Comparative Study of Filter Banks to Improve the Performance of Voice Disorder Assessment Systems using LTAS Features

Purva Barche* Krishna Gurugubelli* and Anil Kumar Vuppala *
* Speech Processing Laboratory, LTRC, KCIS
International Institute of Information Technology, Hyderabad, India.
E-mail:{purva.sharma, krishna.gurugubelli}@research.iiit.ac.in, and anil.vuppala@iiit.ac.in

***Abstract*—Objective assessment of voice disorders is widely explored as an early diagnosis tool for the classification of voice disorders. Voice disorders affect the pitch, loudness and voice quality, which are perceived at the suprasegmental-level in the speech signal. For the detection and assessment of voice disorders, this study explores the effectiveness of Long Term Average Spectral (LTAS) features using four state-of-the-art filter banks designed with critical-band, constant-Q, gammatone, and single-frequency filtering approaches. Moreover, the performance of the systems is compared with state-of-the-art statistical-average and openSMILE features. Voice disorder detection experiment was carried out on SVD and HUPA database, while only SVD database is used for assessment task. Assessment task is performed in clinical way, in which four binary classifiers were trained in our study. Voice disorder detection and assessment tasks were carried out using the support vector machine classifier. From the results, it was observed that constant-Q filter bank based LTAS features performed better among all LTAS features with classification accuracy of 78% and 81.4% for voice disorder detection task on SVD and HUPA database, respectively. Further, the combination of LTAS features with OpenSMILE features improved (89.6% and 86.6% for SVD and HUPA database, respectively) the performance.**

I. Introduction

Speech is a perceptual phenomenon, and its production is a complex process that involves the coordination of various muscles of the voice box, respiratory system, and resonating system all controlled by the brain. Voice disorders are mainly due to abnormalities in the larynx and its associated structure. It is characterized by abnormal voice production, change in voice quality, pitch, and loudness inappropriate to age and gender [1]. Instrument assessment, auditory-perceptual assessment and objective assessment are very popular methods for diagnosing the voice disorders [2]. In the instrument assessment method endoscope like laryngoscope, stroboscopes are used, but these methods are expensive and painful [3]. Auditory-perceptual method done by Speech-Language Pathologists (SLPs) is considered as a golden standard for the detection of voice disorder [4]. The decision taken in the subjective intelligibility test vary with experience of SLPs, type of scale used, and also depend on the examiner's experience [5]. Due to these reasons, objective or automatic assessment of voice disorders is explored a lot in literature. Objective assessment derives the acoustic features from the speech signal, hence reliable, economical and can be used by SLPs as pre-diagnosis measure [6].

Acoustic features, used for discrimination of healthy voice from the disordered voice include perturbation measures like jitter, shimmer and noise measures like harmonic to noise ratio, glottal to noise excitation [7], [8]. Further cepstrum features like Mel Frequency Cepstral Coefficients (MFCC), and Linear Prediction Cepstral Coefficients (LPCC) [9], excitation source features like glottal parameters [10] and intonation features [11] were also investigated in the research. The voice quality like breathiness, roughness, loudness, and intonation from the speech signal perceived is in the long term [12]. Hence these features can be captured by Long Term Average Spectrum (LTAS). LTAS captures the static characteristic of the speaker's voice instead of the short time variation present in the speech. Many researchers used LTAS in clinical application, as well as in quantification of voice quality. Some studies claim that LTAS can be used for voice classification [13]. Some researchers used LTAS as a good acoustic measure to differentiate the male and female speakers [14]. In [15], LTAS also used to study voice quality changes before and after surgery. Other works related to LTAS were finding difference related to age [16], professional singers, different styles of singing [17], speaking and singing [18], and quantifying the quality of voice [19].

For the extraction of the LTAS features speech signal should be decomposed into the multiple frequency components for that filter banks should be used. In the literature, LTAS features were extracted using the critical band filter bank [14], [16], and single frequency filter bank [20]. In our study, along with these two filter banks, we used auditory filter bank like constant-Q and gammatone filter banks for the extraction of LTAS features. The highlights of this study include:

- Along with detection (which mostly done in literature), this study also performed an assessment task for the discrimination of voice disorders on SVD database [21].
- In light of the importance of LTAS features, various state-of-the-art filter banks were explored.
- Further, this study compared the performance of LTAS features with state-of-the-art open-SMILE features and statistical features obtained from the vocal-tract system and excitation source components of speech.

- Moreover, N-way analysis of variances (ANOVA) was performed to investigate the relationship between the LTAS features and perceptual scale available for HUPA database.

This paper is organized as follows: In Section II, filter banks used for LTAS feature extractions is discussed. The experimental setup which describes feature extraction, database and classifier is discussed in the Section III. Results obtained are presented in the Section IV. Conclusion and summary of this study are described in Section V.

II. Filter banks For LTAS Feature extraction

In this paper, LTAS based features are used to capture information related to voice disorders. This section describes four state-of-the-art filter banks used in this study for voice disorder detection and assessment, along with the extraction of the LTAS features.

A. State-of-the-art Filter banks

The filter banks considered in this study, namely critical band, gammatone, Constant Q, and single frequency filter banks, are described as follows.

1) Critical band filter bank: Critical band filter bank (CBFB), also referred to as octave band filter bank, is used to mimic human perception. Octave band filters are set of bandpass filters in which highest frequency is twice of the lowest frequency [14]. Octave band is mainly used in music, in which one octave is difference between same notes with double it's frequency.

2) Gammatone filter bank: The gammatone filters are the most widely used auditory filters to model the human auditory system. In the term gammatone, gamma is referred to function mostly used in probability, and tone refers to the cosine term. Gammatone filter bank (GFB) models the cochlea by overlapping bandpass filter with impulse response given by the product of a rising polynomial, a decaying exponential function, and a cosine wave [23]. The impulse response of a gammatone filter $g(t)$ is given by,

$$g(t) = at^{(N-1)}e^{-2\pi bt}cos(2\pi f_c t + \phi) \ for \ t \geq 0. \quad (1)$$

Here, N is the order of the filter which determines the slope of the filter's skirts, b is the bandwidth of the filter, f_c is center frequency, a and ϕ are the scaling factor and phase of the cosine wave, respectively. In general, the order of the gammatone filter is chosen in between 3 to 5 to model the human auditory system [24]. The bandwidth b correspond to each f_c, is obtained using the Equivalent Rectangular Bandwidth (ERB) scale which is given by [25],

$$b = ERB(f_c) = 24.7(4.37f_c + 1) \quad (2)$$

where, b is in Hz and f_c is in kHz.

3) Constant-Q filter bank: Constant-Q filter bank (CQFB) is geometrically spaced filter bank with constant-Q factor (i.e. ratio of center frequency to the resolution is constant), such that resolution of the filters can be approximated to musical notes [26]. The k^{th} center frequency of constant Q transform is given by

$$f_k = f_0 \ 2^{k/B} \quad (3)$$

where, f_0 is minimum frequency, and B is number of bins per octave. The bandwidth of the filter b is given by

$$b = f_k \ (2^{1/B} - 1). \quad (4)$$

Constant-Q filters has high temporal resolution at high frequency and high frequency resolution at low frequency which also mimic the human auditory system.

4) Single frequency filter bank: The single frequency filter bank (SFFB) (as discussed in [27]), is based on single frequency filtering which provides good time-frequency resolution [28]. In single frequency filter bank approach speech signal is passed through a set of complex band pass filters to decompose signal into different frequency bands. The transfer function of the k^{th} filter is given by,

$$H_k(z) = \frac{1}{1 - a_k z^{-1}} \ k = 0, 1, 2, ...M \quad (5)$$

where, $a_k = ae^{-jw_k}$, a represents pole location, w_k is k^{th} frequency component, f_s corresponds to sampling frequency and M is total number of filters. The value of a which can be selected in between 0 to 1, determines the bandwidth of the filter. The narrow filters are designed to provide high spectral resolution by choosing the value of 'a' between 0.95 to 0.995.

B. Extraction of Long term average spectral features

The long term average spectrum features capture the static information like voice quality, gender information and age-related features from the speech signal [14]. To extract these features, first, the speech signal $s[n]$ is passed through the bank of filters (design of various filter banks will be discussed in Section 3) to decompose it into multiple time-frequency components. If $h_i[n]$ is filter's impulse response then the output of the filter is given by

$$s_i[n] = h_i[n] * s[n] \ i = 1, 2.....N \quad (6)$$

where N is the number of filters. All the N band signals along with original full-band signal in total $N + 1$ components are framed using a non-overlapping rectangular window of 20 ms. Then root mean square energy is calculated for each frame denoted by $s_{RMSi[k]}$ correspond to the k^{th} frame of i^{th} band. Finally, 10 statistical averages like normalized mean, standard deviation, range, skewness and kurtosis are calculated, the resulting $((N+1)*10-1)$ dimension feature vector is denoted as LTAS feature.

III. Experimental setup

This section describes the method to extract the various features used for studying voice disorder detection and assessment. In our previous work [11], voice disorder assessment task was performed in the clinical perspective where disorders were categorized into structural, neurogenic, functional and psychogenic from SVD database. Further details of the database, baseline features, and classifier used for this study are presented in the following section.

A. Database

Databases used in our study are saarbruecken voice disorder (SVD) dataset [21], and Hospital Universitario Prıncipe de Asturias (HUPA) database [22].

- The SVD[1] dataset contains voice recordings of more than 2000 subjects and 71 different voice disorders. Recordings are available for three vowel /a/, /i/, and /u/ in normal, high, low and rising-falling pitch. Moreover, the speech samples are also recorded using the German sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). In this study, the speech samples corresponding to voice disorders from SVD database were grouped into four classes as used in our previous study [11], namely, *Structural, Neurogenic, Functional and Psychogenic*. In this study 625 samples were considered from healthy class and total of 950 voice samples were considered from different voice disorders category for vowel /a/, /i/, and /u/ in normal, high, low and rising-falling pitch.
- The HUPA database contains recordings of the vowel /a/ for a total of 440 subjects. Out of total 366 recordings, 239 recordings are from pathological subjects, and 201 recordings are from normal subjects. It contains organic pathologies like Bilateral Reinke's edema, Polyp, Cyst, Bilateral nodule, Recurrent nerve paralysis etc. Auditory-perceptual ratings according to GRBAS scale is available for HUPA database. It contains the five different components, Grade of hoarseness (G), Roughness (R), Breathiness (B), Asthenia (A), and Strain (S). Each component is rated as 0,1,2 or 3, where 0 indicates normal, 1 mild, 2 moderate and 3 indicates more severe degree of voice disorder.

Table I describes the number of healthy and voice disorders samples of SVD and HUPA database used for detection task. Table II describes the different categories of voice disorders available for SVD database used in our study for the assessment of the voice disorders.

TABLE I
Details of the number of sample used for the detection task in our study from SVD and HUPA database.

SVD database		HUPA database	
Healthy	Voice Disorder	Healthy	Voice Disorder
659	950	239	201

[1]It is freely available at http://www.stimmdatenbank.coli.uni-saarland.de/

TABLE II
Details of the different classes of SVD database and number of sample used in our experiment for the assessment task. Here SD: Structural voice Disorder, NVD: Neurogenic Voice Disorder, FVD: Functional Voice Disorder, PVD: Psychogenic Voice Disorder

Disorder type	Disorder name	#Samples
Organic Voice Disorder	SD	352
	NVD	253
Non-organic Voice Disorder	FVD	254
	PVD	91

B. Feature Extraction

The features explored in this study include the LTAS features obtained by using the state-of-the-art filter banks, statistical averages of the short time features (LPCC, MFCC, PLP, etc.) and state-of-the-art openSMILE features such as eGEMAPS and ComParE. The extraction of these features is presented as follows.

1) LTAS based features: The parameters of each filter bank considered for extracting the LTAS features are described in the following subsection.

- CBFB-LTAS feature is calculated using 9-octave band signals and one full band speech signal. To get the time-frequency decomposition of the speech signal, first, the signal is passed through 9-octave band filters with the minimum centre frequency of 30 Hz and a maximum frequency of 3840 Hz. Finally, 99 (10*10-1) dimension CBFB-LTAS vector is obtained.
- For extraction of CQFB-LTAS feature vector, the speech signal is passed through the CQFB with 106 constant Q spaced filters. The CQFB is realized using f_{min} of $10Hz$, f_{max} of $4000Hz$, and number of bins per octave b of 12 [29]. In total, 107 components are used, resulting in 1069 (107*10-1) dimension LTAS feature vector.
- In case of GFB-LTAS feature extraction, the speech signal is decomposed by passing it through the 32 gammatone-tone filters [30]. The minimum and maximum frequency are selected as 0 Hz and 4000 Hz, respectively, which results in 329 (33*10-1) dimension feature vector.
- To extract the SFFB-LTAS feature vector, the speech signal is passed through 201 single pole band pass filter with minimum center frequency $f_{min} = 0Hz$ and maximum frequency $f_{max} = 4000Hz$. The pole location a of 0.98 and frequency resolution of 20 Hz were used to realize the SFFB (as in [20]). Total of 202 components (201 filter responses and speech signal) are used, results in 2019 (202*10-1) dimension LTAS feature vector.

2) Statistical averages of the state-of-the-art features: To compute the statistical averages, first, frame-level features were computed using a hamming window of size 25 ms with 10 ms frame shift. First m static cepstral coefficients and their delta, and delta-delta features were computed yielding in $d = 3*m$ dimension feature vector. Finally, statistical averages such as mean, standard deviation, kurtosis and skewness were derived from these frame-level features resulting in $D = (d*4)$

dimension feature vector named as STAT features as in [20]. Conventional MFCC, LPCC, PLP, and CQCC features, which captures the vocal tract information are used to compute corresponding STAT features, namely MFCC-STAT, LPCC-STAT, PLP-STAT, and CQCC-STAT. CQCC features were calculated from the CQT-transform with f_{min} of 100 Hz, f_{max} of 4000 Hz and bins per octave of 192 [26].

Along with the system features, we also explored the excitation source evidence such as LP-residual and zero frequency filtered (ZFF) signal to compute the STAT features. In this regard, MFCC features were computed from LP-residual and ZFF-signal as in [10]. Then corresponding STAT features were computed and are named as MFCC-WR-STAT and MFCC-ZFF-STAT, respectively. MATLAB implementation of the features used along with supporting material is provided in https://github.com/Purva-Barche/LTASfilterbankcodes.

3) OpenSMILE features: This work explored two state-of-the-art feature sets obtained from openSMILE tool kit [31] as baseline features. The first feature set is extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) which is low dimension knowledge-based acoustic feature [32]. It is 88 dimension feature set mainly used for extraction of emotion. The second set used is Computational Paralinguistic Challenge (ComParE) feature set which is brute-forced set [33]. It has a dimension of 6373 feature which are usually designed to extract paralinguistic information from the acoustic signal.

C. Classifier

The classifier used in our study for detection and assessment of voice disorders is the support vector machine (SVM) which is a supervised binary classifier. The detection and assessment of voice disorders were also done by using several other classifiers like decision tree, k-nearest neighbour, ensemble classifier and logistic regression. SVM is selected among all other classifiers due to its best classification accuracy. Among all different kernels like linear, radial basis functions, and polynomial, polynomial kernel with a polynomial degree of 2 outperformed in this study. Moreover, the grid search algorithm was performed to select the optimum value of kernel parameters. Further, five-fold cross-validation was performed to find the classification accuracy.

IV. Results and discussion

In our previous study [11], we have performed assessment of voice disorder in clinical way by using excitation source evidences. Among the individual excitation source features the intonation features derived from ZFF signal and MFCC-WR provided best classification accuracy of 69.3% and 70.8% for detection and assessment task, respectively. In continuous to our previous studies, the present study explored the significance of long term average spectral features using state-of-the-art filter banks for voice disorder detection and assessment tasks in the similar way to improve the performance of both the tasks.

Also, the performance of the detection and assessment system is compared with state-of-the-art openSMILE features and statistical averages of frame-level features. The detection system performs a binary classification to discriminate the speech samples corresponding to healthy and voice disorders. On the other hand, assessment is a multi-level classification problem in which three binary classifiers were used to identify the type of voice disorder. Total of four experiments were carried out in our paper.

- Experiment 1 (Voice disorder detection) was performed to discriminate healthy voice samples from the voice disorder sample of all the classes.
- In experiment 2, organic voice disorder samples were classified from non-organic voice disorder samples.
- In experiment 3 Organic voice disorder samples were further classified into structural and neurogenic voice disorder .
- Experiment 4 was conducted to classify functional voice disorders from psychogenic voice disorder category.

Voice disorder detection and assessment experiments were performed on the SVD dataset, whereas only detection task was performed on HUPA dataset as samples of different categories of voice disorders are not available for HUPA database. All the experiments were performed using the SVM classifier. Performance of the detection and assessment systems with individual baseline features and LTAS features obtained from various filter banks is reported in Table III for SVD database. Table IV shows the voice disorder detection (experiment 1) result for HUPA database. In addition, the performance of detection and assessment systems was evaluated using the combination of filter bank features with the state-of-the-art openSMILE features, and the results are presented on SVD and HUPA database in Table V. Further, the relation between the LTAS features and perceptual scale was evaluated using N-way analysis of variances (ANOVA).

From Table III, it is evident that, among all STAT features PLP-STAT features shows better classification accuracy of 78.4% and 74.7% for experiment 1 and 3 respectively . Further,

TABLE III
PERFORMANCE OF VOICE DISORDER DETECTION AND ASSESSMENT SYSTEMS IN TERMS OF CLASSIFICATION ACCURACY (IN %) FOR INDIVIDUAL FEATURE SET ON SVD DATABASE. HERE, EXP. 1: CLASSIFICATION OF HEALTHY AND VOICE DISORDER, EXP. 2: CLASSIFICATION OF ORGANIC AND NON-ORGANIC VOICE DISORDERS, EXP. 3: CLASSIFICATION OF STRUCTURAL AND NEUROGENIC VOICE DISORDERS, EXP. 4: CLASSIFICATION OF FUNCTIONAL AND PSYCHOGENIC VOICE DISORDERS, S1 STATISTICAL AVERAGE FEATURE SET, S2 OPENSMILE FEATURE SET, S3 LTAS FEATURES.

	Feature	Exp.1	Exp.2	Exp.3	Exp.4
S1	MFCC-STAT	76.1	71.6	69.9	68.2
	PLP-STAT	78.4	71.2	74.7	66.2
	LPCC-STAT	75.6	68.6	70.4	65.3
	CQCC-STAT	74.4	70.3	71.2	**70.8**
	MFCC-WR-STAT	72	70.1	66.3	65.9
	MFCC-ZFF-STAT	71.3	69.3	70.6	69.1
S2	eGeMAPS	80.7	71	70.6	64.5
	ComParE	**85.9**	**75.7**	**76.5**	69.4
S3	CBFB-LTAS	74.3	69.9	68.6	66.2
	GFB-LTAS	76.9	71.4	69.9	67.9
	CQFB-LTAS	78	70.8	71.2	65.9
	SFFB-LTAS	76.8	69	69.1	65.9

TABLE IV
PERFORMANCE OF VOICE DISORDER DETECTION SYSTEMS IN TERMS OF CLASSIFICATION ACCURACY (IN %) FOR HUPA DATABASE. HERE, S1 STATISTICAL AVERAGE FEATURE SET, S2 OPENSMILE FEATURE SET, S3 LTAS FEATURES.

	Features	Accuracy (%)
S1	MFCC-STAT	69.2
	LPCC-STAT	69.2
	PLP-STAT	73.7
	CQCC-STAT	62.3
	MFCC-WR-STAT	70.2
	MFCC-ZFF-STAT	69.9
S2	eGeMAPS	76.1
	ComParE	**82.1**
S3	CBFB-LTAS	75.9
	CQFB-LTAS	81.4
	GFB-LTAS	79.2
	SFFB-LTAS	74.9

TABLE V
PERFORMANCE OF VOICE DISORDER DETECTION AND ASSESSMENT SYSTEMS IN TERMS OF CLASSIFICATION ACCURACY (IN %) FOR COMBINATION OF FEATURE SETS ON SVD AND HUPA DATABASE. HERE, EXP. 1: CLASSIFICATION OF HEALTHY AND VOICE DISORDER, EXP. 2: CLASSIFICATION OF ORGANIC AND NON-ORGANIC VOICE DISORDERS, EXP. 3: CLASSIFICATION OF STRUCTURAL AND NEUROGENIC VOICE DISORDERS, AND EXP. 4: CLASSIFICATION OF FUNCTIONAL AND PSYCHOGENIC VOICE DISORDERS.

Features	SVD				HUPA
	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 1
CBFB-LTAS+eGeMAPS	**89.6**	71.9	72.9	69.1	80
CBFB-LTAS+ComParE	86	76.1	77.2	67.1	85.4
GFB-LTAS+eGeMAPS	87.5	73	70.2	64.5	82.1
GFB-LTAS+ComParE	85.8	77.2	77	**69.7**	81.1
CQFB-LTAS+eGeMAPS	84.2	72.9	69.9	67.6	83
CQFB-LTAS+ComParE	87.2	78.3	75	67.9	**86.6**
SFFB-LTAS+eGeMAPS	84.1	68.7	69.9	64.5	78.2
SFFB-LTAS+ComParE	86.9	**78.9**	**77.4**	68.5	81.3

ComParE feature set outperformed for all the experiments. Among all LTAS features, CQFB-LTAS performed better for experiment 1 and 3, while GFB-LTAS performed better for experiment 2 and 4. Moreover, the performance of the CQFB-LTAS features (78%, 70.8% and 71.2%) is comparable to the baseline eGeMAPS features (80.7%, 71% and 70.6%) for three experiments.

Table IV shows the voice disorder detection (only experiment 1) results on HUPA dataset using the different baseline features and LTAS based features. From the table it is evident among all the STAT features PLP-STAT features shows better classification accuracy of 73.7% for HUPA datset. Further, the best performance is obtained in term of classification accuracy of 82.1% for ComParE feature sets. Among the filter bank based LTAS features, CQFB-LTAS performed best with a classification accuracy of 81.4%.

Among baseline feature sets, the openSMILE features showed better classification accuracy compared to statistical feature sets; hence, the performance was also observed by combining the LTAS feature sets with openSMILE feature sets as reported in Table V for SVD (all the experiments) and HUPA (only experiment 1) database. It can be observed from the Table V for the detection task best classification accuracy of 89.6% is obtained when CBFB-LTAS features combined with eGeMAPS features for SVD database. For HUPA database the best classification accuracy of 86.6% is observed when constant-Q based LTAS features were combined with ComParE feature sets. SFFB-LTAS features when combined with ComParE performed best among all other (for SVD samples) combinations for experiment 2 and 3. It can also be observed that even by combining the different features, classification accuracy for the experiment 4 is not increased significantly, as psychogenic voice disorder samples mostly confused with functional voice disorder.

To assess the relationship with the perceptual scale used by SLPs, statistical analyses were computed with N-way ANOVA by considering the LTAS feature as a dependent variable and perceptual ratings of Grade of hoarseness, Roughness, Breathiness, Asthenia and Strain as independent variables. ANOVA was computed on the HUPA dataset which has a perceptual rating according to the GRBAS scale. Out of 99 LTAS features, 35 features show the minimum value of p for the perceptual scale of Roughness, 31 features indicate the minimum value for Asthenia. Remaining 14 features out of 99 LTAS features indicates the least value of p (very smaller than 0.5) for overall degree of hoarseness, while 11 LTAS features and 8 LTAS features shows the minimum value of p for perceptual scale of breathiness and strain, respectively. Moreover, multivariate ANOVA was also obtained for different frequency ranges. Two frequency ranges were considered, one from 0 to 1 KHz and other above 1 KHz. It was observed that for the frequency range below 1 KHz, 31 and 27 LTAS features out of 69 features indicate the minimum value of p for perceptual scale R (Roughness) and Asthenia respectively. For the frequency range above 1 KHz perceptual rating, G(Overall severity) and S (Strain) indicate the minimum value of p for most of the LTAS features. Thus from this ANOVA analysis we can conclude that LTAS features indicate the stronger correlation with roughness (which might be due to degradation in the voice quality) and asthenia (indicates the degree of vocal weakness) compared to other perceptual characteristics.

V. SUMMARY AND CONCLUSION

This paper explores the state-of-the-art filter bank-based LTAS features for the detection and assessment of voice disorder. From the experimental results, it can be verified that classification accuracy for an assessment system is less compared to detection system, as different disorders may share a common acoustic space. More interestingly, it was observed that the choice of filter bank in the extraction of LTAS features play an important role in the classification of voice disorders. In [20], SFFB based LTAS features showed the best performance for hyper-nasality detection, whereas, in this study, the SFFB-LTAS features showed better performance than CBFB-LTAS for the detection task. The CQFB-LTAS and GFB-LATS features showed better classification accuracy for the detection and assessment of voice disorders, perhaps due to the underlying filter banks (constant Q filters and Gammatone filters) that were designed to mimic the human auditory system. In

addition, an improvement in the performance of detection and assessment systems was observed with the combination of feature sets, which highlights the complementary nature of filter bank-based LTAS features. Further, we evaluated the relation between LTAS features and perceptual measure (GRBAS scale available for HUPA database) using ANOVA analysis. The results from this experiment suggested that, most of the LTAS features have least value of the p (less than 0.5) for roughness and asthenia compared to grade, breathiness and strain. Compared to our previous study [11], significant improvement of performance for all the experiments was observed which might be due to the reason that, long term features can capture the voice disorders information in a better way as compared to the short term variations.

In future work, we intend to study which frequency band are more important for voice disorder detection and assessment system. LTAS feature may capture age and gender-related information which should be normalized to improve the performance.

ACKNOWLEDGMENT

Authors would like to thank IHub-Data Technology Innovation Hub (TIH) @ IIIT Hyderabad for providing the research fellowship.

REFERENCES

[1] A. E. Aronson, "Clinical voice disorders," *An Interdisciplinary Approach*, 1985.

[2] B. Barsties and M. De Bodt, "Assessment of voice quality: current state-of-the-art," *Auris Nasus Larynx*, vol. 42, no. 3, pp. 183–188, 2015.

[3] L. Sulica, "Laryngoscopy, stroboscopy and other tools for the evaluation of voice disorders," *Office Procedures in Laryngology, An Issue of Otolaryngologic Clinics-E-Book*, vol. 46, no. 1, p. 21, 2012.

[4] M. P. Karnell and et al., "Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders," *J. of Voice*, vol. 21, no. 5, pp. 576–590, 2007.

[5] R. D. Kent, "Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders," *American J. of Speech-Lang. Pathology*, vol. 5, no. 3, pp. 7–23, 1996.

[6] D. D. Mehta and R. E. Hillman, "Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods," *Current opinion in otolaryngology & head and neck surgery*, vol. 16, no. 3, p. 211, 2008.

[7] J. Laver, S. Hiller, and J. M. Beck, "Acoustic waveform perturbations and voice disorders," *Journal of Voice*, vol. 6, no. 2, pp. 115–126, 1992.

[8] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[9] J. C. Saldanha, T. Ananthakrishna, and R. Pinto, "Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features," *Journal of medical imaging and health informatics*, vol. 4, no. 2, pp. 168–173, 2014.

[10] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. of Selected Topics in Signal Process.*, vol. 14, no. 2, pp. 367–379, 2019.

[11] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach," *Proc. Interspeech 2020*, pp. 2537–2541, 2020.

[12] T. Leino, "Long-term average spectrum in screening of voice quality in speech: untrained male university students," *J. of Voice*, vol. 23, no. 6, pp. 671–676, 2009.

[13] B. Hammarberg, B. Fritzen, J. Gauffin, and J. Sundberg, "Acoustic and perceptual analysis of vocal dysfunction," *Journal of phonetics*, vol. 14, no. 3-4, pp. 533–547, 1986.

[14] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (ltas)," *Journal of voice*, vol. 10, no. 1, pp. 59–66, 1996.

[15] K. Tanner, N. Roy, A. Ash, and E. H. Buder, "Spectral moments of the long-term average spectrum: sensitive indices of voice change after therapy?" *Journal of Voice*, vol. 19, no. 2, pp. 211–222, 2005.

[16] D. Sergeant and G. F. Welch, "Age-related changes in long-term average spectra of children's voices," *J. of Voice*, vol. 22, no. 6, pp. 658–670, 2008.

[17] G. Kovačić, P. Boersma, and H. Domitrović, "Long-term average spectra in professional folk singing voices: A comparison of the klapa and dozivački styles," *Proc. Inst. of Phonetic Sciences, Univ. of Amsterdam*, vol. 25, pp. 53–64, 2003.

[18] T. F. Cleveland, J. Sundberg, and R. Stone, "Long-term-average spectrum characteristics of country singers during speaking and singing," *Journal of voice*, vol. 15, no. 1, pp. 54–60, 2001.

[19] K. Peter, "LTAS criteria pertinent to the measurement of voice quality," *J. of Phonetics*, vol. 14, no. 3-4, pp. 477–482, 1986.

[20] M. H. Javid, K. Gurugubelli, and A. K. Vuppala, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6754–6758.

[21] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[22] J I Godino-Llorente and et al., "Acoustic analysis of voice using wpcvox: a comparative study with multi dimensional voice program," *European Archives of Oto-Rhino-Laryngology*, vol. 265, no. 4, pp. 465–476, 2008.

[23] R. D. Patterson and et al., "Complex sounds and auditory images," in *Auditory Physiology and Perception*. Elsevier, 1992, pp. 429–446.

[24] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[25] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[26] J. C. Brown, "Calculation of a constant q spectral transform," *JASA*, vol. 89, no. 1, pp. 425–434, 1991.

[27] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6410–6414.

[28] G. Aneeja and Y. Bayya, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 23, no. 4, pp. 705–717, 2015.

[29] Y. Panagakis and C. Kotropoulos, "Music classification by low-rank semantic mappings," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–15, 2013.

[30] H. K. Maganti and M. Matassoni, "Auditory processing-based features for improving speech recognition in adverse acoustic conditions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–9, 2014.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[32] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[33] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," in *Sixteenth annual conference of the international speech communication association*, 2015.