EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Significance of parametric spectral ratio methods in detection and recognition of whispered speech

Arpit Mathur, Shankar M Reddy and Rajesh M Hegde*

Abstract

In this article the significance of a new parametric spectral ratio method that can be used to detect whispered speech segments within normally phonated speech is described. Adaptation methods based on the maximum likelihood linear regression (MLLR) are then used to realize a mismatched train-test style speech recognition system. This proposed parametric spectral ratio method computes a ratio spectrum of the linear prediction (LP) and the minimum variance distortion-less response (MVDR) methods. The smoothed ratio spectrum is then used to detect whispered segments of speech within neutral speech segments effectively. The proposed LP-MVDR ratio method exhibits robustness at different SNRs as indicated by the whisper diarization experiments conducted on the CHAINS and the cell phone whispered speech corpus. The proposed method also performs reasonably better than the conventional methods for whisper detection. In order to integrate the proposed whisper detection method into a conventional speech recognition engine with minimal changes, adaptation methods based on the MLLR are used herein. The hidden Markov models corresponding to neutral mode speech are adapted to the whispered mode speech data in the whispered regions as detected by the proposed ratio method. The performance of this method is first evaluated on whispered speech data from the CHAINS corpus. The second set of experiments are conducted on the cell phone corpus of whispered speech. This corpus is collected using a set up that is used commercially for handling public transactions. The proposed whisper speech recognition system exhibits reasonably better performance when compared to several conventional methods. The results shown indicate the possibility of a whispered speech recognition system for cell phone based transactions.

Keywords: Parametric spectrum estimation, Linear prediction, Minimum variance distortion less response, Bayesian information criterion, HMM, MLLR, Whisper detection, Automatic speech recognition (ASR)

Introduction

Speech has been the most primitive modes of communication between all higher forms of life. However it is interesting to note that even while the basic organs that regulate our speech are the same, speech varies with the speaker. This variation transcends grammar and vocabulary. This difference is accounted for by prosody which is defined as a science of pitch, loudness, tempo, rhythm and intonation of speech. It is on account of these differing prosodic features that robust automatic speech recognition (ASR) systems are still a challenge. By and large models based on a large collection of regional databases have indeed proven very effective to counter the regional variations. This has led to attempts to even match prosodic features to subject's mother tongue as in [1]. However the vulnerability of prosodic features to emotional changes is still a challenge. The same speaker can have different prosodic features under different emotional states that can lead to different modes of speech. The ineffectiveness of the usual speech engines over changes in speech modes is evident from the study in [2]. Whisper is one such natural mode of speech. It is a regular response to situations that require secrecy. It can selectively exclude potential listeners of the message. Patients with a collapsed larynx or who have undergone laryngectomy due to cancer of larynx may have to resort to a form of speech that is very close

*Correspondence: rhegde@iitk.ac.in
Department of EE, Indian Institute of Technology Kanpur, Kanpur, India

to whispered speech. This form is known as esophageal speech.

The classification of speech can be done into five categories depending upon the modes of speech production i.e. difference in vocal efforts [3]. They are whispered speech, soft speech, neutral speech, loud speech and shouted speech. Whispered speech is produced in the absence of vocal cord vibrations, however larynx movements are the same as in neutral speech. This mode of speech generated has high noise-like characteristics. Soft speech is produced when the listener and the speaker are very close by and there is an element of secrecy or quietness to be maintained, for example talking to a friend in the library. Both the vocal cord and the larynx vibrate in this mode. Neutral speech is the regular baseline mode of speech spoken at leisure with people. Loud speech is the mode of speech employed when addressing a large gathering or when in a noisy environment. The increased effort is often accompanied by increased length of speech segments to increase intelligibility. Shouted speech is produced in a state of extreme anger or when addressing a person over a large distance. It is usually accompanied by extreme articulation by the glottis, however extreme articulation usually leads to less intelligibility. Of all these modes, modeling whispered speech is most challenging because of its high noise-like content. It also suffers from lack of a harmonic structure due to absence of vocal cord excitation [4]. Hence Spkr-ID systems have been reported to perform worst in case of whispered speech [2]. Therefore, detecting whispered speech segments in a normally phonated speech signal and recognizing the detected whispered speech segments separately can improve the performance of speech recognition systems. Several methods have been developed in detecting whisper-islands embedded in a normally phonated speech signal [5,6]. In this work we used linear prediction (LP) to minimum variance distortionless response (LP-MVDR) spectral ratio based features [7] for whispered speech detection. Whisper-island detection can be done using features extracted from linear predictive residual (LPR) and Bayesian information criteria (BIC) [8]. Chi and Hanse [9] proposes a method which detects whisper-islands in an audio clip via BIC/T2-BIC using a 4-D feature set. In some works MVDR is used for speech recognition [10] and spectral coding of speech [11]. Yapanel and Dharanipragada [12] used perceptual MVDR (PMVDR) based features in speech recognition.

Acoustic aspects of neutral and whispered speech

It is important to understand both articulatory and acoustic aspects of neutral speech and whispered speech to differentiate one from the other. However since this article relates only to the acoustic aspects, a brief discussion of the acoustic aspects that differentiates neutral from whispered speech follows herein. Neutral speech is primarily characterized by its formants. The formant structure can be changed by changing the vocal tract length which is done by moving the lips, tongue, teeth or by closing or opening the nasal cavity. On the other hand whispered speech does not have the vibrating vocal cords. Hence there are very few or no glottal pulses. However oral features do the articulation and produce the required characteristic sound [13]. This sound does not have a definite formant structure. The formants that are present are shifted to higher frequencies as compared to their neutral speech counterparts [14]. Because of turbulence created at the vocal folds, there is a shift in spectral powers to the higher frequencies in whispered speech. Figure 1, illustrates a spectrogram of a neutral speech and a whispered speech utterance. The shift in spectral powers to the higher frequencies can be noted in the spectrogram of whispered speech when compared to neutral speech.

Brief review of techniques for detection of whispered speech

From Section "Acoustic aspects of neutral and whispered speech", we realized that whispered speech is significantly different from the neutral speech on many grounds. Whispered speech is produced without vocal cords motion that leads to lack of a formant structure (though not the absence of it). The formant structure whatever is present is also found to have shifted to higher frequencies. The air is throttled at the glottis leading to turbulence and the corresponding shift [5]. All this in the spectral power terms reflect a shift of concentration of power to higher frequency bands. Based on these characteristics, few detection techniques have been proposed. These techniques were explored for their performance on whispered speech. They are briefly described in this section.

Spectral energy ratio

Energy ratio is a fast and easy to implement method that can detect large segments of whisper [15]. The method uses the shift of spectral energies to higher frequencies to detect whisper. Quite simply, spectral energy ratio is defined as ratio of frequencies of higher band to lower band. Hence

$$E_{\text{rat}} = \frac{E_{\text{HB}}}{E_{\text{LB}}} \tag{1}$$

The two bands' definition is subjective. Usually lower band is taken to be from 0–1 kHz and higher band corresponds to 2.5 kHz-end frequency. The method is crude and only conceptual. The detection rate is very low with only large segments being detected. Also this method is expected to perform badly in noisy conditions as noise will have a direct effect on the spectral powers (Figure 2).

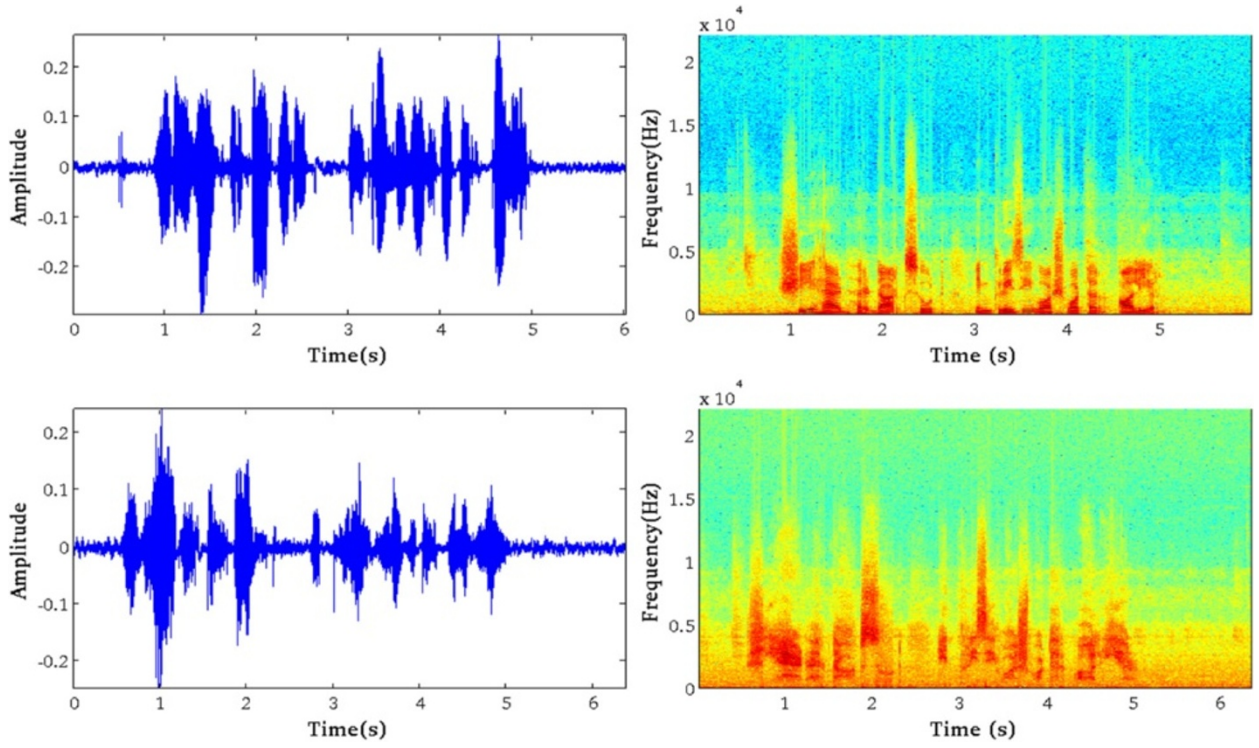

Figure 1 Time domain waveforms (left) and Spectra (right) for the same speaker speaking 'She had her dark suit in greasy wash water all year' in neutral (top) and whispered (bottom) speech indicating shift of spectral energies from lower bands to higher bands spectrograms. Also addition of noise-like content and lack of harmonics can be observed for whispered speech.

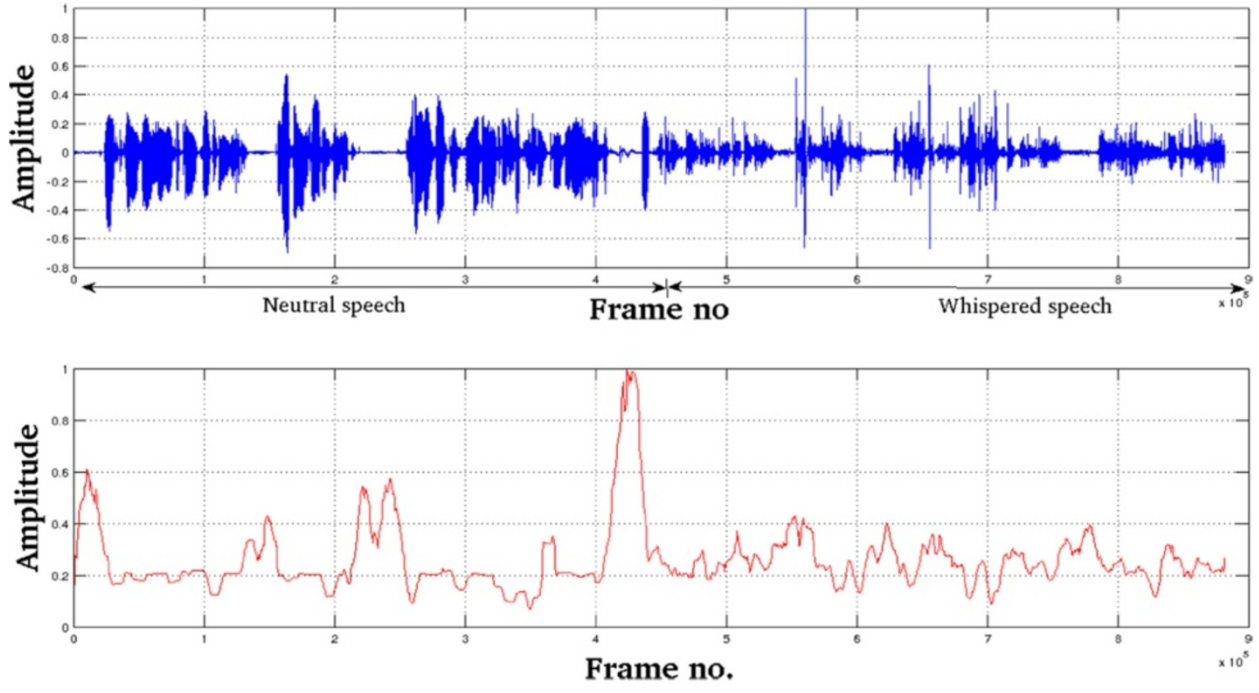

Figure 2 Plot of spectral energy ratio feature (bottom) with the corresponding speech signal (top). Whispered speech has higher spectral energy ratio on account of shift of spectral power to higher frequencies.

Spectral flatness

Spectral flatness is a sense of the 'flatness' of the speech spectrum. If we wanted to explore the presence of white noise in the given signal; then a spectral flatness measure (SFM) must be devised. The measure at the same time should take into account the particular structure of speech. This means that penalties are to be assigned accordingly [16,17] (Figure 3).

Let s_n be a finite real time speech signal. Thus the energy of the signal with the application of parseval's theorem gives

$$E = \sum_{k=1}^{n} s_k^2 = \int_{-\pi}^{\pi} |S(e^{j\theta})|^2 \frac{d\theta}{2\pi} \tag{2}$$

The SFM that we need must have the following normalized boundary conditions namely one value for a perfectly flat spectrum and another for a spectrum with peaks and troughs. Let us consider an integrand

$$M(\theta) = \log\left(\frac{|S(e^{j\theta})|^2}{E}\right) \tag{3}$$

We can observe that M(θ) will give a value of zero for a perfectly flat spectrum and non zero finite values otherwise. Hence M(θ) is a candidate for measuring spectral flatness. However having a simple criteria like mean square criteria will weigh positive and negative deviations from the mean value as the same. This is not acceptable in the context of speech processing as positive deviations are indicative of harmonic content. Hence these should be weighed higher making the spectrum look to be less flat. Thus we consider $e^{M(\theta)} - 1 - M(\theta)$. It can be seen that it has the required property for speech signals. Consider the integration over some interval

$$\gamma = \int_{-\pi}^{\pi} |e^{M(\theta)} - 1 - M(\theta)| \frac{d\theta}{2\pi} \tag{4}$$

This integral basically comes down to

$$\gamma = -\int_{-\pi}^{\pi} |M(\theta)| \frac{d\theta}{2\pi} \tag{5}$$

Thus the mathematically complex looking integrand upon integration gains a very simple form. Hence $e^{[-\gamma]}$ from previous equation will be one for a perfectly flat spectrum and lie between 0 and 1 otherwise. Hence this measure has also been used to detect unvoiced segments of speech [18]. Also it can be written as

$$\gamma = \frac{\exp \int_{-\pi}^{\pi} \log(|S(e^{j\theta})|^2) d\theta}{\int_{-\pi}^{\pi} |S(e^{j\theta})|^2 d\theta} \tag{6}$$

The above equation in discrete time can be interpreted as geometric mean to arithmetic mean of the spectral powers. Thus we have our SFM as

$$\text{SFM} = \frac{\left(\prod_{k=0}^{N_{\text{FFT}}-1} |S(e^{j\omega_k})|^2\right)}{\frac{1}{N_{\text{FFT}}} \sum_{k=0}^{N_{\text{FFT}}-1} |S(e^{j\omega_k})|^2} \tag{7}$$

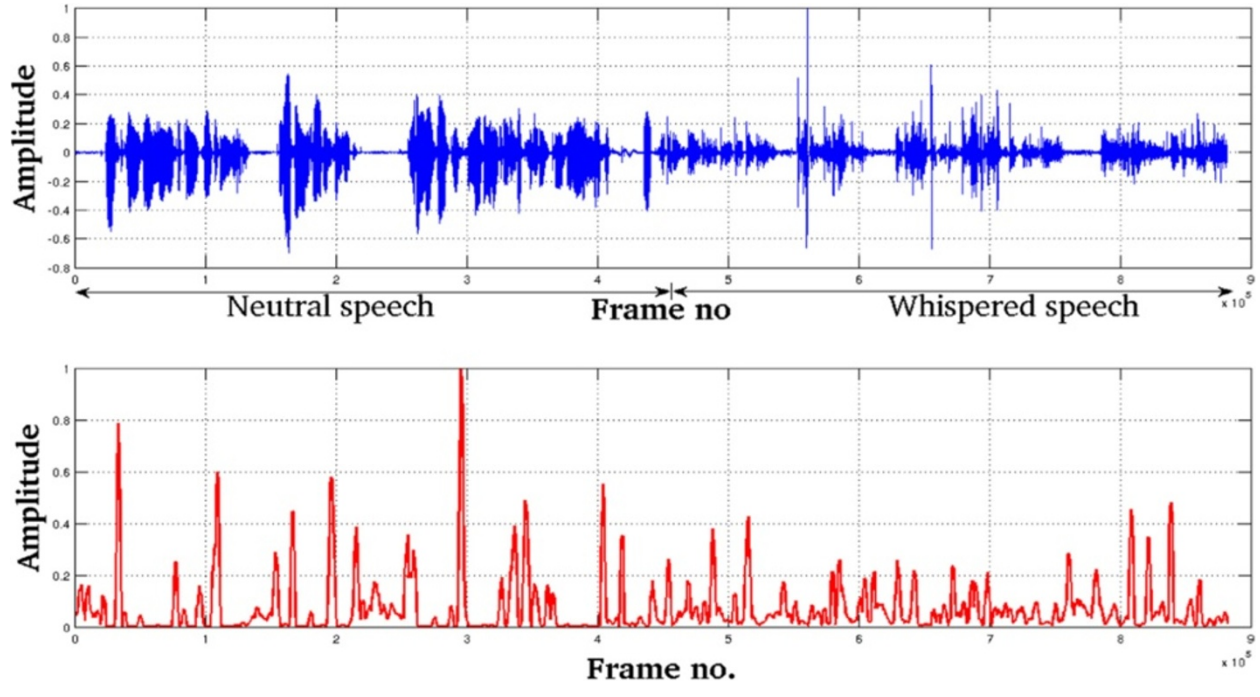

Figure 3 Plot of SFM (bottom) with the corresponding speech segment (top). Whispered speech shows higher SFM than neutral speech. Also the short pauses exhibit a very high SFM.

LP-MVDR spectral ratio method for detection of whispered speech

Linear prediction [19], is the most widely used and accepted method, for whisper detection [15] and also whisper island detection [5]. Other methods like vocal effort change point detection have also been used for improved whisper detection within normally phonated audio streams in this context [20]. These methods in general parameterize an AR spectra using a least square errors method. However the LP model works well only for speech formants as in voiced speech and more specifically at low frequencies. The most undesirable effect of these techniques is that the LP model tends to overestimate the power at formant frequencies. Moreover increasing the model order increases the overestimation rather than correcting it. Hence this method is able to resolve harmonics but is poor at estimating the power at the formant frequencies in the spectrum. Hence it leads to poor characterization of the vocal tract transfer function. The MVDR spectra [21], on the other hand, is capable of modeling the power of the spectra efficiently at all harmonic frequencies due to the nature of the estimation method. The MVDR spectra also responds to increase in model order and improves the model at higher harmonic frequencies [22]. As whisper is characterized by formant shifts and overall increased concentration of power in high frequency bands, the models must essentially provide good modeling results in those bands. Thus exploiting the use of MVDR spectra is desirable in robust whisper detection. Moreover, MVDR coefficients can be computed from LP coefficients themselves [23], making the process computationally less expensive. In the following section, a brief description of the significance of LP and the MVDR methods in spectrum estimation of whispered speech is discussed followed by the proposed parametric spectral ratio method.

LP method for detection of whispered speech

In LP, we approximate speech as an auto-regressive process. Hence a model $\hat{s}(n)$ discrete time signal $s(n)$ is given by

$$\hat{s}(n) = \sum_{k=1}^{m} a_k s(n-k) \tag{8}$$

where 'm' denotes the order of the AR model and a_k's are the model constants that are to be estimated. When predicting a signal using the AR model, the error from the actual signal $e(n)$ can hence be written as

$$e(n) = s(n) - \sum_{k=1}^{m} a_k s(n-k) \tag{9}$$

To obtain the parameters a_k, least square error technique is used. The idea is to minimize the total error with respect to each of the parameters. Hence

$$\text{LSE} = \sum_n e(n)^2 = \sum_n \left(s(n) - \sum_{k=1}^{m} a_k s(n-k) \right)^2 \tag{10}$$

To find parameters a_k;

$$\frac{\partial(\text{LSE})}{\partial a_k} = 0 \quad \forall\, k \in [1, m] \tag{11}$$

using Equation (10) in Equation (11) and expanding, we get

$$\sum_{k=1}^{m} \left(a_k \sum_n s(n-k)s(n-i) \right) = \sum_n s(n)s(n-i) \quad \forall\, i \in [1, m] \tag{12}$$

Equation (12) is an equation set with 'm' variables and 'm' unknowns and hence can be solved to get the required parameters. Interms of autocorrelation function Equation (12) then can be written as

$$\sum_{k=1}^{m} a_k R(i-k) = R(i) \quad \forall\, k \in [1, m] \tag{13}$$

where R(i) denotes the autocorrelation over signal $s(n)$

$$R(i) = \sum_{n=-\infty}^{\infty} s(n)s(n-i) \tag{14}$$

The autocorrelation matrix hence formed by R(i–k) is a Toeplitz matrix. In matrix form Equation (13) can be written as

$$\mathbf{Ra=r} \tag{15}$$

The solution to the Equation (13) is found using Levinson-Durbin recursion algorithm. The method requires only 2m storage and the time complexity is of the order of $p^2 + O(p)$.

Inadequacy of LP method for detection of whispered speech

To understand the inadequacy of LP in the context of whispered speech; we need to analyze the spectrum in spectral domain. The inadequacy can be traced back to the least square error method that was employed to determine the parameter a_k's. The merit of least squares method as described above in any error analysis is its ability to amplify big errors and diminish small errors of parameter. To begin with, in spectral domain Equation (9) becomes

$$E(z) = A(z)S(z) \tag{16}$$

where

$$A(z) = 1 - \sum_{k=1}^{m} a_k z^{-k} \tag{17}$$

The total error E using Equation (10) can be written as

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega \tag{18}$$

Also the real Power

$$P(\omega) = |S(e^{j\omega})|^2 \tag{19}$$

Putting in terms of $P(\omega)$, the total error E can be written as

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega)A(e^{j\omega})A(e^{-j\omega})d\omega \tag{20}$$

Following the procedure of error minimization using least squares method, we can arrive at

$$R(i) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega)\cos(i\omega)d\omega \tag{21}$$

This is because $P(\omega)$ is real and even. Further, from Equations (19) and (16)

$$P(\omega) = \left|\frac{E(e^{j\omega})}{A(e^{j\omega})}\right|^2 \tag{22}$$

It is well known that an infinite order AR model can always model the signal arbitrarily closely. But limiting the order of the AR model to 'm', leads to an approximation. Also as the signal is finite, the error can only be minimized at best with a finite order AR model as seen before. Let the minimum possible error for this model be $E_{\min}$. Hence the equation becomes

$$\hat{S}(z) = \frac{E_{\min}}{1 - \sum_{k=1}^{m} a_k z^{-k}} \tag{23}$$

with non zero $E_{\min}$ and the transfer function being $A(z)$. The estimated power $\hat{P}(\omega)$ of the speech signal $\hat{S}(z)$ can now be calculated by taking the square of the modulus as

$$\hat{P}(\omega) = \frac{|E(\omega)|^2}{|1 - \sum_{k=1}^{m} a_k e^{-j\omega k}|^2} \tag{24}$$

The total error in prediction as in Equation (10) can hence now be written as

$$E = \frac{1}{2\pi}\int_{\frac{-\pi}{T}}^{\frac{\pi}{T}} |E(\omega)|^2 \, d\omega \tag{25}$$

which gives net error in terms of the errors in spectral domain. To convert this equation in terms of the predicted and actual powers, using substitution from Equations (23) and (16) we arrive at

$$E = \frac{E_{\min}^2 T}{2\pi}\int_{\frac{-\pi}{T}}^{\frac{\pi}{T}} \frac{P(\omega)}{\hat{P}(\omega)} d\omega \tag{26}$$

Thus minimizing the total error E is equivalent to minimizing the ratio of the actual power to the estimated power integrated over the entire interval. Hence the LP problem can now be viewed in terms of the ratio of power spectra and their minimization over the interval.

The coefficients that are to be subsequently calculated via the least squares method will be found using the R(i)'s whose relation with the power spectra has already been established in Equation (21). In analogy, the expected correlation coefficients can be found using predicted power spectra using

$$R(i) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{P}(\omega)\cos(i\omega)d\omega \tag{27}$$

This concludes that since the spectrum depends upon the autocorrelation; the window over which it is computed becomes immensely important. This is because the process of generating the signal must be such that it can be considered stationary for that duration of time over which it is calculated.

Another important and more significant conclusion in the light of whispered speech can be analyzed through the error equations. The integral of the ratio of powers is to be brought out to some fixed value. This means that the ratios will be less than one for some time and greater than one for another to compensate the two. Moreover no regard is paid to the fact that some parts of the spectra will have low energy and others will have high energy and the minimization is applied uniformly. For that let us make two cases with respect to the spectral ratio.

Case 1: estimated power ($\hat{P}(\omega)$) is over estimated by ϵ times the actual power ($P\omega$))

The ratio within the integral in this case turns out to be

$$\frac{(P(\omega))}{(P(\omega))(1+\epsilon)} = \frac{1}{1+\epsilon} = 1 - \frac{\epsilon}{1+\epsilon} \tag{28}$$

Case 2: estimated power ($\hat{P}(\omega)$) is under estimated by ϵ times the actual power ($P(\omega)$)

The ratio within the integral in this case turns out to be

$$\frac{(P(\omega))}{(P(\omega))(1-\epsilon)} = 1 + \frac{\epsilon}{1-\epsilon} \tag{29}$$

Comparing the two above mentioned cases, the effect on error is more when $P(\omega) > \hat{P}(\omega)$ than when the inequality sign is reversed. Hence the error in *Case*2 is much larger than in *Case* 1. However error minimization strategies do not take these cases into account resulting in overestimation at certain frequencies. Since LP gives a relatively smooth polynomial, the compounded 'less than one' parts of the spectra are liable to compensate at the harmonics. This leads to overestimation at crucial harmonic frequencies is liable to give poor modeling results. As in whisper detection as well as recognition, the focus is mainly on correct detection of 'how good' the harmonics are, an overestimation is an undesirable result. The inaccurate modeling at high frequencies by LP also compounds the problem for whisper detection [22].

The MVDR method for detection of whispered speech

Minimum variance distortion less response spectrum was introduced by Capon, and is also known as the Capon spectrum, or the maximum likelihood method (MLM) spectrum. The MVDR spectrum is a well-known method in array processing applications, and appears to be promising in other applications such as speech. MVDR can be looked upon as a filter design topology in which the aim is to design a filter bank that satisfies a certain constraint known as the 'distortion less' constraint centered at one of the analysis frequencies [23,24]. This means at the frequency of interest(ω_l); the gain of the transfer function should be unity i.e.

$$H(e^{\omega_l}) = \sum_{k=0}^{M} h(k)e^{-jk\omega_l} = 1 \tag{30}$$

In addition to this, the output's variance must be minimized subjected to the above constraint. Hence the resultant filter for a frequency ω_l is obtained by solving the optimization problem

$$\min_{h(n)} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{uu}(e^{j\omega}) d\omega \right) \quad \text{subj. } to \;\; H(e^{j\omega_l}) = 1 \tag{31}$$

where $S_{uu}(e^{j\omega_l})$ is the spectral power at the frequency ω_l. This ensures that MVDR will faithfully preserve the input signal at the frequency ω_l. This is a major difference between MVDR and LP because it can give faithful response even at high frequencies. However the filter designing problem in case of MVDR is just conceptual. It can be shown that MVDR spectrum for all frequencies can be calculated as

$$R_M^{\text{mvdr}}(e^{j\omega}) = \frac{1}{\mathbf{v}^H(\omega)\mathbf{R}_x{}^{-1}\mathbf{v}(\omega)}, \tag{32}$$

where $\mathbf{R}_x$ is the $(M)\times(M)$ data autocorrelation matrix and

$$\mathbf{v}(\omega) = [\,1, e^{j\omega}, e^{j2\omega}, e^{j3\omega}, \ldots\ldots, e^{j(M-1)\omega}]^T. \tag{33}$$

This estimate has some interesting properties which we briefly mention below. It can be efficiently computed exploiting the relationship with LP methods as

$$R_M^{\text{mvdr}}(e^{j\omega}) = \frac{1}{\sum_{k=-M}^{M} \mu(k)e^{-j\omega k}} \tag{34}$$

where the parameters $\mu(k)$ are obtained by a simple non iterative computation involving the LP coefficients [25] by minimizing the prediction error variance R_e, as

$$\mu(k) = \begin{cases} \frac{1}{P_e}\sum_{i=0}^{M-k}(M+1-k-2i)a_i a_{i+k}^* & k = 0 \ldots M \\ \mu^*(-k) & k = -M.\ldots -1 \end{cases} \tag{35}$$

The MVDR method is analogous to the periodogram whose estimate at any frequency can be interpreted as the output of a band pass filter centered at that frequency. The periodogram can hence be interpreted as a band of band pass filters which are data and frequency independent like most non-parametric methods. MVDR can also be considered as a band of filters constrained to a set of conditions. However this filter is both data and frequency dependent. Its robustness as a recognition feature has been verified in [26] in this context.

Dependence of the MVDR method on model order in the context of whispered speech

Let us assume voiced speech signal to be perfectly harmonic for a short duration. Here we are oversimplifying the speech for analysis just at the harmonics [24]. The harmonics can be modeled as

$$b(n) = \int_{k=1}^{L} c_k \cos(\omega_0 kn + \phi_k) \tag{36}$$

where L is the number of harmonics. The pitch of such a system can be clearly seen to be $f_0 = \frac{\omega_0}{2\pi}$. The correlation will hence be

$$r_{uu}(m) = \sum_{k=1}^{L} \frac{|c_k|^2}{4} \cos(\omega_0 km) \tag{37}$$

The MVDR filter $h_l(n)$ designed for lth harmonic will hence try to preserve the input power at that frequency faithfully while trying to have minimum power at the other harmonics.

$$P_{MV}(\omega_0 l) = \sum_{k=1}^{L} (|H_l(e^{j\omega_k})|^2 + |H_l(e^{-j\omega_k})|^2) \frac{|c_k|^2}{4} \tag{38}$$

For MVDR filter at the frequency ω_l;

$$P_{MV}(\omega_0 l) = \frac{|c_l|^2}{4} + |H_l(e^{-j\omega_l})|^2 \frac{|c_l|^2}{4} + \sum_{k=1, k\neq l}^{L} (|H_l(e^{j\omega_k})|^2 + |H_l(e^{-j\omega_k})|^2) \frac{|c_k|^2}{4} \tag{39}$$

If the MVDR filter has M filter zeros and $M > 2L-1$ then the MVDR has enough zeros to cancel all the other input exponentials and can estimate the exact power at the harmonic. Otherwise there will be a positive bias to the spectral estimate. Hence spectral estimates are bound to get better at every frequency with increase of model order. This is in direct contrast to LP spectra where increase in order leads to more over-estimation at the harmonics [22].

Proposed LP-MVDR spectral ratio method for detection of whispered speech

It is important to note that MVDR spectrum is a smoother spectrum when compared to the LP spectrum. This is on

account of the fact that an MVDR spectrum at any frequency can be represented as a harmonic average of the LP spectra of a particular order [27].

$$\frac{1}{P_{MV}(\omega)} = \sum_{k=0}^{p} \frac{1}{P_{LP(k)}(\omega)} \tag{40}$$

This averaging effect smooths out the spectrum at the regions of sharp rise i.e. at the harmonics. Thus the MVDR spectrum tends to have lower amplitude than that of corresponding LP spectra at the harmonics. Figure 4 shows the various spectra for a short segment of neutral speech. From the illustration in Figure 5, it is clear that a LP to MVDR ratio spectrum can be used to identify the whisper segments in speech, since this ratio is expected to be high where the speech signal has significant harmonics in the higher frequency region than in the neutral speech spectrum. In the context of whispered speech where the harmonic shifts are prominent, this ratio is expected to be high. Also the ratio is expected to be robust to wide band noise because the LP spectrum of high order can still model the spectrum and the averaging effect of MVDR will eliminate the effect of wide band noise when a spectral ratio is taken. Formally the the LP-MVDR spectral ratio is defined as

$$X(\omega) = \frac{\hat{P_{\mathrm{LP}}}(\omega)}{P_{\mathrm{MV}}(\omega)} = \frac{(|E(\omega)|^2)\left(\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}\right)}{\left|1 - \sum_{k=1}^{m} a_k e^{-j\omega k}\right|^2} \tag{41}$$

The LP-MVDR ratio spectrum is computed on a short time basis for each speech data window. The ratio spectrum is further smoothed and a threshold is decided depending on the penalties fixed based on the false alarm rate and detection failure rate. The whispered segments can then be segmented from the normally phonated segments of speech. The salient steps used for whisper detection using the LP-MVDR ratio spectrum is listed below.

- Hamming Window the test speech waveform using a frame size of 20 ms and a frame overlap of 50%.
- Compute the MVDR coefficients using Equation (35), for each frame.
- Compute the smooth MVDR Power spectrum for sufficient number of frequency points (1024 in our work).
- Compute the linear predictor coefficients for each frame.
- Compute the LP power spectrum for the same number of frequency points as used for computing the MVDR power spectrum.

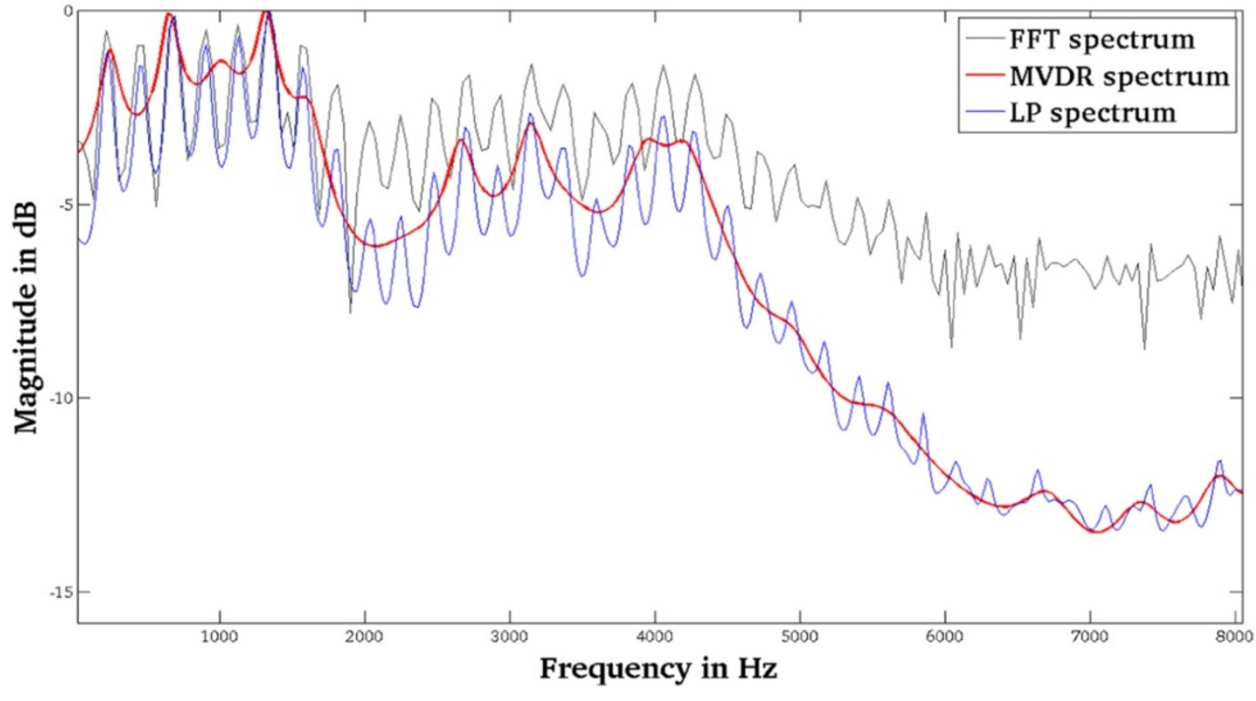

Figure 4 Diagram illustrating the various spectra for a short segment of neutral speech. FFT (black), MVDR (red) and LP (blue) spectra. Overshoot by at the harmonics can be observed clearly.

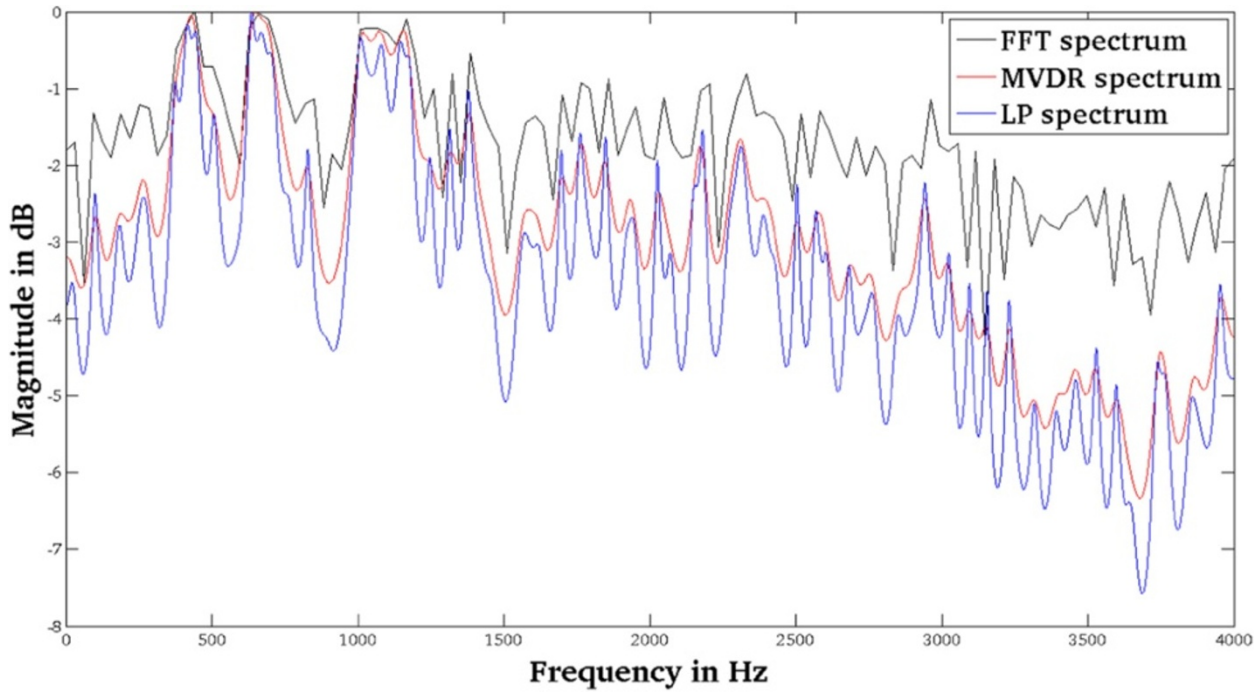

Figure 5 Diagram illustrating the various spectra for a short segment of whispered speech. FFT (black), MVDR (red) and LP (blue) spectra. No clear harmonics can be observed an also the overshoots are not pronounced as in neutral speech.

- Compute the LP to MVDR ratio spectrum in each frame.
- Calculate the maximum of ratio values in each frame and using these values form a vector called maxratio.
- Scale maxratio by its maximum value.
- Select the threshold according the penalties required for false alarm rate and detection failure rate.
- Segment whispered speech at boundaries of change.

The flow diagram illustrating the proposed sequence of steps is given in Figure 6: A long segment of speech containing both neutral and whispered segments is considered. The LP-MVDR ratio spectrum is computed for such a speech segment. Figure 7 illustrates a smooth LP-MVDR Ratio spectrum for a speech signal containing both neutral and whispered segments. From the LP-MVDR ratio histograms of neutral speech and whispered speech, shown in Figure 8, it is clear that the whispered segment exhibits a higher value of spectral ratio than the neutral segment.

However the proposed method is not able to detect the exact points where the speaker shifts from normal speech to whispered speech and vice versa. This can lead to segmentation issues in the subsequent recognition process. Hence there is a need to explore the possibility of fine tuning the position where the speaker has shifted from one mode of speech to another. Hence we propose to use the BIC. This method is discussed in the subsequent section.

Refining the whisper segment boundaries using BIC

To overcome the segmentation issues in detecting neutral to whisper or whisper to neutral change point, we apply BIC on LP-MVDR ratio obtained for the speech signal. The BIC is a model selection criteria that helps to maximize the model performance. Model selection includes the best possible selection of variables to come up with the best statistical model for given variable set. Issues like over-fitting might eventually lead to less efficient models. Another criteria called Akaike information criteria is also used but it imposes lower penalty increase of model parameters and hence is not preferred [28]. Use of BIC as listed in Equation (42), in speaker segmentation has been extensively used in literature [29,30].

$$\mathrm{BIC} = \ln L - \frac{1}{2} \cdot \lambda \cdot k \cdot \ln(n) \tag{42}$$

where L is the maximized value of likelihood function for the estimated model, n is the number of data points, k is the number of free parameters to be estimated and λ is the penalty factor assumed 1 in our case. Given any two estimated models, the model with the lower value of BIC is preferred. Under the assumption that the model errors are i.i.d and that the boundary condition indicating the derivative of the log likelihood in respect to the true variance is equal to zero, the BIC criterion can be rewritten as

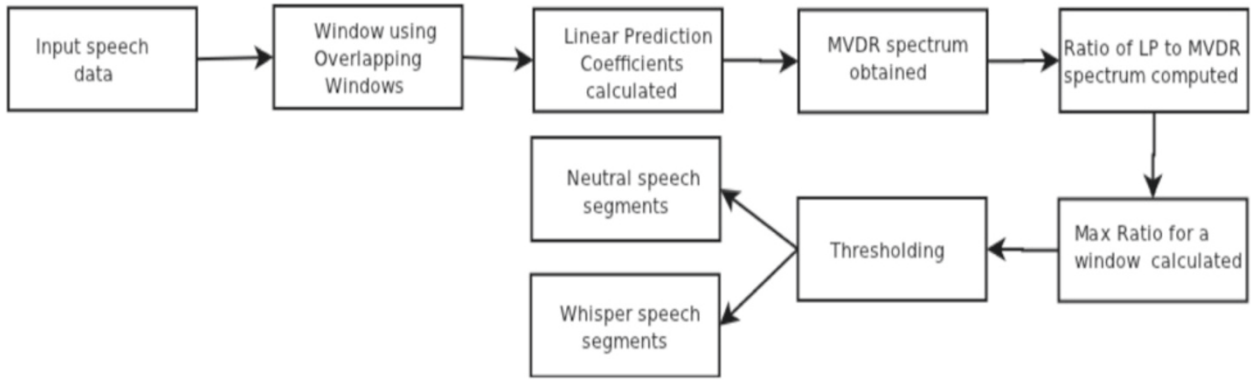

Figure 6 Block diagram showing the scheme for detection of whispered speech using proposed LP-MVDR technique. Overlap is ideally taken as 50%.

$$\mathrm{BIC} = n \cdot \ln(\hat{\sigma}_e^2) + k \cdot \ln(n) \tag{43}$$

Hence ΔBIC can be effectively used to find the change in vocal effect. Given two audio segments, $X = x_1, x_2 \ldots x_{n_x}$ *and* $Y = y_1, y_2 \ldots y_{n_y}$, the binary hypothesis testing problem in this context of whisper detection can be formulated as

$$\begin{aligned} H_0 :\ & x_1, x_2 \ldots x_{n_x}, y_1, y_2 \ldots y_{n_y} \sim N(\mu, \Sigma) \\ H1 :\ & x_1, x_2 \ldots x_{n_x} \sim N(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \\ & \text{and } y_1, y_2 \ldots y_{n_y} \sim N(\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}}) \end{aligned} \tag{44}$$

H_0 is the hypothesis that claims X and Y to belong to the same multivariate Gaussian distribution while H_1 claims it to be derived from two distinct multivariate Gaussian Distributions. Hence ΔBIC values can be computed as the difference in the BIC values of H_1 and H_0.

$$\begin{aligned} \Delta\mathrm{BIC} &= \mathrm{BIC}(H_0) - \mathrm{BIC}(H_1) \\ &= \ln(p(X|\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})) + \ln(p(Y|\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}})) \\ &\quad - \ln(p(Z|\mu, \Sigma)) - \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\ln(n) \\ &= \frac{n}{2}\ln(|\Sigma|) - \frac{n_x}{2}\ln(|\Sigma_{\mathbf{x}}|) - \frac{n_y}{2}\ln(|\Sigma_{\mathbf{y}}|) \\ &\quad - \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\lambda ln(n) \end{aligned} \tag{45}$$

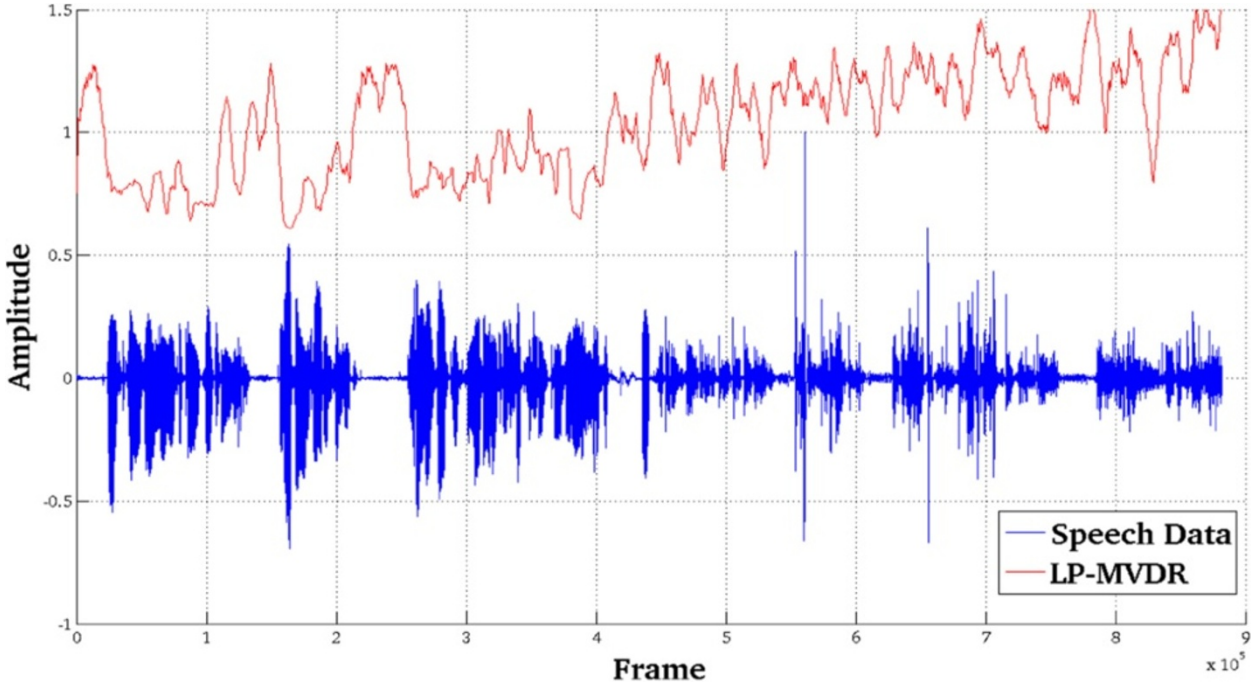

Figure 7 Plot of smoothened LP-MVDR Ratio spectrum (top) with the speech signal (bottom) containing both neutral and whispered segments. Whispered segment exhibits a higher ratio value than the neutral segment.

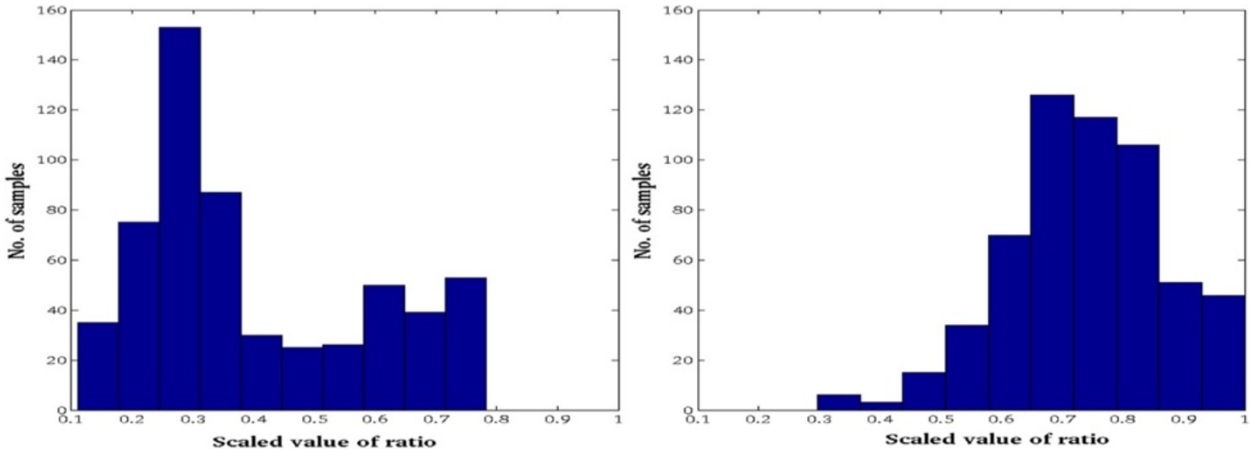

Figure 8 LP-MVDR ratio histograms for neutral speech (left) and whispered speech (right) of the same speaker speaking same sentence.

where d is the dimension of the feature vector, λ is the penalty factor assumed 1 in our case. The larger the value of ΔBIC, any two segments can be considered the most dissimilar. For a single known change point detection the boundary segments for $i_{\min} < i < n - i_{\min}$ can be written as

$$\Delta\mathrm{BIC}(i) = \frac{n}{2}\ln(|\Sigma|) - \frac{i}{2}\ln(|\Sigma_{\mathbf{x_i}}|) - \frac{n-i}{2}\ln(|\Sigma_{\mathbf{y_i}}|) - \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\ln(n) \tag{46}$$

If $\Delta\mathrm{BIC}(i) > 0$, then the change is occurred and the point where change is occurred is the time index which has maximum value of ΔBIC, else no change point in the data. It is to be noted that a minimum value of $i_{\min}$ needs to be ascertained for which the number of data points are insufficient to give a reliable estimate of BIC. Hence empirically it is sufficient to keep $i_{\min}$ to be 30–50 data points [31,32] in the context of whisper detection, so we used a fixed search window with initial length of 50 data points. From Figure 9, it is clear that, if there is a change point then $\Delta\mathrm{BIC}(i) > 0$.

Whispered speech recognition using LP-MVDR spectral ratio method and MLLR

A general methodology followed in whispered speech recognition systems is to segment speech at the whisper boundaries and subsequently use statistical models like the hidden Markov model (HMM) to perform recognition. A block diagram of proposed ASR system incorporating the speech mode changes between neutral and whispered speech is illustrated in Figure 10. Developing large databases for whispered speech is neither practical nor cost-effective. Hence a system that can use a small number of files to model transformations from one mode to another is important. An attempt in this direction is made in the form of application of existing maximum likelihood linear regression (MLLR) adaptation which has been primarily used to model the differences due to changes in environment or non nativeness of speakers. In the following section a brief description of the MLLR adaptation technique is described.

Model adaptation using MLLR

Speaker adaptation methods aim to adapt data independent statistical models to more specific models (for a particular speaker or environment) using only a small amount of new adaptation data. MLLR is one such technique. It estimates some linear transformations for groups of model parameters to maximize the likelihood of adaptation data. The linear transformations shift the component means and alter the variances in the independent system so that each state in the HMM system is more likely to generate the adaptation data [33,34].

Methodology

A simple transformation method illustrated in Equations (47) and (48), generally applies the transformation to the mean with a fixed bias while keeping the covariance matrix the same.

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b} \tag{47}$$

$$\hat{\Sigma} = \Sigma \tag{48}$$

where **A** can either be a diagonal or a full transformation matrix. Other techniques modify both means and variances using a bias and a transformation matrix. The method proposed in this work uses a transformation matrix that computes an adapted mean and is given by

$$\hat{\mu} = \mathbf{W}\xi \tag{49}$$

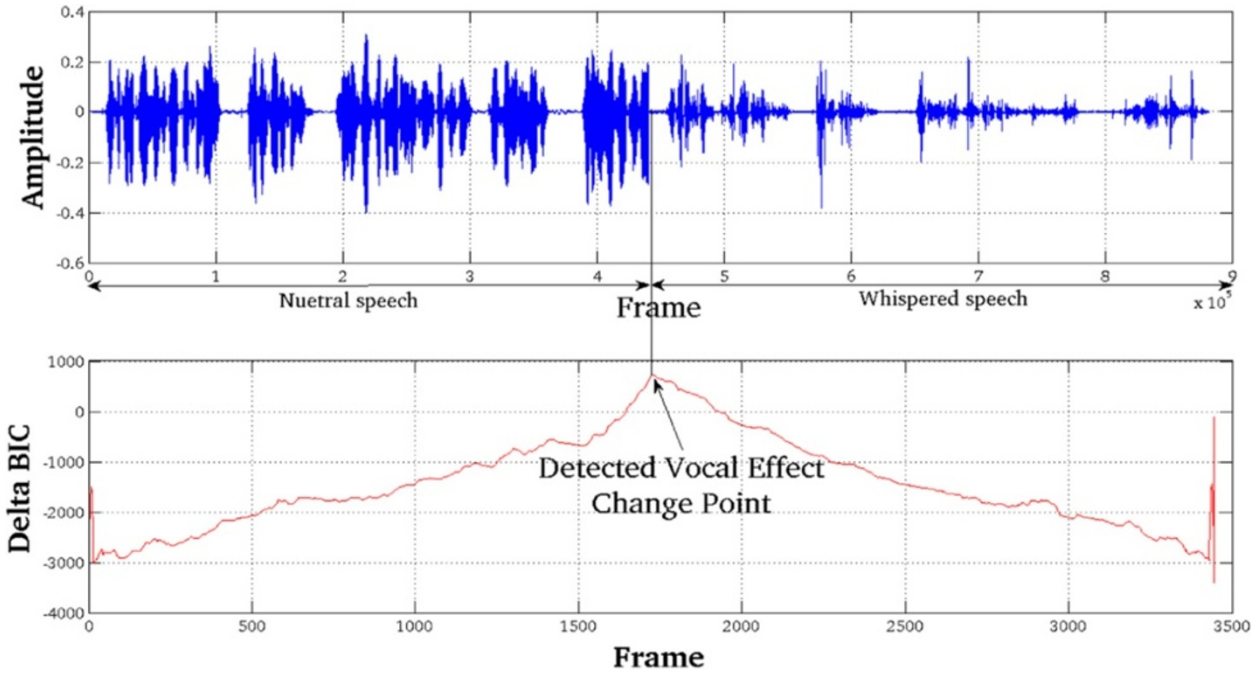

Figure 9 Plot of BIC value (bottom) with the speech signal (top) containing neutral and whisper speech. The BIC value is greater than zero in sections that are probable vocal effect change points with the maxima being the detected change point.

where $\mathbf{W}$ is the $n \times (n+1)$ transformation matrix and ξ is the extended mean vector given by

$$\xi = [1 \quad \mu_1 \quad \dots \quad \mu_n]^T \tag{50}$$

Therefore

$$\mathbf{W} = [\mathbf{bA}] \tag{51}$$

For linear normalization the transformed variance is of the form [35]

$$\hat{\Sigma} = \mathbf{H}\Sigma\mathbf{H}^{\mathbf{T}} \tag{52}$$

where $\mathbf{H}$ is the $n \times n$ transformation matrix. For modeling the speech data according to the above described procedure [36], HMM based normalization of means and variances was used. in the succeeding section the results of performance evaluation conducted on both detection and recognition of whispered speech are discussed.

Performance evaluation: detection of whispered speech

Two databases were used to analyze the performance of the proposed spectral ratio method for whispered speech detection namely, the CHAINS corpus and a whispered speech corpus collected over the cell phone. In the subsequent sections, the description of the databases used and the experimental results on whisper detection using ROC curves are described. The performance of the proposed whisper detection algorithm is also presented as the whisper diarization error rate.

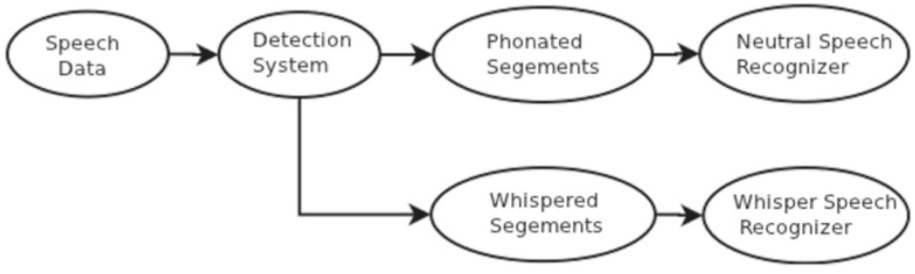

Figure 10 Block diagram of proposed system for speech recognition incorporating the speech mode changes between neutral and whispered speech.

CHAINS database

CHAINS is a research project funded by Science Foundation Ireland from April 2005 to March 2009 [37]. Its goal is to advance the science of speaker identification by investigating those characteristics of a persons speech that make them unique. The corpus was recorded in the early phase of the project. The corpus freely available to researchers, both through its website and through the linguistic data consortium. The corpus features approximately thirty six speakers recorded under a variety of speaking conditions, allowing comparison of the same speaker across different well-defined speech styles. Speakers read a variety of texts alone, in synchrony with a dialect-matched co-speaker, in limitation of a dialect-matched co-speaker, in a whisper, and at a fast rate. There is also an unscripted spontaneous retelling of a read fable. The bulk of the speakers were speakers of Eastern Hiberno-English.

Recording conditions

The solo recording session was carried out in a professional recording studio in December 2005 and speakers were recorded in a sound-attenuated booth. The recordings in the released corpus were done using a Neumann U87 condenser microphone. The whisper recording session from March 2006 to May 2006 was carried out in a quiet office environment, using an AKG C420 headset condenser microphone. In solo reading the speakers were asked to read a prepared text at a comfortable rate and volume. In whisper recording, the speakers read all text in a whisper. Any involuntary switch to modal voicing was interpreted as a disfluency and led to a restart of the phrase.

Text categories in the CHAINS corpus

The corpus texts can be divided into two categories. First category contains famous fables recorded in continuous speech. The second section contains short sentences. In order to provide good phonetic coverage, there are thirty three individual sentences wherein nine are selected from the CSLU Speaker Identification corpus, and twenty four from the TIMIT corpus.

Cellphone whispered speech data corpus

This corpus was developed at Indian Institute of Technology Kanpur. The speakers were fluent in English with an Indian accent living in on campus. The speakers belonged to different parts of India and hence of varied Indian accents. The data was recorded using cell phone calls placed over an open source Asterisk server.

Hardware setup

The IVRS system was hosted on a Intel Orgi. G31 machine. The telephony card used was Sangoma's A200/-Remora FXO/FXS Analog AFT card. This card supports up to 2.048 Mbps of full duplex data through-put and up to thirty voice calls over a E1 line. The files were recorded in uncompressed wave file format. The sampling rate was 8 kHz with 16-bit values per sample. A single channel was used for the recording. The encoding of data was in 16-bit signed integer Pulse-code modulation (PCM).

Text categories in cellphone whispered speech data corpus

The corpus texts are divided into two categories. The first category contains recordings of the five TIMIT sentences recorded in neutral and whispered modes. The sentences spoken are listed in Table 1. Second category contains digits from 0–9 spoken one at a time for both modes. A total of seventy speakers performed the recording.

Experiments on detection of whispered speech

Whisper detection performance was evaluated using the fables part of the CHAINS corpus and sentences section of the cellphone whispered speech data corpus. One neutral part and one whispered part (10 s each) of the same speaker were concatenated at a fixed interval. Thirty such concatenated sentences, fifteen from Cellphone Whispered Speech Data Corpus and fifteen from CHAINS database, were formed and detection was performed on each. Scaled additive white Gaussian noise (AWGN) is used to simulate the various signal to noise ratios (SNRs) and the SNRs are calculated from the concatenated sentence. To calculate SNRs we used the equation $S_n[n] = S_c[n] + \alpha w[n]$, where $S_n[n]$ is the noisy signal, $S_c[n]$ is the clean signal, $w[n]$ is the AWGN noise and α is a variable. By varying α we get different SNRs. Two performance measures were used in the evaluation. Receiver operating characteristic (ROC) curve which is a plot of true positive rate (TPR) versus false positive rate (FPR) and the whisper diarization error rate (WDER) were used herein. However TPR is defined as

$$\text{TPR} = \frac{\text{No. of correctly detected whisper segments}}{\text{Total no. of whisper segments}} \tag{53}$$

while the FPR is defined as

$$\text{FPR} = \frac{\text{No. of wrongly detected whisper segments}}{\text{Total no. of neutral segments}} \tag{54}$$

Table 1 TIMIT sentences spoken in cellphone whispered speech data corpus in neutral and whispered modes

S. no.	TIMIT sentences
1.	Don't ask me to carry an oily rag like that
2.	Call an ambulance for medical assistance
3.	Special task forces rescue hostages from kidnappers
4.	A boring novel is a superb sleeping pill
5.	The sermon emphasized the need for affirmative action

On the other hand WDER computes the diarization error similar to that used in speaker diarization. A performance index similar to the one in [20], is used in this work. The possible errors in whisper detection are generally the false alarm (FA) and the detection failure (DF). Hence the WDER using the above sources of error can be defined as

$$\text{WDER} = \frac{C_1 \cdot \text{FA} + C_2 \cdot \text{DF}}{N_f} \tag{55}$$

where N_f denotes the total number of speech frames, and C_1 and C_2 are the weights assigned to FA and DF respectively, where $FA = FPR$ and $DF = 1 - TPR$. Note that C_1 and C_2 are selected based on the penalties fixed for false alarm rate and DF rate, respectively.

Experimental results on CHAINS database

The receiver operating curves (ROC) are computed for the proposed and other conventional techniques like SFM and Linear predictive residual spatial audio coding (LPRSAC) under different SNRs. From Figure 11, we can see that all the techniques perform reasonably well in the absence of noise. LP to MVDR (LP-MVDR) technique performs better than both the conventional techniques, SFM and LPRSAC. The initial poor response of spectral flatness can be attributed to presence of short pauses in the data which show a high spectral flatness value. This can often lead to their classification as whispered speech segments. To calculate WDER, individual thresholds are set on each method such that all the methods give equal TPR and the FPR at this TPR is calculated from the ROC curve. WDER is determined from the FPR and TPR. The WDER after choosing appropriate thresholds is given in Table 2.

The results are indicative of a better performance by LP-MVDR spectrum over other conventional techniques. Without addition of noise, in Figure 11 we can see that all the methods seem to perform well. At an SNR of 50dB shown in Figure 12, the performance of all the techniques worsens. The LPRSAC seems to fail at this noise with very poor response. However LP-MVDR and spectral flatness methods still perform reasonably well. At a lower SNR of SNR = 45 dB, no change is observed in spectral flatness spectrum as shown in the Figure 13. The LPRSAC almost coincides with the reference line in an ROC curve and hence gives completely random results. LP-MVDR still performs reasonably well with a degradation as can be observed from Figure 12. At a SNR of 35 dB, LPRSAC completely fails and in fact shows an inverted characteristic shown in Figure 14. The spectral flatness response still remains consistent almost following a curve similar to that for a SNR of 45 dB. However the LP-MVDR performance is degraded severely for high TPRs and falls below spectral flatness for these zones. This is probably due to the presence of short pauses which were not removed in the experiments.

Discussion

It was observed that while spectral flatness is very consistent with its performance on addition of white Gaussian noise, LP-MVDR almost always performs better. LPRSAC performs poorly in comparison to either of these techniques under noisy conditions. An initial dip in spectral

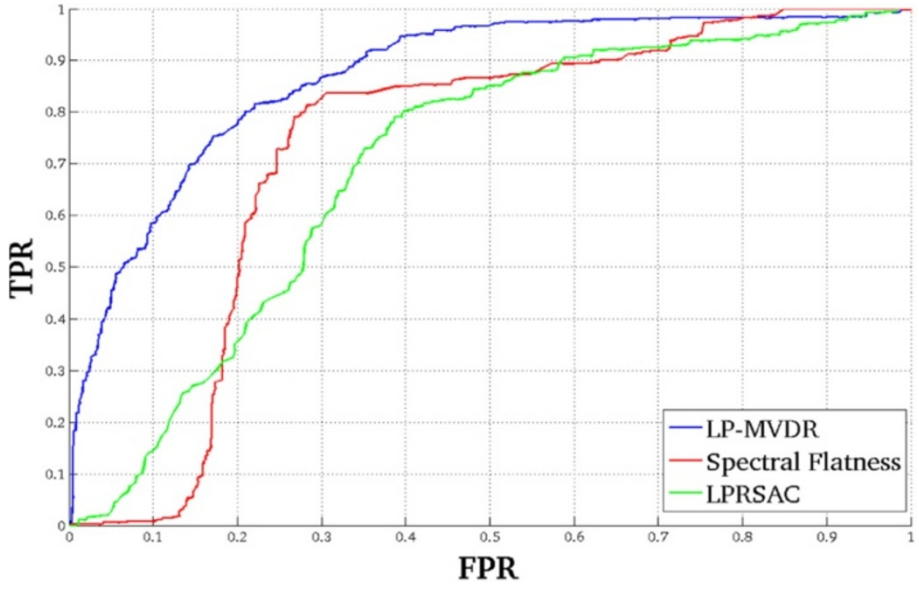

Figure 11 ROC curve for normal speech signal from CHAINS database containing whisper and neutral segments. Blue color indicates LP-MVDR technique, Red indicates Spectral Flatness and Green LPRSAC.

Table 2 WDER for different techniques at different SNRs on CHAINS database

Condition	LP-MVDR	LPRSAC	Spectral flatness
Normal	0.12	0.3	0.26
SNR = 50 dB	0.14	0.46	0.29
SNR = 45 dB	0.14	–	0.29
SNR = 35 dB	0.40	–	0.30

flatness is observed that can be explained by presence of short pauses in speech waveform which is mostly the background noise. Since spectral flatness will give a high value for white noise because of its wide band flatness, the segments are categorized as whispered although they might have been parts of neutral speech. LP-MVDR shows a sharp rise in TPR with little rise in FPR in all cases. This is a desirable result and hence validates the efficacy of the proposed method.

Experimental results on cellphone whispered speech data corpus

Similar whisper detection experiments were conducted on the cellphone whispered speech data corpus. The results obtained are presented in this section. The LPRSAC performs poorly for the cell phone corpus. This is probably due to the effect of frequency cut off in telephones that is usually a band pass filter with 300–3300 Hz with center frequency at around 1 kHz. This frequency range not only affects the first harmonic adversely as it is around 400 Hz but also cuts down a large portion of high frequency component of noise that is a characteristic of whispered speech. From Figure 15, it can be noted that LPRSAC method fails even in clean conditions. The proposed spectral ratio method performs poorly at high FPR values. At a SNR of 50 dB as shown in Figure 16, the performance of LP-MVDR improves. The spectral flatness method also shows a steep rise just like in the case the case of clean signal. At a lower SNR of 45 dB, the LP-MVDR spectrum worsens a bit with TPR changing less for a change in FPR. Spectral Flatness is still almost the same with the same steep rise in TPR. At a SNR of 35 dB, the LP-MVDR performance improves. This is probably due to pronounced effect of noise on the whisper segments as discussed before. The spectral flatness method also shows a small improvement in performance (Figures 17 and 18).

Discussion

The results in Table 3, indicate an overall improvement in WDER results is observed as the SNR is increased. This trend is seen in both spectral flatness as well as LP-MVDR ratio. In order to corroborate these results with the ROC, it can be seen that the TPR rises rapidly with little increase in FPR leading to a very low false alarm rate at a high TPR. The results can also be alluded to the fact that the detection methods were designed to detect the non harmonic content in the signal. However it must be noted that these results are for stationary noise. The analysis of these methods in non stationary and non Gaussian type of noise has not been studied in this work. Given the nature of the proposed technique when compared to other conventional techniques, the results are interesting since the performance of the whisper detection improves in noise. On the other hand the phonated parts of speech would be effected in exactly the opposite manner.

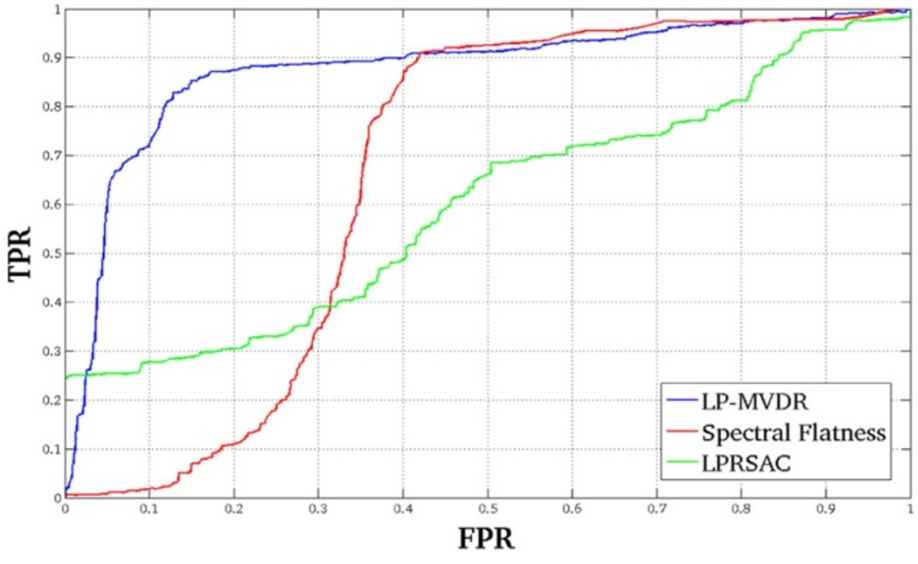

Figure 12 ROC curve for speech signal with SNR= 50 dB from CHAINS database containing whisper and neutral segments. Blue color indicates LP-MVDR technique, Red indicates Spectral Flatness and Green LPRSAC.

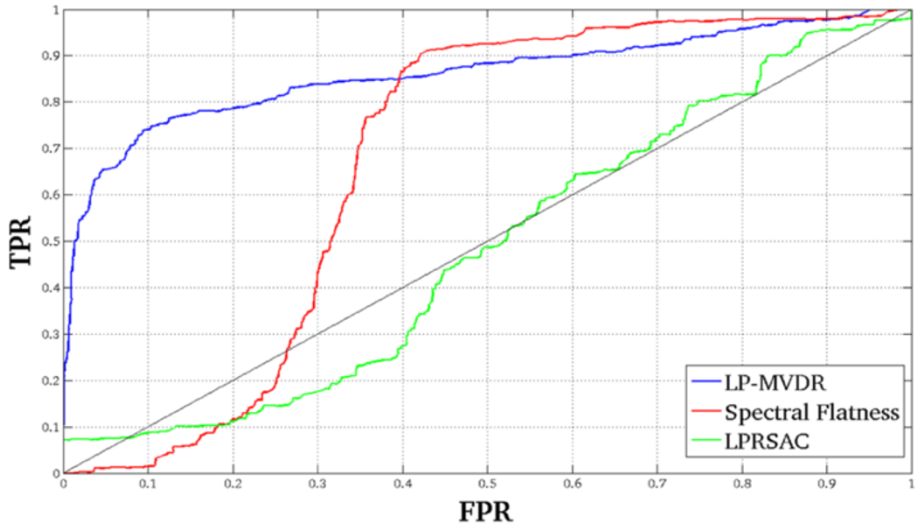

Figure 13 ROC curve for speech signal with SNR= 45 dB from CHAINS database containing whisper and neutral segments. Blue color indicates LP-MVDR technique, Red indicates Spectral Flatness and Green LPRSAC curve for speech signal with SNR = 45.

Performance evaluation: whispered speech recognition

As discussed earlier in Section "Whispered speech recognition using LP-MVDR spectral ratio method and MLLR", a whispered speech recognition system includes both segmenting the speech signal at the whisper boundaries and subsequently use statistical models like the adapted HMMs to perform automatic whispered speech recognition. The purpose is also to use minimal changes in the standard speech recognition engines to implement a whispered speech recognition system. A block digram illustrating the the process of whispered speech recognition is shown in Figure 19. The important blocks as illustrated in Figure 19 are the extraction of features from the speech signal and the MLLR method of adapting neutral speech models to whispered speech models. The adaptation is carried out using MLLR and has been described in Section "Whispered speech recognition using LP-MVDR spectral ratio method and MLLR". The features used in this work are the Mel frequency cepstral coefficients (MFCC). MFCCs are the most commonly used features in speech recognition systems. This makes them the ideal

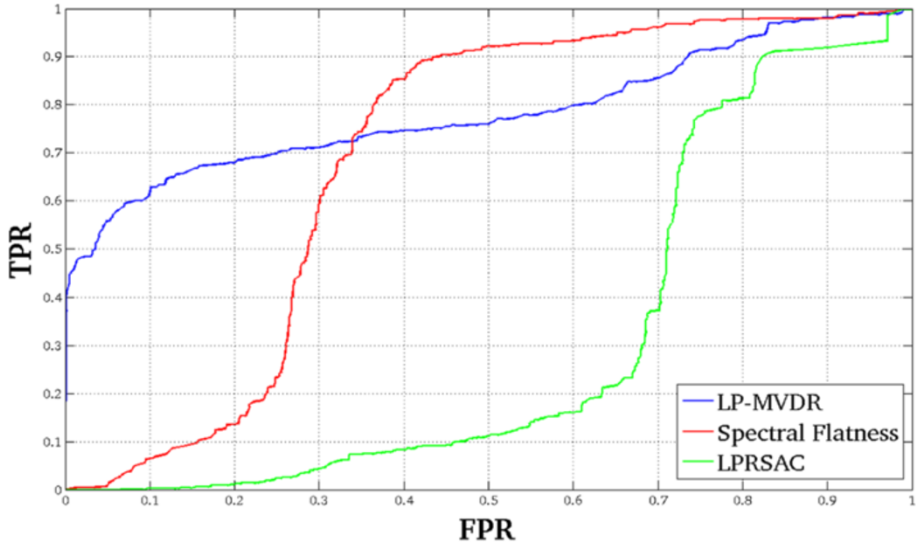

Figure 14 ROC curve for speech signal with SNR= 35 dB from CHAINS database containing whisper and neutral segments. Blue color indicates LP-MVDR technique, red indicates spectral flatness and green LPRSAC.

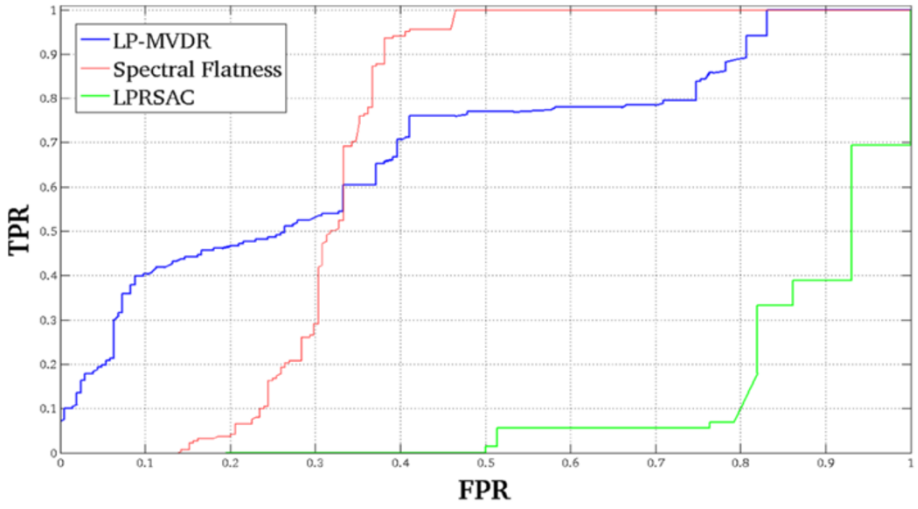

Figure 15 ROC curve for normal speech signal from cellphone whispered speech data corpus containing whisper and neutral segments. Blue color indicates LP-MVDR technique, red indicates spectral flatness and green LPRSAC.

choice for investigating the feasibility of testing the effect of MLLR on whisper recognition. The MFCC features are calculated using the procedure given in [38].

Experiments on whispered speech recognition on the cellphone whispered speech data corpus

Recognition was performed for digits comprising the cellphone whispered speech data corpus. HMM with five states and three mixtures with no tied states were trained for digits spoken in neutral speech in the first stage. Fifty training files were used for each digit in the training process. This HMM was then adapted to whispered speech using MLLR adaptation as discussed earlier. Twenty five whisper files per digit were used in the adaptation process. Twenty one whisper files were used for each digit in the digit recognition process. The experimental results for the recognition scheme are shown in Figure 20, as listed in Table 4. The recognition performance is computed as

$$\text{Recognition} = \frac{\text{Number of correctly recognized digits}}{\text{Total number of digits}} \times 100 \tag{56}$$

The following test cases have been evaluated in the experiments conducted on automatic whispered speech recognition

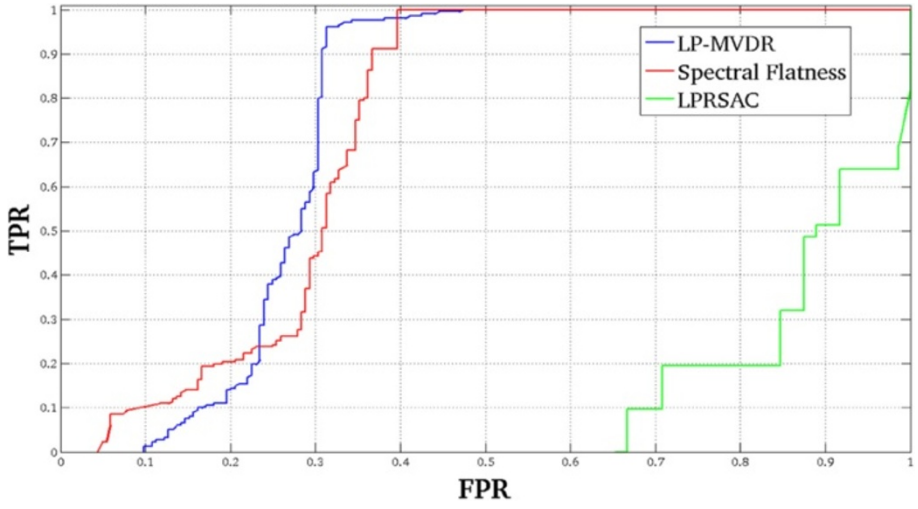

Figure 16 ROC curve for speech signal with SNR=50 dB from cellphone whispered speech data corpus containing whisper and neutral segments. Blue color indicates LP-MVDR technique, red indicates spectral flatness and green LPRSAC.

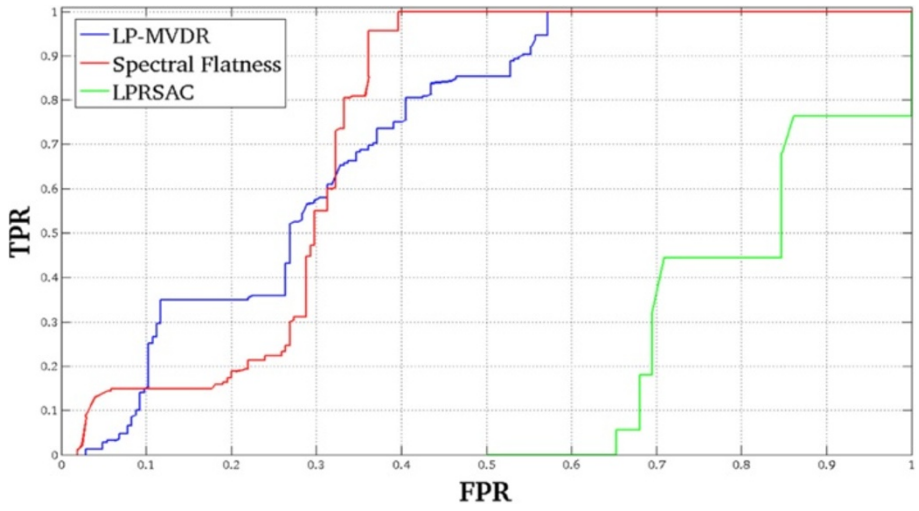

Figure 17 ROC curve for speech signal with SNR=45 dB from cellphone whispered speech data corpus containing whisper and neutral segments. Blue color indicates LP-MVDR technique, red indicates spectral flatness and green LPRSAC.

- **Case 1:** Neutral speech recognition with same train and test data
- **Case 2:** Neutral speech recognition with different train and test data
- **Case 3:** Whispered speech recognition using neutral speech models
- **Case 4:** Neutral speech recognition using neutral speech models adapted to whispered speech data (on similar train and test data)
- **Case 5:** Neutral speech recognition using neutral speech models adapted to whispered speech data (on dissimilar train and test data)
- **Case 6:** Whispered speech recognition using neutral speech models adapted to whispered speech data (on whisper adaptation data)
- **Case 7:** Whispered speech recognition using neutral speech models adapted to whispered speech data (on whisper test data)

The baseline results in Table 4 show a poor performance when neutral speech HMMs are used to recognize whispered speech as expected. Models adapted on whispered speech show an improvement of 12.73% over the baseline. Also it is seen that the performance of adapted models on neutral speech recognition is very poor. Hence

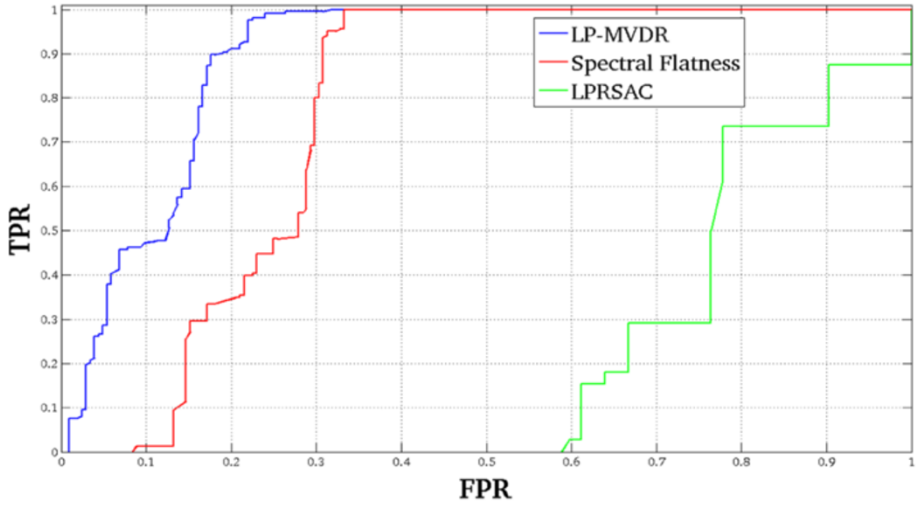

Figure 18 ROC curve for speech signal with SNR=35 dB from cellphone whispered speech data corpus containing whisper and neutral segments. Blue color indicates LP-MVDR technique, red indicates spectral flatness and green LPRSAC.

Table 3 Whisper diarisation error rates for different techniques at different SNRs on cellphone whispered speech data corpus

Condition	LP-MVDR	LPRSAC	Spectral flatness
Normal	0.31	-	0.31
SNR = 50	0.30	-	0.27
SNR = 45	0.28	–	0.27
SNR = 35	0.21	–	0.24

treating adapted models as a unified model encompassing both whisper and neutral speech elements may not be reasonable as indicated by the experimental results.

Conclusions and future scope

The work presented herein proposes a parametric spectral ratio method for whisper detection. The method performs reasonably better than conventional methods used in whisper detection both in terms of the ROC performance and the WDER. The usefulness of this method in automatic whispered speech recognition is also discussed in a MLLR adaptation framework. However availability of whispered speech data is an issue. It is also difficult to collect whispered databases on a large scale without proper supervision because of the human tendency to jump to neutral speech in a natural environment. Since whispered speech is characterized by noise and shift in formants to higher frequencies, newer model adaptation techniques can play a major role in this context. The work presented herein has also led to the development of a new database cellphone whispered speech data corpus which comprises whisper data collected through an IVRS system over standard mobile phones in an uncontrolled environment. Implementing whisper detection systems in environments like hospitals and prisons can address possible emergency situations. This can also prove to be useful for people with a collapsed larynx, or congenital diseases of the vocal chords. Recognition of the other natural modes of speech

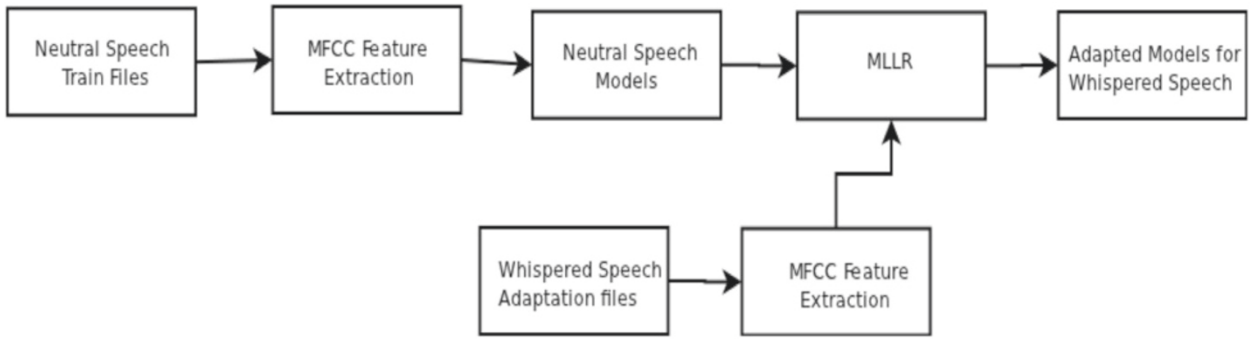

Figure 19 Block diagram showing the scheme followed for making of recognition models using hmm on cellphone whispered speech data corpus.

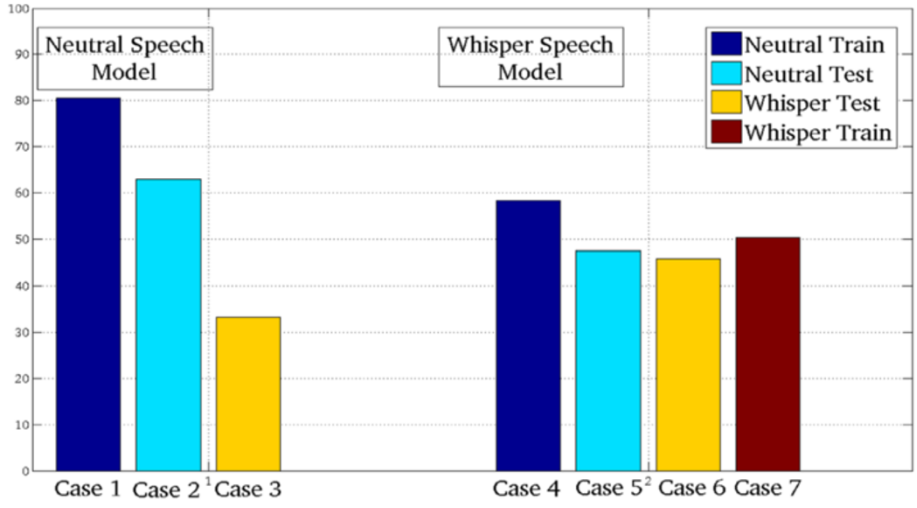

Figure 20 Recognition results for neutral and adapted digit models over different sets of voice sample files. The adapted model shows better whisper recognition results than the neutral model.

Table 4 Results of % recognition for different cases by the HMM models

Case	Number of files	% Recognition
Case 1	493	80.53
Case 2	200	63.00
Case 3	250	33.20
Case 4	493	58.42
Case 5	200	47.50
Case 6	250	50.40
Case 7	209	45.93

like shouted speech is currently being explored. The possibility of utilizing the spectra derived for whisper detection as features in the recognition process is being explored.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the BSNL-IIT Kanpur Telecom Center, MCIT, and QUALCOMM.

Received: 6 August 2011 Accepted: 22 June 2012
Published: 23 July 2012

References

1. R Ng, T Lee, C Leung, B Ma, H Li, Analysis and selection of prosodic features for language identification. in *2009 International Conference on Asian Language Processing,* (IEEE, 2009), pp. 123–128
2. C Zhang, J Hansen, Analysis and classification of speech mode: Whispered through shouted. in *Eighth Annual Conference of the International Speech Communication Association,* (2007)
3. TF Quatieri, *Discrete-Time Speech Signal Processing* (Pearson Education India, 2002)
4. V Tartter, What's in a whisper? J. Acoust. Soc. Am. **86**, 1678 (1989)
5. C Zhang, J Hansen, Advancements in whisper-island detection within normally phonated audio streams. in *Tenth Annual Conference of the International Speech Communication Association,* (2009)
6. SJ Wenndt, EJ Cupples, RM Floyd, A study on the classification of whispered and normally phonated speech. in *In ICSLP-2002,* (2002), pp. 649–652
7. A Mathur, R Hegde, Significance of the LP-MVDR spectral ratio method in whisper detection. in *National Conference on Communications (NCC),* (2011), pp. 1–5
8. C Zhang, J Hansen, Advancements in whisper-island detection using the linear predictive residual. in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010,* (2010), pp. 5170–5173
9. Z Chi, J Hansen, Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. IEEE Trans. Audio Speech Lang. Process. **19**(4), 883–894 (2011)
10. S Seyedin, S Ahadi, Robust MVDR-based feature extraction for speech recognition. in *7th International Conference on Information, Communications and Signal Processing, 2009, ICICS 2009,* (2009), pp. 1–5
11. M Murthi, B Rao, MVDR based all-pole models for spectral coding of speech. in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999,* vol. 2, (1999), pp. 669–672
12. U Yapanel, S Dharanipragada, Perceptual MVDR-based cepstral coefficients (PMCCs) for robust speech recognition. in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, ICASSP '03,* vol. 1, (2003), pp. I–644–I–647
13. VA Petrushin, LI Tsirulnik, V Makarova, Whispered Speech Prosody Modeling for TTS Synthesis. in *Speech Prosody 2010-Fifth International Conference,* (2010)
14. S Jovicic, Formant feature differences between whispered and voiced sustained vowels. Acta Acustica United with Acustica. **84**(4), 739–743 (1998)
15. M Carlin, B Smolenski, S Wenndt, Unsupervised detection of whispered speech in the presence of normal phonation. in *Ninth International Conference on Spoken Language Processing,* (2006)
16. A Gray Jr., J Markel, A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Trans. Acoust. Speech Signal Process. **22**(3), 207–217 (1974)
17. A Ali, J Van der Spiegel, P Mueller, An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants. in *ICASSP-1998,* vol. 2, (IEEE, 1998), pp. 961–964
18. R Yantorno, K Krishnamachari, J Lovekin, D Benincasa, S Wenndt, The spectral autocorrelation peak valley ratio (SAPVR)-a usable speech measure employed as a co-channel detection system. in *Proceedings of IEEE International Workshop on Intelligent Signal Processing (WISP),* (2001)
19. J Makhoul, Linear prediction: a tutorial review. Proc. IEEE. **63**(4), 561–580 (1975)
20. C Zhang, J Hansen, Effective segmentation based on vocal effort change point detection. in *ITRW,* (Aalborg, 2008)
21. PJ Sherman, KN Lou, On the family of ML spectral estimates for mixed spectrum identification. IEEE Trans. Signal Process. **39**(4), 644–655 (1991)
22. M Murthi, B Rao, Minimum variance distortionless response (MVDR) modeling of voiced speech. in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97,* vol. 3, (IEEE, 1997), pp. 1687–1690
23. M Murthi, B Rao, All-pole modeling of speech based on the minimum variance distortionless response spectrum. IEEE Trans. Speech Audio Process. **8**(3), 221–239 (2000)
24. M Wolfel, J McDonough, Minimum variance distortionless response spectral estimation. IEEE Signal Process. Mag. **22**(5), 117–126 (2005)
25. S Haykin, *Adaptive filter theory (ISE)*, (2003)
26. S Dharanipragada, B Rao, MVDR based feature extraction for robust speech recognition. in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001 (ICASSP'01),* vol. 1, (IEEE, 2001), pp. 309–312
27. J Burg, The relationship between maximum entropy spectra and maximum likelihood spectra. Geophysics. **37**, 375 (1972)
28. H Acquah, Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. J. Develop. Agri. Econ. **2**(1), 001–006 (2010)
29. A Tritschler, R Gopinath, Improved speaker segmentation and segments clustering using the Bayesian information criterion. in *Proc. Eurospeech,* vol. 2, (Citeseer, 1999), pp. 679–682
30. B Zhou, J Hansen, Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. in *Sixth International Conference on Spoken Language Processing (ICSLP),* (2000)
31. S Cheng, H Wang, H Fu, BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. IEEE Trans. Audio Speech Lang. Process. **18**(1), 141–157 (2010)
32. M Cettolo, M Vescovi, R Rizzi, Evaluation of BIC-based algorithms for audio segmentation. Comput. Speech Lang. **19**(2), 147–170 (2005)
33. MF Gales, The Generation And Use Of Regression Class Trees For MLLR Adaptation, Cambridge University Engineering Department, Tech. Rep., 1996
34. M Gales, P Woodland, Mean and variance adaptation within the MLLR framework. Comput. Speech Lang. **10**(4), 249–264 (1996)
35. M Tamura, T Masuko, K Tokuda, T Kobayashi, Speaker adaptation for HMM-based speech synthesis system using MLLR. in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*. (Citeseer, 1998), pp. 273–276
36. S Young, G Evermann, M Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, The HTK book (for HTK version 3.4), Cambridge University Engineering Department, vol. 2, no. 2, (2006), pp. 2–3
37. F Cummins, M Grimaldi, T Leonard, J Simko, Proceedings of SPECOM, *vol. 6*, (2006), pp. 431–435
38. F Zheng, G Zhang, Z Song, Comparison of different implementations of MFCC. J. Comput. Sci. Technol. **16**(6), 582–589 (2001)

doi:10.1186/1687-6180-2012-157
Cite this article as: Mathur *et al.*: Significance of parametric spectral ratio methods in detection and recognition of whispered speech. *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:157.