International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Dysfluency Recognition by using Spectral Entropy Features

Vinay N A, Bharathi S H

Abstract: ***In recent decades, speech recognition technology has improved effectively and significantly, but it is restricted only for a stream of words. These recognition systems assist human's effectively for structured speech but for unstructured speech this assistance is not so effective for humans to communicate with machines, because unstructured stream of word lacks in providing useful information about pronunciation and punctuation. Recovering of such structural information by detecting the position of each phones in a sentence by locating the sentence boundaries, repeated words and missing phones in each phrase. The proposed work investigates the spectral entropy features, for the automatic detection of voiced and non-voiced regions, in the process of dysfluent speech recognition. The entropy features are estimated by normalizing the Fourier transform spectrum as Probability mass function (PMF). For clear formants of speech, the value of entropy is low and the value of entropy is high for flat distribution of silence part or if there is any noise in speech sample. A comparison of entropy features with Word Error Rate is presented in the proposed work.***

Index Terms: Dysfluency, MHFCC, Spectral Entropy, Speech recognition, IMF, Phones, Word Error Rate (WER).

I. INTRODUCTION

The speech recognition process includes conversion of speech into words based on its features, but recognition of spontaneous is slight difficult because of its dysfluent characteristics in terms of time, pitch, pronunciation and punctuation. There are many speech recognition algorithms, among that most widely used algorithm is Mel-Frequency Cepstrum Coefficient (MFCC). But in dysfluent speech it is necessary to find end point of each phrases, so Mel Hilbert Frequency Cepstral Coefficients (MHFCCs) is adopted in this paper. Along with this, to know the time taken by each words or phones to get settle down Reverberation Time is calculated in dB, which helps to identify the dysfluent part in a complete sentence.

II. MEL HILBERT FREQUENCY CEPSTRAL COEFFICIENTS (MHFCCS)

In MHFCC, the analysis of dysfluent speech is done in time and frequency is done by using Hilbert Transform. The Hilbert Transform is applied to speech in two ways: Empirical Mode Decomposition (EMD) and Hilbert Spectral (HS) Analysis. In EMD method, the dysfluent data is decomposed into finite and smaller number of Intrinsic Mode Functions (IMF). This IMF is defined as function of zero crossings, local maxima and minima values. Since EMD based on local characteristic time scale data of non-stationary (speech) signals along with Hilbert transform results in energy-frequency-time distribution designated as Hilbert Spectrum (HS). The front-end design of MHFCC is as shown in Figure 1.

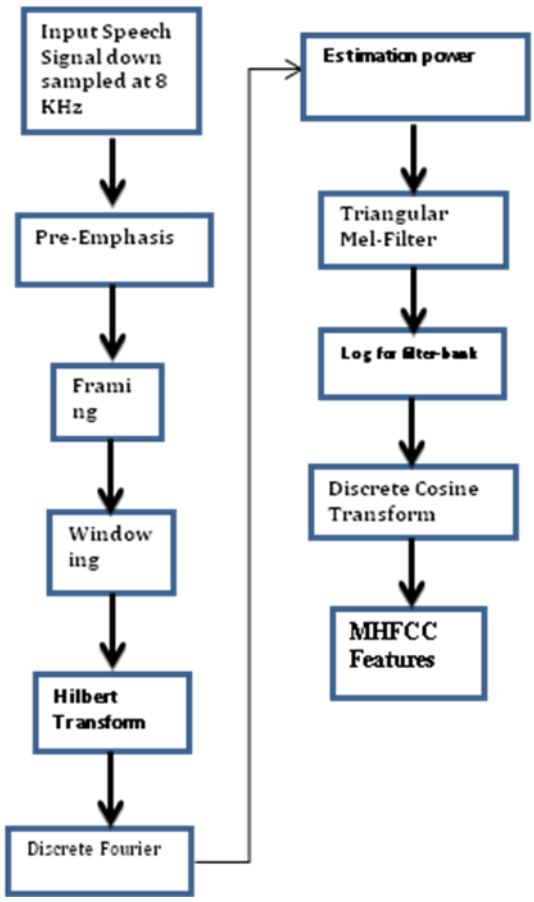

Figure 1: Steps in MHFCC

Pre-emphasis:

The frequency of speech sample is 41 KHz, nevertheless most of the speech constraints are not in this high-frequency range(which are not audible for human eras), so the input speech frequency is ascended down to 8 KHz which can be used for processing further. Then by a passing this high-frequency signal through a low pass filter, the higher frequency components, and some DC distortion components are removed. The transfer function of such filter is given by [19].

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Vinay N A*, School of ECE, REVA University, Bangalore, India
Bharathi S H, School of ECE, REVA University, Bangalore, India

$$S'(n) = s(n) - as(n-1) \quad (1)$$

By taking Z- transform on both sides:

$$\frac{S'(Z)}{s(Z)} = H(Z) = 1 - aZ^{-1} \quad (2)$$

Splitting of speech into frames:

The down sampled speech signal is splited into frames of 5ms and each frame is having 256 samples.

Windowing:

In order to give continuity between first and last frames, each frame is convolved with a window function. To get low-side band attenuation hamming window is used [8].

$$Z(n) = f(n) * w(n) \quad (3)$$

Where $Y(n) - Window\ output$

$f(n) - Frmes\ of\ raw\ speech\ signal$

$w(n) - Window\ function$

For hamming window: Window function is

$$w(n) = 0.54 - 0.46COS\left(\frac{2\pi n}{N-1}\right)$$

Where N-Number of samples=256 and $0 < n \leq N$-1

Discrete Fourier Transform (DFT):

Spectral information entails the vitality levels at various frequencies in the given window. Time-space data is changed over into frequency space to attain the spectral data using DFT. The Discrete Fourier Transform (FFT) is given [8]:

$$Z(K) = \sum_{i=0}^{N} z(n) e^{\frac{-j2\pi ki}{N}} \quad ;0 \leq K \leq N-1 \quad (4)$$

$$|Z(K)| = \sqrt{Re(Z(k))^2 + Im(Z(k))^2} \quad (5)$$

For the real part of DFT signal, power spectrum is determined by taking square of modulus value [6].

$$PS = |V(K)|.\wedge|V(K)| \quad (6)$$

Hilbert Transform:

The spectral estimation using Hilbert transform, extracts Hilbert spectrum envelope which gives immediate values of: amplitude, frequency, and phase, thus both discrimination and anti-noise of feature extraction process are good.

The Discreet Hilbert Transform is given by:

$$\hat{Z}(K) = \begin{cases} -jZ(K), K = 1 \ldots\ldots \frac{N}{2} - 1 \\ jZ(K), K = \frac{N}{2} + 1 \ldots\ldots N-1 \end{cases} \quad (7)$$

The elapsed time of endpoint detection of various speech samples using IMF is tabulated in table 1.

Mel-Scale Smoothing

In this step, the spectrum is rescaled to the range of mel-scale. This will be achieved by multiplying the power spectrum values with a set triangular band pass filters. The positions of these filters are uniformly spaced, as [6]:

$$B_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \quad (8)$$

Where f (m)-center frequency, m- Number of filters, k-0, 1, ---------$\frac{N}{2-1}$

Each filter in filter bank as to satisfy: $\sum_{m=0}^{N-1} B_m(k) = 1$

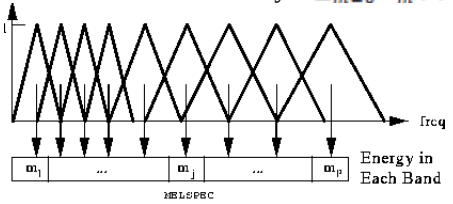

Figure 2: Triangular Band Pass Filter Bank

To reduce the noise and to estimate the errors as [5], Log has been taken for each frame power spectrum values.

$$L(m) = ln\left(\sum_{k=0}^{N-1} |PS|^2 B_m(k)\right) \quad (9)$$

The log values are segmented as mel-scale frequency bins. Mel frequency is calculated as [6]

$$f_{mel} = 2595 log_{10}\left(\frac{f}{100} + 1\right) \quad (10)$$

In this process, the lower frequency components become more important as the frequency bins at higher frequency become larger.

Discrete Cosine Transform (DCT):

DCT coefficients are proportional to DFT coefficients but double the length. Discrete Cosine Transform discards the higher coefficients, the lower coefficients gives this smooth spectral shape.

To train machine learning algorithm, the MHFCC's are fed as features, because lower order coefficients represent exact spectral shape and gives good features of speech but higher order coefficients represents noisy like features. In addition to these Mel spectrum magnitudes of particular amplitude at different frequencies is not so important than the general shape of the spectrum. Finally mean values of MHFCC vectors were considered as feature component for each frame and MHFCC vectors are constructed (column wise) for each frame.

$$D(m) = \sum_{n=0}^{M} L(m) COS\left[\pi m\left(\frac{n-0.5}{M}\right)\right] \quad (11)$$

Where L (m)-logarithmic value, M-Number of filters and $0 < m \leq M$

Table 1: Elapsed time for the detection of end point

Sl.No	Age and Gender of the speaker	Elapsed time in Sec	Type of dysfluency detected	Dysfluent word detected
1.	Male(14 Yrs)	0.39 Sec	Prolongation	Market
2.	Female(8 Yrs)	0.26 Sec	Prolongation	Flower
3.	Female(18 Yrs)	0.286 Sec	Repetition	Friend
4.	Male(20 Years)	0.45	Prolongation, Repetition	After and fruits
5.	Male(12 Years)	0.95	Silence	--

III. SPECTRAL ENTROPY

The discontinuity and peak value of speech distribution is measured by using Spectral Entropy. The spectral entropy ensures the voiced and non-voiced region in speech, it captures the peak value of a speech, which gives the formants and location of speech. Based on this two-values speech recognition is done, because the Fourier spectrum of speech is converted into Probability Mass Function (PMF), by stabilizing the spectrum in each sub-band. Equation (12) is used for sub- band normalization.

$$x_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \quad \text{For i=1 to N} \quad (12)$$

The i^{th} frequency component of the spectrum X_i represents the energy of the signal and x_i indicates the PMF of the spectrum. The sum of area under stabilized spectrum should equal to 1.

The entropy of stabilized spectrum is computed with equation (13).

$$H(x) = \sum_{x \in X} x_i \bullet \log\left(\frac{1}{x_i}\right) \quad (13)$$

To separate the voiced and non-voiced regions, entropy is determined. The max entropy value represents the presence of
flat distribution of silence or noise and minimum entropy value represent the clean speech.
The maximum and minimum entropy values in Fig. 2 show that, frame 5, frame 14, frame 24, frame 28, frame 30 has dysfluency. In this speech sample female speaker of age 18 years telling her friends about her feelings. Here the dysfluency is prolongation of various words in different frames. As mentioned in table 1, the elapsed time to detect this dysfluency 0.286 sec, so total elapsed time taken to recognize the dysfluency in all 30frames is 8.58 sec.

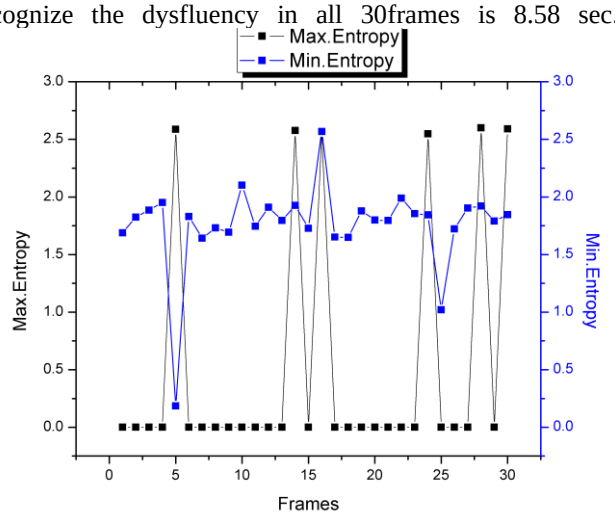

Figure 2: Maximum and Minim Entropy values for 30 frames.

Sub-Bands Entropy

To improve the firmness, the Fourier spectrum is divided into K number of non-overlapping sub-bands of equal size as: 2, 3, 4, 5, 6, 8, 12 and 13 sub-bands.
Then Word Error Rate is computed for 13 sub-bands.
Word Error rate is a communal way to analyses the process of speech recognition, the typical exertion in the analyses of speech recognition performance is identifying of word error rate of different length. For calculation of WER we use Levenshtein distance word level. Levenshtein distance is a minimal quantity of insertions, deletions and substitutions of words for conversion of a hypothesis to a reference.

$$WER = \frac{D(H,R)}{N}$$

Where *D (H,R)* is a Levenshtein distance between *H* and *R*, and *N* is the number of words in the reference *R*. *H* and *R* are cell arrays of words or cells with word sequences or strings. Types of *H* and *R* may be different.

Table 3: Word Error Rates for sub-band entropy features

Sub-bands	WER
Full-band	47.8
2 sub-bands	38.72
3 sub-bands	35.65
4 sub-bands	32.56
5 sub-bands	31.96
6 sub-bands	26.45
8 sub-bands	23.75
12 sub-bands	23.8
13 sub-bands	24.68

Table 3 illustrates the word error rates (WERs) for spectral entropy features up to 13 sub-bands. We could notice the impact of entropy features with better firmness. Table 3, clearly shows that as the dimension of sub-band increases the value of WER is decreased.
The speech samples taken from UCLASS database
consists of both related and not related utterances. The algorithm is trained with a dataset, consists of 14 male and 4 female speakers

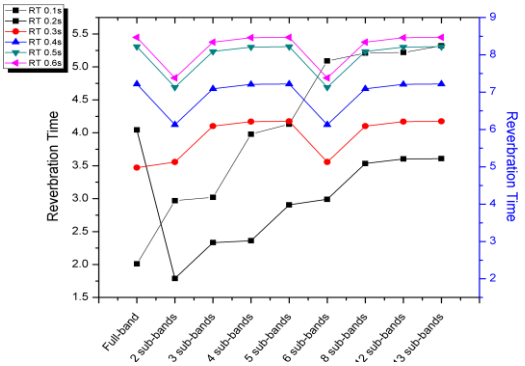

Figure 4: SNR for MHFCC with entropy features

Among this, for testing purpose 6 male and 2 female speakers were considered. Every sub-set composed of 67 words. The signal-to-noise ratios (SNR) are measured from 0dB to 60dB at the interims of 10 dB. Then, the speech is convolved with the RT60 room impulse response and the number of filter coefficients is adjusted as per the reverberation time. Figure 4 shows the SNR for MHFCC with entropy features in additive noise environment. All the spectral entropy features contributed to increase the performance of speech recognition in terms of SNR. The role of sub-band entropy features was important for 12 and 13 sub-bands entropy. These entropy features have shown improvement in performance in additive
noise at different level of SNRs.
Then speech recognition is done with MHFCC and spectral entropy features in reverberant environments. Reverberation is the determination of sound in a particular space after the original sound is produced. A reverberation, or reverb, is formed when a sound is created in a surrounded space producing a large number of echoes to build up and then slowly decay as the sound is captivated by the walls and air. This is most evident when the sound source stops but the reflections remain, decreasing in amplitude, until they can no longer be heard. Table 5 records the reverberation time for MHFCC 0 and entropy features in reverberant atmospheres, the initial results in reverberation of RT 0.1s and RT 0.2s did not show any significant influence or robustness from spectral entropy features, thus it is extended up to 0.6s. The performance of the entropy features greatly deteriorated as the reverberant level increased.
Then speech recognition is done with MHFCC and spectral entropy features in reverberant environments. Reverberation is the determination of sound in a particular space after the original sound is produced.

A reverberation, or reverb, is formed when a sound is created in a surrounded space producing a large number of echoes to build up and then slowly decay as the sound is captivated by the walls and air. This is most evident when the sound source stops but the reflections remain, decreasing in amplitude, until they can no longer be heard. Table 5 records the reverberation time for MHFCC 0 and entropy features in reverberant atmospheres, the initial results in reverberation of RT 0.1s and RT 0.2s did not show any significant influence or robustness from spectral entropy features, thus it is extended up to 0.6s. The performance of the entropy features greatly deteriorated as the reverberant level increased.

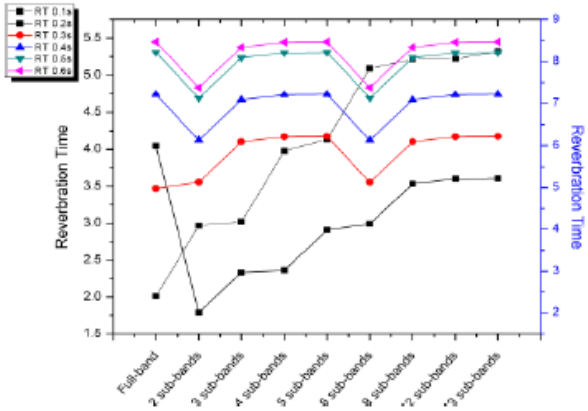

Figure 5: Reverberation time for MHFCC 0 and entropy features in reverberant environments

CONCLUSION:

In this paper we effectively located the boundaries of each segmented words, by finding end point of it using MHFCC and spectral entropy validates the result of MHFCC. The estimation of Word Error Rate(WER) and Reverberation Time(RT) accurately recognized the dysfluency in each phrases and it gives 92% efficiency in recognition of dysfluency based on spectral entropy features.

REFERENCES:

1. Reza Sahraeian and Dirk Van Compernolle, "Cross-Entropy Training of DNN Ensemble Acoustic Models for Low Resource ASR", IEEE/ACM Transactions on Audio, Speech, and Language Processing, DOI 10.1109/TASLP.2018.2851145.
2. Gustav Eje Henter, Member, IEEE, and W. Bastiaan Kleijn, Fellow, IEEE, "Minimum Entropy Rate Simplification of Stochastic Processes: Supplemental Material", IEEE Transactions on Pattern Analysis and Machine Intelligence, DOI: 10.1109/TPAMI.2016.2533382,
3. Yishan Jiao, Visar Berisha, Julie Liss, Sih-Chiao Hsu, Erika Levy, and Megan McAuliffe, "Articulation Entropy: An Unsupervised Measure of Articulatory Precision", IEEE SIGNAL PROCESSING LETTERS, VOL. 24, NO. 4, APRIL 2017.
4. Ji Wu, *Senior Member, IEEE*, Miao Li, And Chin-Hui Lee, *Fellow, IEEE,* "A Probabilistic Framework For Representing Dialog Systems And Entropy-Based Dialog Management Through Dynamic Stochastic State Evolution", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2015.
5. Karthika Vijayan, *Student Member, IEEE*, and K. Sri Rama Murty, *Member, IEEE,* "Analysis of Phase Spectrum of Speech Signals Using Allpass Modeling", IEEE LATIN AMERICA TRANSACTIONS, VOL. 13, NO. 7, JULY 2015 2135.
6. Chi Zhang and John H. L. Hansen, *Fellow, IEEE,* "Whisper-Island Detection Based on Unsupervised Segmentation With Entropy Based Speech Feature Processing", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 4, MAY 2011
7. Carlos Molina, *Student Member, IEEE*, Nestor Becerra Yoma, *Member, IEEE*, Fernando Huenupán, Claudio Garretón, and Jorge Wuth, "Maximum Entropy-Based Reinforcement Learning Using a Confidence Measure in Speech Recognition for Telephone Speech", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 5, JULY 2010.
8. Chi-Sang Jung, Moo Young Kim, *Member, IEEE*, and Hong-Goo Kang, *Member, IEEE,* "Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 6, AUGUST 2010
9. L. Jin and J. Cheng, "An Improved Speech Endpoint Detection Based on Spectral Subtraction and Adaptive Sub-band Spectral Entropy," 2010 International Conference on Intelligent Computation Technology and Automation, Changsha, 2010, pp. 591-594. doi: 10.1109/ICICTA.2010.309.
10. Vivek Kumar Rangarajan Sridhar, *Student Member, IEEE*, Srinivas Bangalore, and Shrikanth S. Narayanan, *Senior Member, IEEE,* "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 16, NO. 4, MAY 2008.
11. Hagai Aronowitz and David Burshtein, *Senior Member, IEEE,* "Efficient Speaker Recognition Using Approximated Cross Entropy (ACE)", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007.
12. Ming Toh, Aik & Togneri, Roberto & Nordholm, Sven. (2005). Spectral entropy as speech features for speech recognition. Proceedings of PEECS.

AUTHORS PROFILE

Vinay N A Received BE Degree in Electronics and Communication Engineering from VTU Belgaum and M.Tech, in Digital Electronics from India SAHE Tumkur. Presently he is pursuing Ph.D from, REVA University, Bengaluru, Karnataka, India. He is working as faculty in School of Electronics and Communication Engineering, REVA University, Bengaluru, India. His research interests Speech Processing, Image processing, Signal Processing.,etc. He is a Associate Member IETE of (AMIETE) INDIA.

Bharathi S.H has 26 years of teaching and research experience. Received her B.E. in Electronics and M.E. in Electronics and Communication, Bangalore University, Karnataka India. Dr. Bharathi S H was awarded the Ph.D. degree in 2013. Her work in the area of Sensors, VLSI, digital Signal Processing. Presently she is working as Professor, REVA University, Bangalore. She is a member of IEEE and Fellow of IETE societies. She is also a member of IEEE Women in Engineering.