	0	0	3	2	0	1
9	9	0	7	8	4	3	3	2	7	3
10	6	3	9	0	1	2	8	0	0	6
Acc %	69	38	61	24	16	40	39	33	28	37

Table V Recognition results for Articulatory Handicapped Speakers using rasta plp cepstral method

	Correctly recognized words out of 10 samples									
Digit	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
0	10	9	10	5	3	10	10	9	9	8
1	10	9	5	4	7	10	6	9	8	1
2	10	9	10	3	3	10	10	6	6	0
3	8	5	9	3	0	0	4	0	0	0
4	10	7	9	9	1	5	5	7	2	1
5	3	7	2	0	4	7	5	1	0	6
6	0	0	0	0	0	0	0	0	6	1
7	6	6	0	1	6	1	7	6	3	5
8	0	3	5	0	4	0	9	1	1	1
9	9	10	8	5	3	9	9	4	10	6
10	7	2	0	0	0	5	9	0	0	1
Acc %	66	61	53	27	28	52	67	39	41	27

Table VI Overall accuracy using spectral & Cepstral features for Articulatory Handicapped Speakers

Features RASTA-PLP	Overall accuracy %
Spectral	38.5
Cepstral	46.2

Table VII Comparision of recognition accuracy of normal and Articulatory Handicapped Speakers

Features RASTA-PLP	Overall accuracy %
Spectral for Normal speakers	62.11
Cepstral for Normal speakers	75.11
Spectral for Articulatory Handicapped Speakers	38.5
Cepstral for Articulatory Handicapped Speakers	46.2

The results show that the overall accuracy of recognition using RASTA PLP spectral features is 62.11% for normal speakers and 38.5% for Articulatory Handicapped Speakers. The result shows that the overall

accuracy of recognition using RASTA PLP Cepstral features is 75.11% for normal speakers and 46.2% for Articulatory Handicapped Speakers.

CONCLUSIONS

We developed hybrid model of feature extraction using RASTA and PLP technique namely RASTA-PLP and tested recognition of word for normal persons and articulatory handicapped persons. The accuracy of recognition is more for normal speakers compared to the articulatory handicapped speakers. Also the recognition results using RASTA PLP Cepstral features are better compared to RASTA PLP spectral features. We can extend our study to improve accuracy of system by using additional filtering and preprocessing technique.

REFERENCES

[1] R Hynek Hermansky and Nelson Morgon(1994,Oct). "RASTA processing of speech" IEEE transaction on speech and audio processing, Vol 2.

[2] L.R.Rabiner and R.W.Schafer, "Digital Processing of Speech Signals", Pearson Education, 1993

[3] Lawrence Rabiner and Biing-hwang Juang, "Fundamentals of speech recognition", Pearson Education, 2003

[4] H Hermansky, "Perceptual Linear Predictive analysis of speech," J Acoust Soc.Am, pp.1738-1752, 1990

[5] RASTA PLP-SPEECH analysis. International computer science institute, 1991.

[6]Keller E. "Fundamentals of Speech Synthesis and Speech Recognition", John Wiley & Sons, New York, USA, (1994).

[7] Shaila D. Apte "Speech and audio processing", Wiley publication Feb 2012

[8]B.Gold and N.Morgan, Speech and audio signal processing, John Wiley and Sons, England, 1999

[9] Stevens K. (1997) "Articulatory-acoustic-auditory relations," in The Hand-book of Phonetic Sciences, (W. J. Handcastle and J. Lavar eds.), Blackwell, Cambridge, U.K., 462–506.

[10] Atal B. S. and Hanauer S. L. (1971) "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., 50 (2), 637–655.

[11] Atal B. and Schroeder M. (1979) "Predictive coding of speech signals and subjective error criteria," IEEE Trans. Acoustics, Speech & Signal Processing, ASSP-27,247–254.

[12] S. Davis and P. Mermelstein, "Comparison of parametric
representation for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. ASSP, vol. 28,
no. 4, pp. 357–366, 1980.

[13] P. Woodland, M. Gales, and D. Pye, “Improving environmental robustness in large vocabulary speech recognition,” in Proc. ICASSP, vol. 1, 1996, pp. 65–68.
[14] J. Makhoul, “Linear prediction: A tutorial review,” Proceedings of the IEEE, vol. 63, no. 4, pp. 561–580, 1975.