Feature Level Solution to Noise Robust Speech Recognition in the context of Tonal Languages

Utpal Bhattacharjee, Jyoti Mannala

Abstract: ***Performance of a speech recognition system is highly dependent on the operational environments. The mismatched ambient conditions have adverse impact on the performance of an Automatic Speech Recognition (ASR) system. The speech parameterization techniques for tonal speech recognition are different from those used for non-tonal speech recognition. It is due to the fact that tonal speech has two components – basic linguistic unit and tone. The basic linguistic unit with different tones convey different meanings. Therefore, the feature set used for tonal speech recognition must have the capability to representing both of them. Tone is determined by the fundamental frequency of the speech signal which is highly sensitive to noise. Since at the time of parameterization of the non-tonal speech recognition systems, these highly noise-sensitive tone related information are discarded, the traditional noise elimination methods used for non-tonal speech recognition fail to deliver robust performance in tonal speech recognition. In the present study, we have analyze the performance of different commonly used feature sets for noisy tonal speech recognition. Hidden Markov Model (HMM) based speech recognizer has been used for performance evaluation. Noise elimination techniques sub-band spectral subtraction and Wiener filter have been used for noise reduction and their relative performance have been evaluated.***

Keywords :HMM, Noise elimination, Sub-band spectral subtraction, Tonal speech recognition, Wiener Filter

I. INTRODUCTION

Feature extraction is the front-end of any speech recognition system. The feature extraction for a speech recognition system is the process of reliable, compact and robust parameterization of the speech signal. The efficiency of the entire speech recognition system is highly dependent on proper parameterization of the speech signal. The feature vector extracted from the speech signal must have the capability to discriminating among different phonemes and must be robust to the environment and intra-speaker variability. The significance of cepstral features for speech recognition have been established by many researchers [1][2][3]. However, there are practical limitation in the use of cepstral features due to its sensitivity towards the background and channel noises [4].Mel frequency cepstral coefficients (MFCC) and linear predictor cepstral coefficients (LPCC) are two extensively used feature vector in speech science. MFCC feature is based on magnitude spectrum. A perceptually motivated frequency wrapping filter-bank is applied to the magnitude spectrum. The filters are evenly spaced on a perceptually motivated frequency wrapping scale call Mel-scale, first suggested by Stevens and Volkman [5]. The log-energy of each filter output is computed and accumulated. Finally, Discrete Cosine Transformation (DCT) is applied to produce the Mel frequency cepstral coefficients [6]. In the present study, a filter bank of 24 triangular filters spread across the whole frequency range from 0 to Nyquist frequency has been used. The first 12-cepstral coefficients and log energy have been considered as the MFCC feature vector. Linear predictor cepstral coefficient (LPCC) is a feature vector based on Linear predictor coefficient (LPC). The LPC are obtained using a p^{th} -order All-pole approximation in the windowed waveform [7]. The autocorrelation method has been used to evaluate the linear predictor coefficients. The LPCC have been computed directly from LPC as [8]:

$$c_n = \begin{cases} a_n + \dfrac{1}{n}\displaystyle\sum_{m=1}^{n-1} mc_m a_{n-m}, & 1 \le n \le p \\ \dfrac{1}{n}\displaystyle\sum_{m=n-p}^{n-1} mc_m a_{n-m}, & n \ge p \end{cases} \quad \ldots(1)$$

where$p$ is the order of the predictor coefficients and n is the number of cepstral coefficients. In the present study, 10th order LP analysis has been performed and 13 LPCC coefficients have been computed. To capture the dynamic property of the speech signal, along with baseline MFCC and LPCC features their first and second order derivatives are also added. Thus we get a 39-dimensinal MFCC feature set and a 39-dimensional LPCC feature set

Prosody plays an important role in understanding the meaning of a conversation in human to human communications. Prosodic features of speech characterize the paralinguistic information of a conversation like speaker habits, discourse structure, speaker intension, emotion etc. In general, prosody means the organization of a sound. Normally, it is represented by fundamental frequency (F_0), energy and normalized duration of syllable. The prosodic features are very important to identify the tone associated with a syllable. In the present study, in order to use only frame-based features, fundamental frequency and energy have been considered for the representation of prosodic information. Fundamental frequency and frame energy are static features, calculated frame
by frame.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
UtpalBhattacharjee*, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, Pin - 791 112 Email: utpal.bhattacharjee@rgu.ac.in
Jyoti Mannala, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India, Pin - 791 112 Email: mannalajoy@gmail.com

In order to include temporal information, their first (Δ)- and second (ΔΔ)-order derivatives have been calculated and added to the feature set. Thus, we get a 6-dimensional prosodic feature vector for each frame. Left-to-Right Hidden Markov Model (LRHMM) has been used as baseline speech recognition system to recognize the tonal vowels of Apatani language of Arunachal Pradesh of North East India. The main reason for using LRHMM is that it can model the time varying property of the speech signal. A number of HMM states is determined empirically. In the present model, 6 (six) states have been used. Each state is represented by a single Gaussian distribution function given by [9]

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad \ldots (2)$$

Where x is the observation vector, μ is the Gaussian mean vector and σ^2 is the variance. To initialize the model, speech signal from known vowels have been divided into 6 equal parts and each part from left to right has been used to initialize a state. The forward-backward algorithm has been used for training the HMM model. Clean speech signals have been used for the training purpose.

II. NOISE ELIMINATION METHODS

The most commonly used de-noising techniques are based on either spectral subtraction or Weiner's filter. These techniques are based on the assumption that the speech signal $s(n)$and the additive noise $d(n)$are uncorrelated with each other. Therefore, the equation for noisy speech signal $x(n)$ can be represented as [10]

$$x(n) = s(n) + d(n) \quad \ldots (3)$$

The original signal can be estimated from the noisy speech signal by using Wiener filter as:

$$\hat{S}(\omega) = H(\omega).X(\omega) \quad \ldots (4)$$

Where $H(\omega)$, $X(\omega)$, $\hat{S}(\omega)$ are the Wiener filter response function, noisy signal and the estimated clean speech signal in frequency domain respectively. Wiener filter is an optimal filter that minimize the mean square error. The mean square error is represented by the function

$$\begin{aligned} E(\omega) &= S(\omega) - \hat{S}(\omega) \\ &= S(\omega) - H(\omega).X(\omega) \end{aligned} \quad \ldots (5)$$

The $H(\omega)$ value is determined by minimizing the expectation of mean square error, which is obtained by taking first order derivative of the error function with respect to response function of the Weiner filter $H(\omega)$ and equating it to 0. The expectation of mean square error is given by

$$E[|E(\omega)|^2] = E[|S(\omega) - H(\omega).X(\omega)|^2] \quad \ldots (6)$$

where $E[.]$ stands for expectation operation. Taking the derivatives of eq(6) and equating it to 0, we get

$$\begin{aligned} \frac{\delta E[|E(\omega)|^2]}{\delta H(\omega)} &= 2H(\omega)E[|X(\omega)|^2] - 2E[|X(\omega)S(\omega)^*|] \\ &= 2H(\omega)P_X(\omega) - 2P_{XS}(\omega) = 0 \end{aligned} \quad \ldots (7)$$

where$P_X(\omega)$ and $P_{XS}(\omega)$ are power spectra of noisy speech and cross power spectra between noisy speech signal and clean speech respectively.In case of no correlation between the speech signal $s(n)$ and the additive noise $d(n)$, we get

$$\begin{aligned} P_X(\omega) &= E[|X(\omega)|^2] = E[|S(\omega) + D(\omega)|^2] \\ &= E[|S(\omega)|^2] + E[|D(\omega)|^2] + E[|S(\omega)D(\omega)|] \\ &= P_S(\omega) + P_D(\omega) \end{aligned} \quad \ldots (8)$$

Similarly

$$\begin{aligned} P_{XS}(\omega) &= E[|X(\omega)S(\omega)^*|] \\ &= E\left[\left|\left(S(\omega) + D(\omega)\right)S(\omega)^*\right|\right] \\ &= E[|S(\omega)|^2] = P_S(\omega) \end{aligned} \quad \ldots (9)$$

Therefore, the Wiener filter can be represented by:

$$H(\omega) = \frac{P_S(\omega)}{P_S(\omega) + P_D(\omega)} \quad \ldots (10)$$

The signal-to-noise ratio is defined by

$$SNR = \frac{P_S(\omega)}{P_D(\omega)} \quad \ldots (11)$$

Therefore, the impulse response of the Wiener filter can be represented in term of SNR as:

$$H(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1} \quad \ldots (12)$$

In the present work, we have implemented the adaptive Wiener filter based on the model proposed by El-Fattah et al[11] for speech enhancement. The mean m_x and standard deviation ${\sigma_x}^2$of the speech signal have been estimated. It is assumed that the additive noise is of zero mean and variance${\sigma_d}^2$. The variance ${\sigma_d}^2$ has been estimated exploiting the silent period of the speech signal. Thus the power spectrum of noise has been estimated as

$$P_D(\omega) = {\sigma_d}^2 \quad \ldots (13)$$

Considering a small segment of the speech signal, in which speech $x(n)$ is assumed to be stationary, the signal can be modelled as:

$$x(n) = m_x + {\sigma_x}^2 w(n) \quad \ldots (14)$$

wherem_x and ${\sigma_x}^2$ mean and standard deviation of the speech signal for a small segment of the speech signal and $w(n)$ is unit variance noise. Therefore, for a small segment of the speech signal, the Wiener filter transfer function can be represented by:

$$H(\omega) = \frac{\sigma_s{}^2}{\sigma_s{}^2 + \sigma_d{}^2} \qquad \dots (15)$$

Since $H(\omega)$ is constant over this small segment of speech, the impulse response of the Wiener filter can be obtained by

$$h(n) = \frac{\sigma_s{}^2}{\sigma_s{}^2 + \sigma_d{}^2}\delta(n) \qquad \dots (16)$$

The enhanced speech signal $\hat{s}(n)$ for the local segment can be expressed as:

$$\begin{aligned}\hat{s}(n) &= m_x + (x(n) - m_x) * \frac{\sigma_s{}^2}{\sigma_s{}^2 + \sigma_d{}^2}\delta(n)\\ &= m_x + \frac{\sigma_s{}^2}{\sigma_s{}^2 + \sigma_d{}^2}(x(n) - m_x)\end{aligned} \qquad \dots (17)$$

If m_x and $\sigma_s{}^2$ are updated for each segment, we can say

$$\hat{s}(n) = m_x(n) + \frac{\sigma_s{}^2}{\sigma_s{}^2(n) + \sigma_d{}^2}(x(n) - m_x(n)) \qquad \dots (18)$$

when$\sigma_s{}^2$is much larger than $\sigma_d{}^2$, there will be no attenuation and the estimated speech signal $\hat{s}(n)$ will be basically due to $x(n)$. However, if $\sigma_s{}^2$ is smaller than$\sigma_d{}^2$, there will be attenuation and the filtering will be done.The value of $m_x(n)$ can be estimated from the $x(n)$ as:

$$\widehat{m}(n) = \frac{1}{2M+1}\sum_{k=n-M}^{n+M} x(k) \qquad \dots (19)$$

where$(2M + 1)$ is the number of sample in the short segment used for estimation. Since $\sigma_x{}^2 = \sigma_s{}^2 + \sigma_d{}^2$, $\hat{\sigma}_s{}^2(n)$ may be estimated from $x(n)$ as:

$$\hat{\sigma}_s{}^2(n) = \begin{cases}\hat{\sigma}_x{}^2(n) - \hat{\sigma}_d{}^2 & , if\, \hat{\sigma}_x{}^2(n) > \hat{\sigma}_d{}^2\\ 0, otherwise\end{cases} \qquad \dots (20)$$

Where

$$\hat{\sigma}_x{}^2(n) = \frac{1}{2M+1}\sum_{k=n-M}^{n+M}\left(x(k) - \widehat{m}(n)\right)^2 \qquad \dots (21)$$

Another method for de-noising uncorrelated additive noise is spectral subtraction. The power spectra of the corrupted speech signal can be approximated from eq. (3) as

$$|X(k)|^2 = |S(k)|^2 + |D(k)|^2 \qquad \dots (22)$$

where$|S(k)|^2$ and $|X(k)|^2$ are the magnitude spectra of clean and the noise respectively. Since the noise spectra cannot be obtained directly, an estimate $\widehat{D}(k)$is obtained from the silent period [12]. The estimation of clean speech spectrum is obtained by

$$\left|\hat{S}(k)\right|^2 = |X(k)|^2 - \alpha\left|\widehat{D}(k)\right|^2 \qquad \dots (23)$$

whereα is the over subtraction factor, which is a function of SNR. This model is based on the assumption that the noise affects the speech signal uniformly. However in case of real world operational conditions, this assumption is not true. It has been observed that impact of noise is different for different frequency range. Kamath and Loizou [13] proposed a multiband model for spectral subtraction. The entire frequency range of the speech signal is divided into N non-overlapping sub-bands and band-specific subtraction factor is computed for each frequency band. The estimation for clean speech of the i^{th} band is obtained by

$$\left|\hat{S}(k)\right|^2 = |X_i(k)|^2 - \alpha_i\delta_i\left|\widehat{D}(k)\right|^2, b_i \le k \le e_i \qquad \dots (24)$$

whereb_i and b_i are beginning and ending frequency bins of thei^{th} frequency band and δ_i is the tweaking factor for the i^{th}band. The band specific SNR is computed using the magnitude of the noisy spectra and estimated noise spectra as follows:

$$SNR_i = 10\,\log_{10}\left(\frac{\sum_{b_i}^{e_i}|X_i(k)|^2}{\sum_{b_i}^{e_i}\left|\widehat{D}_i(k)\right|^2}\right) \qquad \dots (25)$$

Using the SNR value α_i is computed as:

$$\alpha_i = \begin{cases} 5 & SNR_i < -5\\ 4 - \frac{3}{20}SNR_i & -5 \le SNR_i \le 20\\ 1 & SNR_i > 20\end{cases} \qquad \dots (26)$$

The negative value of the enhanced spectra is floored to the noisy spectra.

III. SPEECH DATABASE

A speech database of Apatani tonal words has been prepared to carry out the experiments. The Apatani language of Arunachal Pradesh of North Eastern India is a tone language. A language is regarded as ‘Tone Language’ if the change in the tone of the word results in changing the meaning of the word [14].Apatani has two lexical tones raising (´) and falling (`) [15]. In addition to these two tones, Apatani has words without any associated tone, which are referred to as normal tone. Except the vowel [ə] all the other vowels have 3 tonal instances namely raising, falling and level. In case of vowel [ə] only level tone has been observed. In the evaluation of the speech recognition system for tonal speech recognition task, the vowel [ə] has not been taken into consideration. The database for the present research consist of 24 isolated tonal words spoken by 20 different speakers (13 males and 7 females).The words chosen for recording are:

Table-1: Tonal words considered for recording

Sl no.	Apatani Tonal Words	Meaning in English
1	tɑ́	Bite
2	tɑ	Listen
3	tɑ̀	Drink
4	khɛ̀	Cry
5	khɛ	To get angry

6	khέ	Remove
7	cɪ	Cut with scissor
8	cɪ̀	Bring together two things
9	jɪ̀	Be black
10	jɪ́	Roll
11	jɪ	Bind
12	ɑ̀lɔ̀	Salt
13	ɑ̀lɔ	Dry
14	kɔrɔ	Day before yesterday
15	kɔ́rɔ	Fence
16	ɑpʊ́	Blossom
17	ɑ́pʊ	Wrap Up
18	kʊ	Beg
19	kʊ̀	Spray
20	kʊ́	Wave like movement
21	ɑnʊ́	Young Brother
22	ɑnʊ	Uncle
23	mədɔ́	Rain
24	mədɔ	Doing

For any tone language, the basic building blocks are tonal syllables. A tonal syllable consist of two components – a syllabic sound unit and an associated lexical tone. If the tone is ignored, it is called base syllable. Each syllable consist of vowel and consonant sounds. Tone is realized in voiced segment, therefore, tonal base units (TBU) in most of the time are voiced vowels [16]. The tone associated with the vowels are sufficient to express the tone associated with the syllable. Therefore, in the present study we will evaluate robustness of a tonal speech recognition system in terms of its capability to recognize tonal vowel at different noise conditions. The words are recorded in a controlled acoustical environment at 16 KHz sampling frequency and 16 bit mono format. A headphone microphone has been used for recoding the database. Each speaker uttered the same words 5 times. From the recorded isolated words, a vowel database has been created by segmenting the vowels from the isolated words. The segmentation has been done by using PRAAT software which is followed by subjective verification. Thus we get at least 100 instances for each tonal vowel. The database has been divided into two parts – training set and testing set. The training set consist of 50 instances of each tonal vowel and the testing set consist of remaining 50 instances of each tonal vowel.

From the clean database noisy versions of the database has been created by adding noise from the AURORA database [17]. The noises added to the database are babble, car, exhibition, restaurant, street, subway and train noises. The noises are added at -15dB, -10dB, -5dB, 0dB, 5dB, 10dB and 15dB signal-to-noise ratio (SNR).

IV. RESULTS AND DISCUSSION

The speech has been analyzed using a Hamming windows of length 25 ms, frame rate 100 Hz and pre-emphasis factor of 0.97. MFCC, LPCC and prosodic features have been extracted from each frame. Now from the extracted features two tonal feature sets have been created by appending the prosodic features with MFCC and LPCC features separately. We call them MFCC tonal feature and LPCC tonal feature respectively. To study the suitability of the feature sets for tonal speech recognition their probability density function (PDF) characteristics have been analyzed. If the same vowel with different tone have different PDF characteristics for a particular feature set, then the feature set will be efficient in recognizing the tonal instances of the vowels. PDF characteristics of the MFCC and LPCC tonal feature sets are given in Fig-1 and Fig-2 respectively. From the Figures it has been observed that both MFCC and LPCC feature sets the peak of the distribution are at different positions. For the vowel [ɔ] the MFCC tonal feature has more tonal phoneme discrimination capability while for vowel [ɑ] the LPCC tonal feature exhibits more tone discrimination capability. In case of vowels [ɛ] and [ʊ], both MFCC and LPCC tonal features display tone discrimination capability. This observation justify the fact that tone discrimination capability of a feature set depends on the underlying vowels.

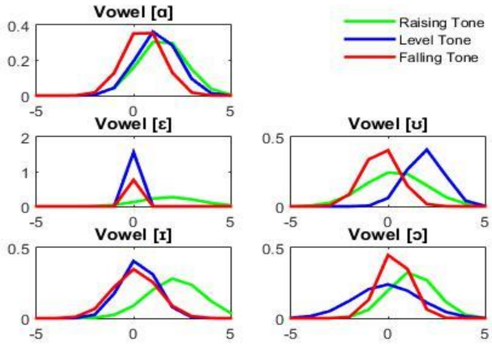

Fig. 1 PDF characteristics of tonal vowels for MFCC tonal feature set

To evaluate the efficiency of the feature set in recognizing the tonal vowels, a Hidden Markov Model based recognizer has been trained using the clean training set. The testing has been done using the testing set and the confusion matrices for recognition of the tonal vowels have been prepared. The confusion matrices for the MFCC and LPCC tonal feature sets based HMM recognizer for recognizing the tonal vowels have been given in Table – 2 and Table -3 respectively.

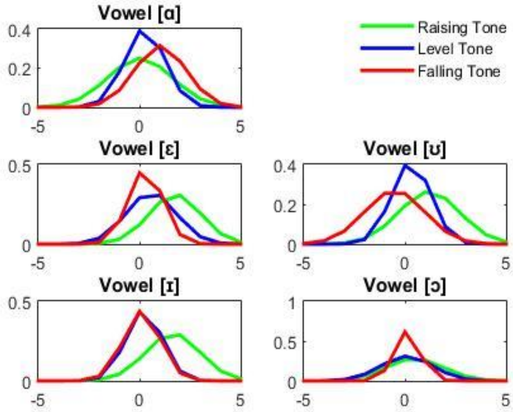

Fig. 2 PDF characteristics of tonal vowels for LPCC tonal feature set

Table – 2: Confusion matrix for tonal phoneme recognition with tonal MFCC and HMM based recognizer (50 test for each tonal vowel)

	[ɑ̄:]	[ɑ́:]	[ɑ̀:]
[ɑ̄:]	41	6	3
[ɑ́:]	4	43	3
[ɑ̀:]	1	2	47
	[ɪ̄]	[ɪ́]	[ɪ̀]
[ɪ̄]	43	7	0
[ɪ́]	6	44	0
[ɪ̀]	0	0	50
	[ɔ̄]	[ɔ́]	[ɔ̀]
[ɔ̄]	49	1	0
[ɔ́]	2	48	0
[ɔ̀]	1	0	49
	[ɛ̄]	[ɛ́]	[ɛ̀]
[ɛ̄]	48	1	1
[ɛ́]	0	48	2
[ɛ̀]	1	1	48
	[ʊ̄]	[ʊ́]	[ʊ̀]
[ʊ̄]	47	1	2
[ʊ́]	1	48	1
[ʊ̀]	0	1	49

Table 3: Confusion matrix for tonal phoneme recognition with tonal LPCC and HMM based recognizer (50 test for each tonal vowel)

	[ɑ̄:]	[ɑ́:]	[ɑ̀:]
[ɑ̄:]	48	1	1
[ɑ́:]	0	49	1
[ɑ̀:]	0	0	50
	[ɪ̄]	[ɪ́]	[ɪ̀]
[ɪ̄]	39	8	3
[ɪ́]	5	40	5
[ɪ̀]	2	2	46
	[ɔ̄]	[ɔ́]	[ɔ̀]
[ɔ̄]	41	6	3
[ɔ́]	9	37	4
[ɔ̀]	2	10	38
	[ɛ̄]	[ɛ́]	[ɛ̀]
[ɛ̄]	47	0	3
[ɛ́]	1	49	0
[ɛ̀]	0	2	48
	[ʊ̄]	[ʊ́]	[ʊ̀]
[ʊ̄]	49	0	1
[ʊ́]	2	48	0
[ʊ̀]	1	1	48

From the above confusion matrices it has been observed that the tonal phoneme recognition accuracy depends on the feature set and the underlying vowel. The recognition accuracy of the HMM based recognizer in tonal phoneme discrimination using different feature sets have been summarized in table-4.

Table -4: Recognition accuracy of the HMM based recognizer for tonal phoneme recognition for different feature sets

Tonal Vowel	MFCC tonal Feature set (in %)	LPCC tonal feature set (in %)
[ɑ̄:]	82	96
[ɑ́:]	86	98
[ɑ̀:]	94	100
[ɪ̄]	86	78
[ɪ́]	88	80
[ɪ̀]	100	92
[ɔ̄]	98	82
[ɔ́]	96	74
[ɔ̀]	98	76
[ɛ̄]	96	94
[ɛ́]	96	98
[ɛ̀]	96	96
[ʊ̄]	94	98
[ʊ́]	96	96
[ʊ̀]	98	96
Average	93.6	90.27

In the next set of experiments, we have considered the noisy versions of the database and their performances have been evaluated using the same HMM model which is trained with clean speech. The recognition accuracy under different noise types and noise levels is given in table-5 and table-6.

Table – 5: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing noisy tonal vowels

Noise Type	-15 dB	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Babble	23.4	25.4	26.7	33.8	37.4	54.2	67.3
Car	24.0	26.5	27.6	32.8	38.6	49.8	69.6
Exhibition	22.6	28.6	29.1	33.1	40.7	52.8	73.3
Restaurant	22.8	25.3	24.6	31.7	34.4	58.0	62.0
Street	28.2	25.9	25.2	35.4	35.3	48.9	63.5
Subway	24.6	29.5	28.8	38.6	40.3	52.3	72.6
Train	25.8	28.2	26.1	37.2	37.8	52.6	68.0

Table – 6: The recognition accuracy of HMM and LPCC tonal feature based speech recognizer for recognizing noisy tonal vowels

Noise Type	-15 dB	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Babble	21.1	22.4	23.8	29.9	33.2	48.0	59.7
Car	23.3	20.7	24.2	27.1	32.9	41.8	58.8
Exhibition	22.4	25.7	27.5	30.5	38.0	49.0	68.2
Restaurant	19.8	21.3	21.0	26.9	29.3	49.3	52.7
Street	24.8	25.4	23.4	33.8	33.3	46.4	60.0
Subway	21.2	23.6	23.9	31.5	33.1	42.8	59.6
Train	23.0	22.3	21.9	30.3	31.3	43.2	56.1

From the above results, it has been observed that the recognition accuracy of the HMM based recognizer degrades considerably when noise presents in the speech signal. The performance deterioration is different for different noise types. Further, it has been observed that MFCC tonal feature outperforms LPCC based tonal feature under all operational conditions. Therefore, MFCC tonal feature may be considered as better parameterization technique for tonal speech recognition under all operational conditions. Therefore, the performance of the de-noising techniques have been evaluated with MFCC tonal feature only. To de-noise the corrupted speech signal, we apply Wiener Filter and sub-band spectral subtraction methods separately and the performance haves been evaluated. The result of the experiments are given in table-7 and table-8.

Table – 7: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing tonal vowels at different noise conditions with Wiener Filter de-noising technique

Noise Type	-15 dB	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Babble	33.7	35.8	38.0	47.9	53.1	66.8	83.0
Car	41.9	37.2	43.5	48.9	59.1	61.9	87.7
Exhibition	35.8	41.2	44.0	48.9	60.8	71.3	89.6
Restaurant	31.7	34.0	33.7	43.0	46.9	65.6	70.3
Street	39.7	40.6	37.5	54.1	53.2	69.0	89.0
Subway	37.2	41.5	42.1	55.4	58.3	60.5	84.3
Train	41.3	40.1	39.5	54.6	56.3	62.6	81.5

Table – 8: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing tonal vowels at different noise conditions with sub-band spectral subtraction de-noising technique

Noise Type	-15 dB	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Babble	29.7	30.8	33.2	41.5	46.2	76.9	95.5
Car	39.8	28.4	37.3	39.6	49.3	75.2	95.9
Exhibition	34.7	36.3	40.7	44.2	55.6	78.4	94.2
Restaurant	27.1	28.0	28.2	35.7	39.1	78.8	84.4
Street	34.2	39.0	34.2	50.6	49.2	74.2	96.1
Subway	31.4	32.6	34.2	44.2	47.0	75.4	97.9
Train	36.0	31.0	32.5	43.6	45.7	77.8	94.9

From the above experiments it has been observed that the Wiener filter gives better performance in high noise condition whereas the sub-band spectral subtraction gives better performance at low noise condition. At 10dB and 15dB noise level, the sub-band spectral subtraction method outperforms the Wiener filter in noise compensation.

In the next experiment, we have combined the sub-band spectral subtraction and Wiener filter method. The speech spectra first goes through a sub-band spectral subtraction method and then Wiener filter is applied. The result of the experiment is summarized in table-9.

Table-9: The recognition accuracy of HMM and MFCC tonal feature based speech recognizer for recognizing tonal vowels at different noise conditions with sub-band spectral subtraction and Wiener filter de-noising techniques

Noise Type	-15 dB	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Babble	38.0	40.0	42.7	53.6	59.6	86.2	89.3
Car	49.0	39.4	48.4	53.1	65.1	82.3	96.8
Exhibition	42.3	46.5	50.8	55.8	69.9	89.8	94.0
Restaurant	35.3	37.2	37.1	47.2	51.6	86.7	77.3
Street	44.4	47.8	43.0	62.8	61.4	85.9	92.6
Subway	41.2	44.5	45.8	59.8	63.2	81.5	94.6
Train	46.4	42.7	43.2	58.9	61.2	84.2	91.2

From the above results it has been observed that under certain noise conditions one de-noising technique gives better performance over the other technique. However, when both the methods are combined together, it gives a consistently optimal performance under all operational conditions.

V. CONCLUSION

In this paper, the robustness issue of MFCC and LPCC features combined with prosodic features has been evaluated for tonal speech recognition. In case of tonal speech recognition only the spectral features like MFCC and LPCC are not sufficient as they does not conation tone related information. Therefore, prosodic features must have to be combined with them. Prosodic feature, which is determined by fundamental frequency and energy is highly sensitive to noise. Therefore, at noisy environmental conditions the performance of the speech recognition system degrades considerably. In the present study it has been observed that under controlled environmental conditions, both MFCC + Prosodic features and LPCC + prosodic features perform well in recognizing the tonal speech. However, with increasing level of noise, the performance degrades considerably. The degradation is more in case of LPCC + prosodic features compared to MFCC + prosodic features. Considering all operational conditions it has been observed that MFCC + prosodic feature is a better option for recognizing tonal speech. Two most commonly used de-noising techniques sub-band spectral subtraction and Wiener filter have been used for noise elimination in the present work.

It has been observed that Wiener filter perform significantly well in high noise conditions whereas sub-band spectral subtraction gives better performance in low noise condition. Combining both the methods, we have observed that the performance has consistently improves in all noise conditions. However, for some noise conditions, this performance is lower than the performance of individual techniques. Considering an optimal operational scenario, we have suggested that sub-spectral subtraction combined with Wiener filter is a viable noise reduction technique for tonal speech recognition.

ACKNOWLEDGMENT

This work is supported by UGC major project grant MRP-MAJOR-COM-2013-40580, Ministry of Human Resource Development, Government of India.

REFERENCES

1. M. Baloul, E. Cherrier, and C. Rosenberger. "Challenge-based speaker recognition for mobile authentication." Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the. IEEE, 2012.
2. N. Desai, K. Dhameliya, and V. Desai. "Feature extraction and classification techniques for speech recognition: A review." International Journal of Emerging Technology and Advanced Engineering 13.12: 367-371, 2013.
3. W. Han, et al. "An efficient MFCC extraction method in speech recognition."2006 IEEE international symposium on circuits and systems. IEEE, 2006.
4. X. Zhao, and D. Wang. "Analyzing noise robustness of MFCC and GFCC features in speaker identification." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
5. S. S. Stevens and J. Volkman, The Relation of Pitch to Frequency, A Revised Scale. In: American Journal of Psychology, 53, 1940.
6. B.J. Shannon and K. K. Paliwal. "A comparative study of filter bank spacing for speech recognition." Microelectronic engineering research conference. Vol. 41. 2003.
7. L. Rabiner and M. Sambur. "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem." IEEE Transactions on Acoustics, Speech, and Signal Processing 25, no. 4: 338-343, 1977.
8. B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification." the Journal of the Acoustical Society of America 55, no. 6 : 1304-1312. 1974.
9. L. Rabiner, et al. "HMM clustering for connected word recognition." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.
10. J. Meyer and K.U. Simmer. "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction." In 1997 IEEE international conference on acoustics, speech, and signal processing, vol. 2, pp. 1167-1170. IEEE, 1997.
11. M.A. Abd El-Fattah, et al. "Speech enhancement using an adaptive wiener filtering approach." Progress in Electromagnetics Research 4: 167-184, 2008.
12. M. Berouti, R. Schwartz and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” Proc.IEEE Int. Conf. Acoust., Speech, Signal Process., pp.208-211, Apr. 1979.
13. S. Kamath and P.Loizou. "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise." In ICASSP, vol. 4, pp. 44164-44164. 2002.
14. M. Yip, The Tonal Phonology of Chinese, New York: Garland Publishing, 1991.
15. J. T. Sun, “Tani languages”, In The Sino-Tibetan Languages, edited by G. Thurgood and R. LaPolla, pp. 456-466, London and New York: Routledge, 2003.
16. P. Sarmah, “Tone Systems of Dimasa and Rabha: A Phonetic and Phonological Study”, Doctoral dissertation, University of Florida, 2009.
17. http://ecs.utdallas.edu/loizou/speech/noizeus/ accessed on 23rd October, 2019.

AUTHORS PROFILE

Utpal Bhattacharjee received his Master Degree from Dibrugarh University, India and Ph.D. from Gauhati University, India in the year 1999 and 2008 respectively. Currently he is working as a Professor in the department of Computer Science and Engineering of Rajiv Gandhi University, Arunachal Pradesh, India. His research interest is in the field of Speech and Natural language Processing and Machine Learning

Jyoti Mannala received her Master Degree from Rajiv Gandhi University,Arunachal Pradesh, India in the year 2012. Presently she is working as a research scholar in the department of Computer Science and Engineering of the University. Her research area is natural language processing.