Low-Audible Speech Detection using Perceptual and Entropy Features

Karthika Senan J P and Asha A S

Department of Electronics and Communication, TKM Institute of Technology, Karuvelil, Kollam, Kerala, India.
karthika.senan@gmail.com, 91-9447712231

Abstract— Low-audible speech detection is important since it conveys significant amount of speaker information and understanding. The performance of Automatic Speaker Recognition (ASR) and speaker identification systems drops considerably when low-audible speech is provided as input. In order to improve the performance of such systems, low-audible speech detection is essential. The production, acoustic and perceptual properties of such speech is different from normal speech and due to this reason, the methods for detection also differs. In the work, low-audible speech detection process involves feature extraction, feature set combination and a detection algorithm. To obtain the speech components perceived by humans, Perceptual Linear Prediction (PLP), RASTA-PLP (Relative Spectral Perceptual Linear Prediction) and Spectral Information Entropy (SIE) features are extracted. These features are combined and a detection algorithm is performed using Gaussian Mixture Model (GMM) classifier. Mobile phone users can convey their credit card information in an open space using low-audible speech in order to access secure services like phone banking, hotel or car reservation etc. Low-audible speech detection can be used in medical field by speech therapists for evaluating voice disorders in aphonic patients. Forensic scientists are able to recognize speaker identities from low-audible speech which is relevant in the area of national security and defense.

Keywords— Low-audible speech, feature extraction, detection algorithm, Perceptual linear prediction, Relative spectral perceptual linear prediction, spectral information entropy, Gaussian mixture model.

INTRODUCTION

Low audible speech or whispering is the mode of speech defined as speaking softly with little or no vocal fold vibration. Thus the passing air does not generate any fundamental frequency, but just a little turbulent noise. Due to the high noise like content and lack of harmonic structure, the modeling of low-audible speech is challenging compared to other modes of speech production. Moreover, it do not reach very far and can be masked easily by environmental noise. Current speech processing systems works well in situations where normally phonated speech is provided as input. When low-audible speech in noisy environment is provided as input to Automatic Speech Recognition (ASR) or Automatic Speaker Verification (ASV) systems, their performance is reduced considerably. There also occurs mismatches between the training and testing phases of such systems. In order to improve the system performance, detection and recognition of low-audible speech is essential.

Low-audible speech differs from normal speech in its physiological production, acoustic and perceptual properties. In normal speech, air from the lungs causes vocal folds of the larynx to vibrate, exciting the resonances of vocal tract. In low-audible speech, the glottis is opened and turbulent flow created by exhaled air passing through this glottal constriction provides a source of sound. Thus the low-audible speech differs from normal speech in its physiological production. Low-audible speech is characterized by the absence of periodic excitation, changes in energy and duration characteristics, shift of lower formant locations, and changes in the spectral slope. The intensity of low-audible speech is significantly lower than that of neutral speech. Due to the absence of vocal fold vibration, fundamental frequency and harmonic components are absent which makes it aperiodic. The location of lower frequency formants in low-audible speech are shifted to higher frequencies as compared to neutral speech counterparts. The spectral tilt is much flatter compared to normal speech. Due to these characteristics, the methods for processing and detection of low-audible speech is quite different from normal speech.

Various methods like calculating the spectral energy ratio, spectral tilt [2], spectral flatness measure [10], linear prediction [14] etc. are useful for low-audible speech detection in silent environment. The spectral energy ratio method uses the shift of spectral energies to higher frequencies to detect whispered speech. The spectral flatness measure is calculated because in whispered speech, the spectral slope becomes flat due to loss in low-frequency content. When noise is present, these methods will not provide adequate results. Therefore other detection methods using features extracted from time waveform or spectral analysis of speech signal like entropy-based features [5],[6],[7], linear prediction residual [8], linear prediction analysis using minimum variance distortionless modeling of speech [9],[10], Mel Frequency Cepstral Coefficients (MFCC) [14] are explored. But these existing methods are not efficient in presence of background noise. The proposed method include features which perform well in presence of background noise. The features which are perceived by humans are extracted and the classifier is trained accordingly. It was found that these features also separate speech from noise, reverberation etc. The Gaussian Mixture Model (GMM) Classifier used in the method give better performance compared to other classifiers and is quite helpful in speaker verification tasks.

The use of multimedia portable devices like smart phones and tablets enables users to communicate in any environment. Many applications allow them to interact with these devices through voice. They can carry out tasks such as unlocking the phone or accessing secure services using voice as an interacting medium. With the help of such devices, they can also make confidential and private conversations even in public places. In such situations, they whisper over the phone to reduce the amount of information being spilled out. In such scenarios, the detection of low-audible speech is essential thereby the users can convey their social security numbers, pin numbers or credit card numbers without being overheard. The application of low-audible speech detection also occurs in spoken document retrieval to preserve historical data. It is useful in the field of medicine where speech scientists use low-audible speech to determine perceptual constants and medical doctors evaluate it for safe recovery of larynx surgery patients.

ACOUSTIC DIFFERENCES OF LOW-AUDIBLE SPEECH FROM NEUTRAL SPEECH

In normally phonated speech, air from the lungs causes the vocal folds of the larynx to vibrate, exiting the resonances of the vocal tract. In low-audible speech, vocal folds do not vibrate and the glottal aperture remains open. The turbulent flow created by the exhaled air passing through the glottal constriction provides a source of sound. This sound source is distributed through the lower portion of the vocal tract and the resulting speech is noise excited. The major differences between whispered and neutral speech [3],[4] are the following:

1. The spectrogram of low-audible speech indicate that it does not have a definite formant structure due to lack of vibrating vocal folds. The formants that are present are shifted to higher frequencies as compared to their neutral speech counterparts.
2. Due to the turbulence created at the vocal folds, there is a shift in spectral powers to the higher frequencies in low-audible speech.
3. The spectral slope of low-audible speech is flatter than that of neutral speech and the duration of it is longer than that of normal speech.
4. Low-audible speech has much lower energy than that of normal speech.

METHODOLOGY

The methodology for low-audible speech detection comprises of feature extraction process, feature set combination and a detection algorithm. The feature extraction process includes the extraction of three features namely: PLP (Perceptual Linear Prediction), RASTA-PLP (RelAtive SpecTrAl Perceptual Linear Prediction) features and entropy based features. The feature set combination process involves combining discriminative capabilities from different sets of features. The detection algorithm is performed using a Gaussian Mixture Model (GMM) based classifier.

The basic block diagram for the methodology adopted in the work is as shown in figure. Figure 1 shows all the major processes involved in the low-audible speech detection.

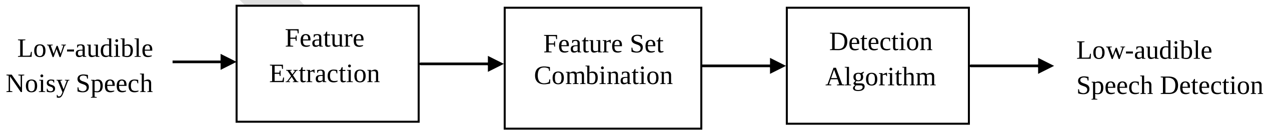


Figure 1: Block Diagram for Methodology

The processing steps and proposed method used in low-audible speech detection are as follows:

Feature Extraction

Feature extraction is the computation of a sequence of feature vectors which provides a compact representation of the given speech signal. This is intended to produce a perceptually meaningful representation of the speech signal. The purpose of feature extraction is to transform the audio data into a space where the observations from the same class will be grouped together and observations of different classes will be separated. Thus the main goal of feature extraction process is to compute a sequence of feature vectors providing a compact representation of the input signal. Prior to feature extraction process, pre-processing steps like framing and windowing are performed in the input speech signals.

Step 1:- Framing

Speech signals are slowly timed varying signals. If speech signals are examined over a sufficiently short period of time (5-100 ms), its characteristics are fairly stationary. In order to analyze speech signals, they are divided into frames. In this step, speech signals are blocked into small frames of *N* samples, with next frames separated by *M* samples *(M<N)* with this the adjacent frames are overlapped by *(N-M)* samples. That is, each frame shares the first part with the previous frame and the last part with the next frame. Studies show that the standard value taken for the samples, *N=256* and for overlapping is *M=100* with the reason of dividing the given speech signal into small frames having sufficient samples to get enough information. If the frame size is smaller than the described size, then the number of samples in the frames will not be enough to get the reliable information and with large size frames it will cause frequent change in the information inside the frame. This process of breaking down the signal into frames will continue until the whole speech signal is broken down into small frames.

Step 2:- Windowing

Windowing is performed to minimize the disruptions at the starting and at the end of the frame. Since the edges add harmonics, it is necessary to tone down the edges using a window. If the window is defined as $W_n(m), 0 \le m \le N_m - 1,$ and N_m stands for the quantity of samples within every frame. The output after windowing the signal will be represented as:

$$Y(m) = X(m)\,W_n(m), \qquad 0 \le m \le N_m - 1 \tag{1}$$

Y(m) represents the output signal after multiplying the input signal represented as *X(m)* and Hamming window represented by $W_n(m)$. Hamming window is applied for carrying out windowing which usually represented as:

$$W_n(m) = 0.54 - 0.46\cos\left(\frac{2\Pi m}{N_m - 1}\right), \qquad 0 \le m \le N_m - 1 \tag{2}$$

Hamming window provides spectral analysis with a flatter pass band and significantly less stop band ripple. This property with the fact that it normalize the signal, so that the energy of the signal will be unchanged through the operation, play an important role for obtaining smoothly varying parametric estimates.

Step 3:- Perceptual Linear Prediction (PLP) Feature Extraction

Perceptual Linear Prediction (PLP) model was developed by Hynek Hermansky [11]. PLP models the human speech based on the concept of psychophysics of hearing. The technique uses three concepts from the pshycophysics of hearing to derive an estimate of the auditory spectrum:

1. The critical-band spectrum resolution
2. The equal-loudness curve
3. The intensity loudness power law

The auditory spectrum is then approximated by an autoregressive all-pole model. PLP analysis is more consistent with human hearing in comparison with conventional linear prediction (LP) analysis. The major disadvantage of LP all-pole model is that autoregressive all-pole model approximates power spectrum equally well at all frequencies of analysis band. This is highly inconsistent with human hearing as the spectral resolution of hearing decreases with frequency. The spectral details of power spectrum are discarded by LP analysis. Therefore a class of spectral transform LP techniques that modify the power spectrum of speech before its approximation by the autoregressive model is introduced.

The steps involved in Perceptual Linear Prediction (PLP) are as follows:

1. The framed and windowed speech signal is subjected to Discrete Fourier Transform (DFT) which transforms it into frequency domain.
2. Computation of Critical Band Spectrum :- The power spectrum obtained (denoted by *P(ω)*) is warped along its frequency axis *ω* into the bark frequency **Ω** by

   $$\Omega(\omega) = 6\ln\left\{\frac{\omega}{1200\Pi} + \left[\left(\frac{\omega}{1200\Pi}\right)^2 + 1\right]^{0.5}\right\} \qquad (3)$$

   where *ω* is the angular frequency in rad/s. The resulting warped spectrum is convoluted with the power spectrum of the simulated critical –band masking curve, *ψ(***Ω***)*. The critical band curve is given by:

   $$\Psi(\Omega) = \begin{cases} 0 & for\ \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & for - 1.3 \le \Omega \le -0.5 \\ 1 & for - 0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & for\ 0.5 \le \Omega \le 2.5 \\ 0 & for\ \Omega > 2.5 \end{cases} \qquad (4)$$

   The Bark filter bank used in the analysis allocates more filters to the lower frequencies where hearing is more sensitive. Also the shape of auditory filters is approximately constant on the Bark scale.
3. The discrete convolution of critical band curve with power spectrum gives samples of the critical-band spectrum.

   $$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\,\Psi(\Omega_i) \qquad (5)$$

   The convolution of relatively broad critical band masking curves *ψ(***Ω***)* reduces the spectral resolution of *θ(***Ω***)* in comparison with the power spectrum *P(ω)*. *θ(***Ω***)* is sampled in approximately 1-Bark intervals.
4. The sampled critical band power spectrum is pre-emphasized by simulated equal-loudness curve. The function is an approximation to the nonequal sensitivity of human hearing at different frequencies.
5. The operation before all-pole modelling is the cubic root compression. This operation is an approximation to the power law of hearing and stimulates the nonlinear relation between the intensity of sound and its perceived loudness.
6. In the final operation of PLP analysis, an all-pole model is estimated using the auto-correlation method. The Inverse Discrete Fourier Transform (IDFT) is applied to yield the dual of autocorrelation function. The IDFT is used since only a few autocorrelation values are needed. The autocorrelation values are used to solve the Yule-Walker equations for the autoregressive coefficients. The autoregressive coefficients are transformed into cepstral coefficients.

Step 4:- Relative Spectral Perceptual Linear Prediction (RASTA-PLP) Feature Extraction

In RASTA-PLP (RelAtive SpecTrAl Perceptual Linear Prediction) feature extraction, each frequency channel is band-pass filtered by a filter with sharp spectral zero at the zero frequency [12]. Since any slowly varying component in each frequency channel is suppressed by this operation, the new spectral estimate is less sensitive to slow-variations in the short-term spectrum. Thus RASTA-PLP features suppresses the spectral components that change more slowly or quickly that the typical range of change of speech.

The initial steps of RASTA-PLP are same as that of the conventional Perceptual Linear Prediction (PLP) speech analysis. RASTA-PLP feature extraction follows step 1 to step 3. The remaining steps are as follows.

1. After computing the critical band spectrum, its logarithm is taken.
2. The temporal derivative of the log critical-band spectrum is estimated.
3. Re-integrate the log critical band temporal derivative using a first order IIR system.

The remaining steps are as explained in the PLP feature extraction (i.e. steps 4 to 6). The block diagram for RASTA-PLP feature extraction is as shown in Figure 2.

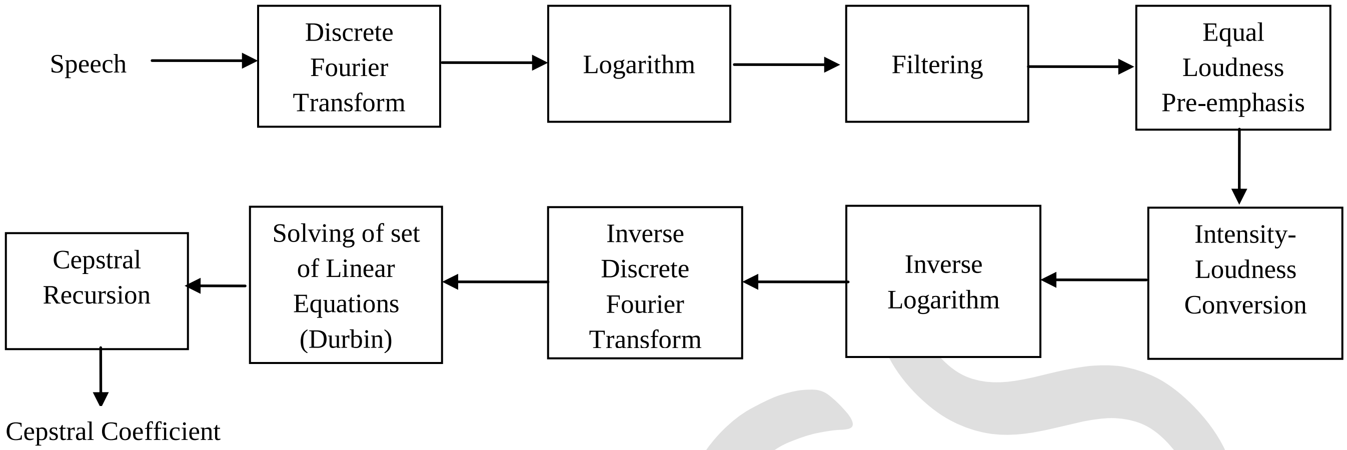

Speech
Discrete Fourier Transform
Logarithm
Filtering
Equal Loudness Pre-emphasis
Intensity-Loudness Conversion
Inverse Logarithm
Inverse Discrete Fourier Transform
Solving of set of Linear Equations (Durbin)
Cepstral Recursion
Cepstral Coefficient

Figure 2. RASTA-PLP Feature Extraction

Step 5:- Spectral Information Entropy Feature Extraction

Entropy based features are considered because short term spectrum is more organized during speech segments than during noise. The spectral peaks of the spectrum are more robust to noise and due to this reason, a voiced region of speech would induce low entropy. The entropy in time-frequency domain known as Spectral Information Entropy (SIE) is found as a useful feature for low-audible speech detection.

The Spectral Information Entropy (SIE) for the input speech frame is measured in the following manner.

1. Let *X(k)* be the power spectrum of the input speech frame *x(n)*, where *k* varies from k_1 to k_m , a specified frequency band; then that portion of the frequency content in the k_{th} band versus the entire frequency response is written as:

$$p(k) = \frac{|X(k)|^2}{\sum_{j=k_1}^{k_M} |X(j)|^2}, \quad k = k_1 \; to \; k_M \qquad (6)$$

2. Since $p(k) = \sum_{k=k_1}^{k_M} p(k) = 1, p(k)$ can be viewed as an estimated probability that describes the energy distribution within this frequency band $(k=k_1,...,k_M)$ can be calculated as

$$H = \sum_{k_1}^{k_M} p(k) \log p(k) \qquad (7)$$

The SIE represents the distribution of energy over the frequency domain rather than the total amount of energy over the entire frequency domain. Even though the original waveform is amplified, the spectral information energy which means that it is not influenced by the amplitude of the original speech signal.

Step 6:- Feature Set Combination

The idea for feature set combination is to use the discriminative capabilities from different sets of features that have been computed on the same basis from the speech recordings. That is, for a given frame, the feature vector *X* is computed (nx_1),i.e *n* features. From the same frame the feature vector *Y*, (mx_1), i.e *m* features, then the combination will be the feature vector *Z=[X;Y]* of dimension $(m+n)\times 1$. This is done for all frames.

Step 6:- Detection Algorithm

Gaussian Mixture Model (GMM) based classifier is used for the detection of low-audible speech. Gaussian Mixture Model (GMM) is a distribution which consists of finite number of Gaussian distributions in the linear way. Gaussian distribution is commonly used because it provides a mathematically straight forward analysis and also it is well qualified to approximate many types of noise in physical systems. GMM is used for unsupervised learning because it can identify the data patterns and cluster those sharing similar data behaviours together. Expectation Maximization (EM) algorithm is a method to estimate the parameters under Maximum a Priori (MAP) or Maximum Likelihood (ML) since hidden variables are involved.

The Expectation Maximization (EM) algorithm computes probabilities for each possible completion of the missing data, using the current parameters. These probabilities are used to create a weighted training set consisting of all possible completions of the data. The EM algorithm alternates between the steps of guessing a probability distribution over completions of missing data given the current model (known as the E-step) and then re-estimating the model parameters using these completions (known as the M-step). The name 'E-step' comes from the fact that one does not usually need to form the probability distribution over completions explicitly, but rather need only compute 'expected' sufficient statistics over these completions. Similarly, the name 'M-step' comes from the fact that model re-estimation can be thought of as 'maximization' of the expected log-likelihood of the data.

A general form of the EM algorithm can be formulated as follows:

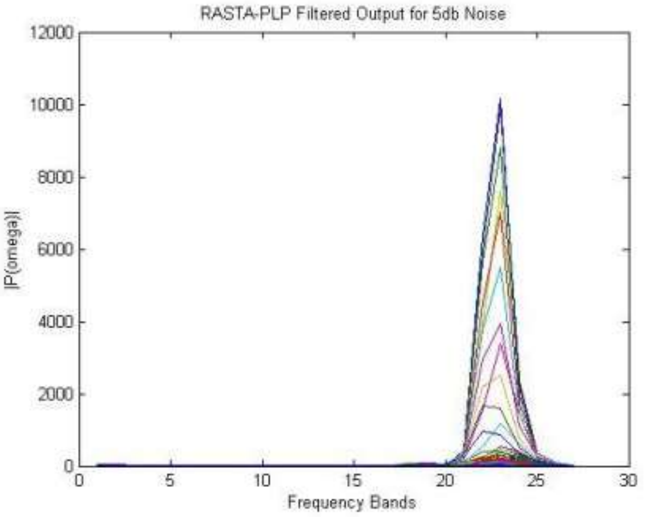
The notation *X* and *Y* is the unobserved data and the observed data corresponding to *X* respectively. θ is the parameters needed to calculate the likelihood *f(Y)*.The goal is to calculate the maximum likelihood θ_{ML} that maximizes $L(\theta) = \log f(Y|\theta)$. Usually the $\log(f(X,Y|\theta))$ has well defined form and thus easy to compute the maximum but it asks the unobserved data *X*; then what the EM algorithm does is to figure out a sequence of θ' and θ'' such that $L(\theta') \succ L(\theta'')$. The calculation steps are the following:

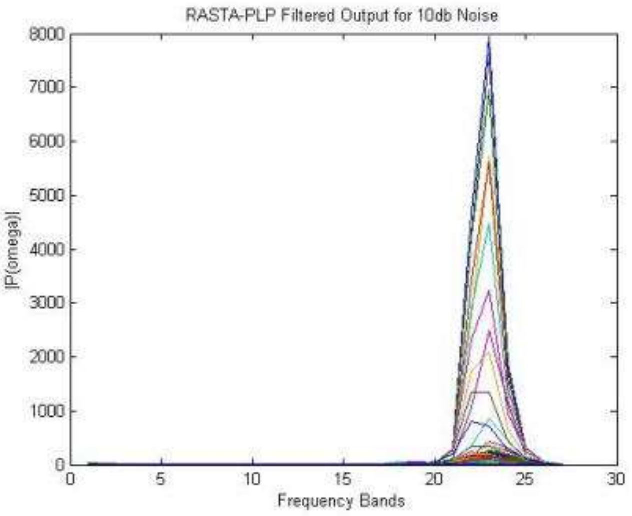
1. Estimation Step: Calculate the expectation of unobserved data $E_{f(X|Y,\theta')}[\log f(X,Y|\theta'')]$.
2. Maximization Step: Find θ'' such that $\theta'' = argmax\left(E_{f(X|Y,\theta'')}[\log f(X,Y|\theta'')]\right)$.

If it hold that: $E_{f(X|Y,\theta'')}[\log f(X,Y|\theta'')] \succ E_{f(X|Y,\theta'')}[\log f(X,Y|\theta')]$ then it is also valid $L(\theta') \succ L(\theta'')$ to achieve the goal of Maximum Likelihood.

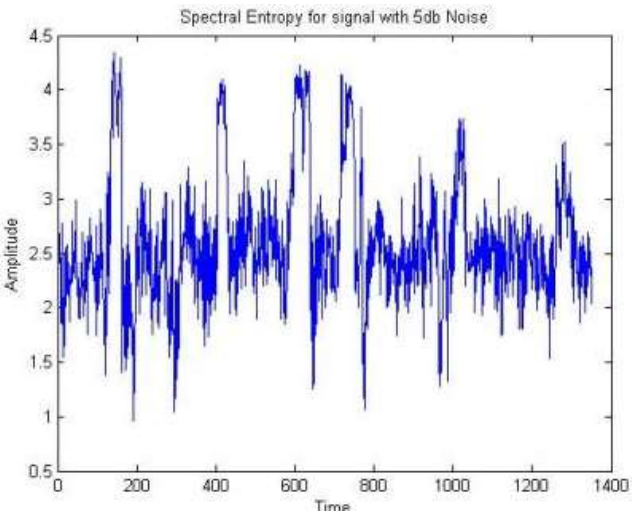
EXPERIMENTAL RESULTS & ANALYSIS

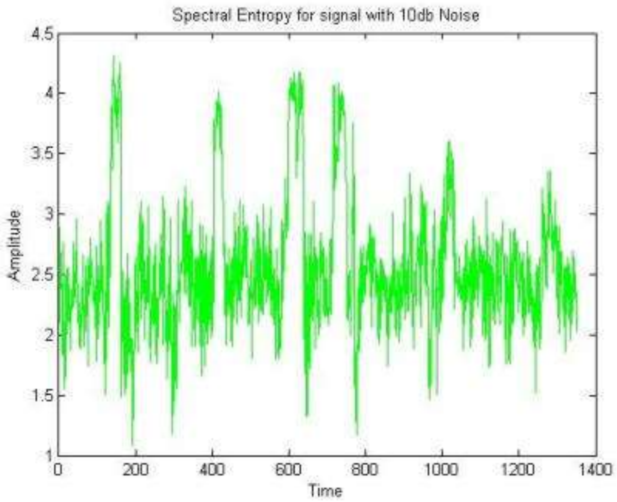
The input speech signal of 3ms duration is obtained from CHAINS Speech Database. The speech signals were created by adding babble noise of different signal strengths. Thus a noisy speech environment was created and changes in the amplitude of input speech was observed. The RASTA-PLP filtering was carried out as the primary feature extraction process are shown in figure. From the filtered output, it is seen that the magnitude of the signal with 10 db noise is less than that of the signal with 5db noise. The components are found in the frequency band ranging from 20-25 Hz. The frequency range of low-audible speech are particularly low frequencies and the audible range of frequencies are filtered.





The spectral entropy for both the speech segments was plotted. Figure shows the spectral information entropy for speech signal with 5db noise and 10 db noise.





CONCLUSION

The RASTA-PLP feature extraction process extracted the speech components perceived by humans. More amount of low-audible speech information was obtained by this feature extraction. The output of RASTA-PLP feature extraction process is components in the low-frequency range since low-audible speech components are present in the low frequency range. The spectral entropy feature was also extracted since it is useful for separating speech from background noise. The low-audible speech can be used for speaker verification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. The speaker verification process comprises of training and recognition phases. The training of the classifier can be made by feature extraction and the same classifier can be used for training. The gender of the speaker can also be identified from the detected speech.

REFERENCES:

[1] Milton Sarria-Paja, Tiago H. Falk, "Whispered Speech Detection in Noise using Auditory-Inspired Modulation Spectrum Features", IEEE Signal Processing Letters, pp. 783-786, vol.20, No.8, Aug. 2013.

[2] C. Zhang, J. H. L. Hansen, "Analysis and Classification of Speech Mode: Whispered through Shouted", Proc. Interspeech '07-ICSLP, Antwerp, Belgium, pp. 2289-2292, 2007.

[3] T. Itoh, K. Takeda, F. Itakura, "Acoustic Analysis and Recognition of Whispered Speech", IEEE Int. Conf. on Acoustic, Speech and Signal Processing, pp. 389-392, vol.1, 2002.

[4] Nicholas Obin, "Cries and Whispers: Classification of Vocal Effort in Expressive Speech", Interspeech, Portland, United States, Sep. 2012.

[5] C. Zhang, J.H.L Hansen, "Effective Segmentation based on Vocal Effort Change Point Detection", Proc. of ITRW, Aalborg, Denmark, June 2008.

[6] C. Zhang, J.H.L Hansen, "Whisper-Island Detection based on Unsupervised Segmentation with Entropy-based Speech Feature Processing ", IEEE Trans. on Audio, Speech & Lang. Processing, pp. 883-894, vol. 19, No.4, May 2011.

[7] C. Zhang, J.H.L Hansen, "An Entropy based Feature for Whisper-Island Detection within Audio Streams", Proc. Interspeech '08, Brisbane, Australia, 2008.

[8] C. Zhang, J.H.L Hansen, "Advancements in Whisper-Island Detection using the Linear Prediction Residual", International Conf. on Acoustics, Speech and Language Processing (ICASSP), pp. 5170-5173, 2010.

[9] A. Mathur, S. Reddy, and R. Hegde, "Significance of Parametric Spectral Ratio Methods in Detection and Recognition of Whispered Speech," EURASIP Journal on Advanced Signal Processing, no.157, 2012.

[10] A. Mathur, S. Reddy, and R. Hegde, "Significance of the LP-MVDR Spectral Ratio Method in Whisper Detection," IEEE National Conf. on Communications (NCC), pp. 1-5, 2011.

[11] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", Journal of Acoustic Society of America, vol.4,pp. 1738-1750, April 1990.

[12] Hynek Hermansky, Nelson Morgan, Aruna Bayya, Phil Kohn,"RASTA-PLP Speech Analysis", Speech Communication, vol. 45, pp. 139–152, 2005.

[13] C. B. Do, S. Batzoglou, "What is Expectation Maximization Algorithm?", Nature Biotechnology, vol.26, no.8, August 2008.

[14] T. F Quatieri,"Discrete-Time Speech Signal Processing", Pearson Education Inc., 2002.

[15] B. Gold,N. Morgan,"Speech and Audio Signal Processing: Processing and Perception of Speech and Music", Wiley, India, 2006.

[16] Matthias Wolfel, John McDonough, "Distant Speech Recognition", John Wiley & Sons Ltd, 2009