PERCEPTUALLY ENHANCED SINGLE FREQUENCY FILTERING FOR DYSARTHRIC SPEECH DETECTION AND INTELLIGIBILITY ASSESSMENT

Krishna Gurugubelli, Anil Kumar Vuppala

Speech Processing Laboratory, LTRC, KCIS
International Institute of Information Technology, Hyderabad, India
krishna.gurugubelli@research.iiit.ac.in, anil.vuppala@iiit.ac.in

ABSTRACT

This paper proposes a new speech feature representation that improves the intelligibility assessment of dysarthric speech. The formulation of the feature set is motivated from the human auditory perception and high time-frequency resolution property of single frequency filtering (SFF) technique. The proposed features are named as perceptually enhanced single frequency cepstral coefficients (PE-SFCC). As a part of SFF technique implementation, speech signal passed through a single pole complex bandpass filter bank to obtain high-resolution time-frequency distribution. Then, the distribution is enhanced by using a set of auditory perceptual operators. Lastly, traditional homomorphic analysis has been carried out on the resulting signal to obtain PE-SFCC feature vector. The performance of proposed features in dysarthric speech detection and its intelligibility assessment has been reported on UASPEECH database. The PE-SFCC features outperformed the state-of-the-art features in dysarthric speech detection and intelligibility assessment.

Index Terms— Dysarthria detection, Intelligibility assessment, Auditory perception, Single frequency filtering

1. INTRODUCTION

Dysarthria is a speech disorder emanating from neurological damage connected with motor control of speech muscles. The abnormalities in the resonance, articulation, respiration, phonation, and prosody of speech are associated with dysarthria leads to poor speech intelligibility. The symptoms of poor speech quality and reduced intelligibility can be used to identify the dysarthria [1, 2]. The subjective intelligibility assessment methods may influence by the listener familiarity with patients, the contextual, suprasegmental factors, and semantic/syntactic features. Moreover, the subjective intelligibility assessment methods are costly and time-consuming [3, 4]. Objective intelligibility assessment methods, on the other hand, are economical, repeatable, reliable and can assist in remote patient rehabilitation monitoring. The growing evidence suggesting that clinicians are becoming more receptive to objective intelligibility assessment systems in which the speech intelligibility can be assessed by the trained acoustic model [5, 6].

In literature wide range of acoustic features capturing the vocal tract dynamics, excitation source information, and prosody information are explored for the objective assessment of dysarthric speech intelligibility. In [7, 8, 9] spectral features (e.g., Mel-frequency cepstral coefficients, spectral centroid, linear prediction cepstral coefficients) are used to represent vocal tract shape and dynamics. In [10] perceptual linear prediction (PLP)features are used in the analysis of Parkinson dysarthria. Excitation source features of speech (e.g., inter-pulse similarity, average magnitude difference function, residual peak prominence, jitter, and shimmer) are used to represent the voice quality [11, 7]. In [12, 13] glottal source parameters (e.g., normalized amplitude quotient, open quotient, closing quotient, and harmonic richness factor) are used to capture the wide variations in the excitation source. In [14, 15, 16], long-term spectral average features, ITU-T P.563 features, long-term temporal dynamics are evaluated as correlates of subjective intelligibility in the analysis of spastic dysarthria. The long-term spectro-temporal dynamic measures are correlated to loudness, intonation, stress, and rhythm of speech. It is also understood that a better feature representation which can capture wide variabilities in dysarthric speech can discriminate the dysarthria accurately. Most of the traditional methods are based on block processing, and they find the average spectrum over the 20-30ms window. Hence, the abnormal spectral-temporal variations present in dysarthric speech can be averaged. Single frequency filtering (SFF) based instantaneous spectral representation is investigated for robust epoch extraction, speaker separation and voice activity detection [17, 18]. The SFF technique gives a good spectro-temporal representation of speech which may capture dysarthric speech information effectively.

In this work, SFF based instantaneous spectral representation is investigated to detect dysarthria from speech. SFF time-frequency representation is further perceptually enhanced based on auditory perception. Post perceptual operations homomorphic analysis is done on SFF time-frequency representation to get the cepstral representation of speech signal. The feature representation is named as perceptually enhanced single frequency cepstral coefficients (PE-SFCC). The proposed features are used to train dysarthric speech detection system using i-vector representation with probabilistic linear discriminant analysis (PLDA) scoring mechanism.

The rest of the paper is organized as follows: Perceptually enhanced single frequency filtering based speech feature extraction is introduced in Section 2. In Section 3, the experimental setup for dysarthric speech detection is discussed. Section 4 describes the results and discussions. Section 5 provides the conclusions of the paper.

2. PROPOSED FEATURE EXTRACTION METHODOLOGY

The proposed feature extraction framework is implemented in two steps. In the first step, time-frequency distribution of speech is obtained using SFF technique. In the next step, auditory perceptual enhancement is done on the time-frequency distribution to improve

the discriminative capability by modeling production irregularities of dysarthric speech. The algorithmic steps involved in the feature extraction process are discussed in the subsequent subsections 2.1 and 2.2.

2.1. Time-frequency Representation of speech signal using single frequency filtering

The generalization of single frequency filtering [17] method in generating the time-frequency distribution of speech signal is now presented. This technique employs a set of complex bandpass filters to decompose the signal into different frequency bands. As a part of the filtering, the speech signal is convolved with a set of complex-valued coefficients. These filter coefficients are derived from the design of low pass filter having an appropriate frequency response. The transfer function of the low pass filter prototype is given by

$$H(z) = \frac{1}{1 - az^{-1}}. \tag{1}$$

In the above equation the value of a determines the pole location. A k^{th} frequency component can be decomposed using a complex band pass filter. It is realized by modulating the single pole low pass filter, by multiplying its impulse response $h[n]$ with a complex sinusoid $e^{(jw_k n)}$. Its transfer function is given by,

$$H_k(z) = \frac{1}{1 - a_k z^{-1}} \tag{2}$$

where $a_k = ae^{-jw_k}$ and w_k represents the k^{th} frequency component. Here $w_k = \frac{\tilde{w}_k * 2 * \pi}{f_s}$ is discrete frequency corresponds to analog frequency $\tilde{w}_k$ and f_s is the sampling frequency. While tracking the energy variations of a particular frequency component, bandwidth of the filter should be kept narrow in order to achieve better spectral resolution. The bandwidth of the filter and a are related as,

$$BW_{H(z)} = cos^{-1}\left(\frac{4a - a^2 - 1}{2a}\right). \tag{3}$$

The value of "a" lies in between 0 and 1. A sampling rate of 8000 Hz and "$a = 0.9875$" are considered to obtain filter bandwidth of 20 Hz approximately. Using equations 2 and 3, a complex bandpass filter bank is realized to decompose M frequency components of a speech signal and it can be represented as,

$$SF_{filterbank} = \begin{pmatrix} H_1(w) \\ H_2(w) \\ . \\ . \\ H_k(w) \end{pmatrix}, k = 1, 2, 3....M \tag{4}$$

where M is the total number of frequency bands. In this work, frequency components are decomposed with a frequency resolution of 20 Hz from 100 Hz to 4000 Hz. The speech signal $x[n]$ is convolved with the impulse responses of different channels of the single frequency filter-bank. The output filtered frequency components in the time-frequency plane are given by

$$S(k, n) = \begin{pmatrix} y_1[n] \\ y_2[n] \\ . \\ . \\ y_M[n] \end{pmatrix}, k = 1, 2, 3....M. \tag{5}$$

where

$$y_k[n] = \sum_{i=1}^{N} h_k[i]\, x[n - i] \tag{6}$$

and N is length of the speech signal, $h[n]$ represents the impulse response of the filter. The envelope of each filtered component is represented as,

$$m_k[n] = \sqrt{y_{kR}^2[n] + y_{kI}^2[n]} \tag{7}$$

$y_{kR}[n]$ and $y_{kI}[n]$ are the real and imaginary parts of the filtered component $y_k[n]$. To demonstrate the efficacy of SFF, time-

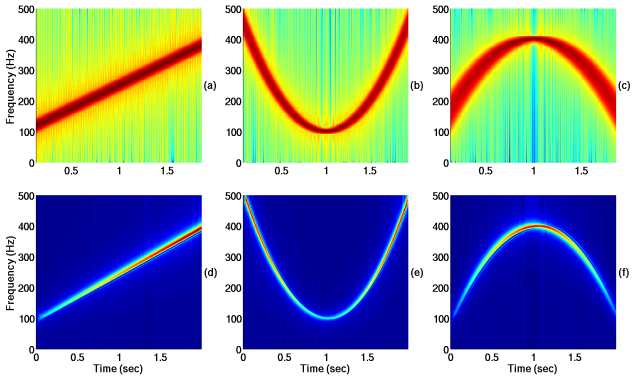

Fig. 1. Time-frequency representation of synthesized linear, quadratic, and convex chirp signals. Short time Fourier transform (Top row: (a)-(c)). Single frequency filtering (Bottom row: (d)-(f))

frequency representation of synthesized chirp signals are depicted in Fig. 1. Top row demonstrates a short time Fourier transform (STFT) based time-frequency representation for a linear, quadratic, and convex chirp signals. The STFT of chirp signals is estimated using a window size of 10 ms with an overlap of 5 ms. The bottom row shows SFF based time-frequency representation (SFF-TFR) for same signals. It can be noticed that the frequency spread is less in case of SFF-TFR as compared to STFT representation. It is also observed that better time-frequency localization can be observed in SFF-TFR. In Fig. 2, SFF based time-frequency representation of speech at time-frequency resolutions is demonstrated for different values of a. Fig. 2.b represent the time-frequency distribution of speech signal with "$a = 0.85$". It possess good time resolution, but the formant information is smeared. While increasing the value of "a" from 0.85 to 1.0, it can be noted that formant information is resolved in a better manner with a decrease in time resolution. A good time-frequency trade-off can be made for the value of "a" in between 0.95 to 0.995, highlighting the formant information with reasonably good time resolution. Hence, speech signals have been analyzed with $a = 0.9875$ and $f_s = 8000Hz$ in this framework.

2.2. Perceptually enhanced SFF based feature extraction

SFF based speech spectrum is enhanced based on the human auditory system to realize the perceptually enhanced SFF spectrum. The block diagram representation of the feature extraction framework is shown in Fig. 3. The feature extraction procedural steps are listed below:

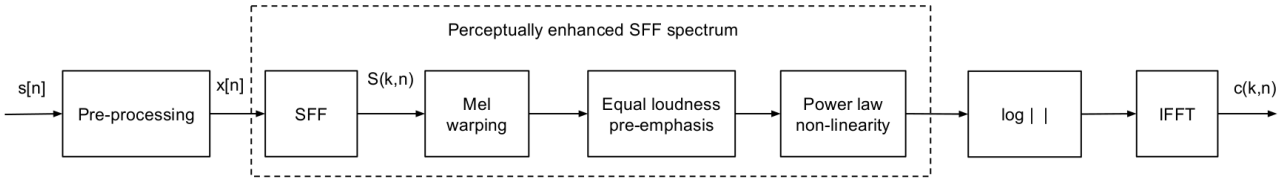

Fig. 3. Block diagram of PE-SFCC feature extraction framework.

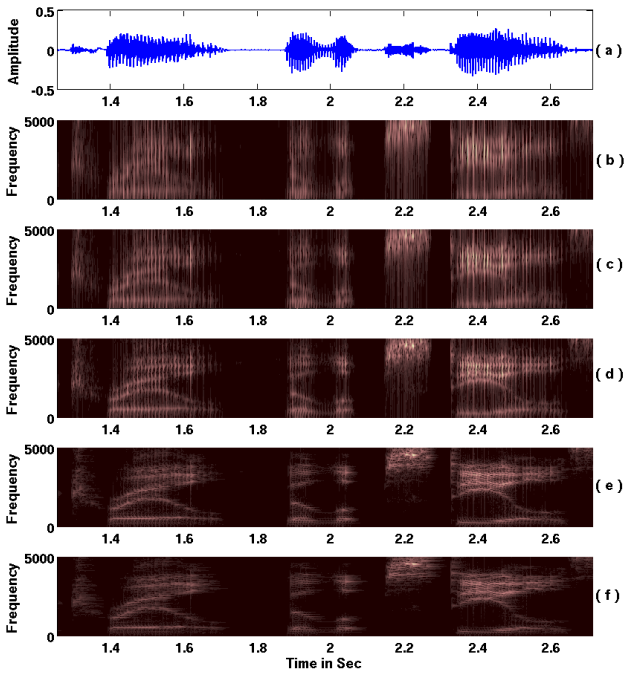

Fig. 2. Time-frequency representation of speech signal with different resolutions. (a) Speech signal. (b)-(f) SFF-TF representations of speech for a = 0.85, 0.90, 0.95, 0.99, and 0.995 respectively.

1. The SFF-TFR $S(k,n)$ of the pre-emphasized speech signal is estimated using SFF technique as discussed in subsection 2.1.
2. The Mel frequency warping is carried out to achieve the non-linear frequency scaling property of the human auditory system [19]. The warped spectrum is given by $S_W(k,n) = \psi_m\{S(k,n)\}$. Here ψ_m represents Mel warping operator.
3. The warped spectrum is further processed using equal loudness pre-emphasis contour [20] to model the non-uniform sensitivity of human hearing at different frequencies. It is given by,
$$S_{WE}(k,n) = \psi_e\{S(k,n)\} \quad (8)$$
here ψ_e represents equal loudness pre-emphasis operator.
4. To simulate the non-linear relationship between sound intensity and perceived loudness, the power law non-linearity with exponent 1/5 is operated on $S_{WE}(k,n)$ [21]. This results,
$$S_{WEP}(k,n) = S_{WE}(k,n)^{1/5} \quad (9)$$
5. Inverse Fourier transform computed for the logarithm of the power spectrum to realize the cepstral representation. A 13 dimensional cepstral feature representation is obtained through liftering operation. The cepstral representation of speech is denoted as,
$$c(k,n) = IFFT\{log\{ S_{WEP}(k,n) \} \} \quad (10)$$

As PE-SFF gives instantaneous cepstral representation $c(k,n)$, subsampling is done to reduce the feature size [22]. A 39 dimensional PE-SFCC includes the 13 cepstral coefficients, first and second time derivatives.

3. EXPERIMENTAL SETUP

In this Section, experimental framework for the aforementioned task using the proposed feature is discussed.

3.1. UASPEECH Database

The UASPEECH is a publicly available dysarthric speech database that consists of 16 dysarthric speakers and 13 healthy speakers. From each speaker, 765 isolated words that include 300 uncommon words and three repetitions of common words, computer commands, digits, and radio alphabets are recorded. Overall intelligibility of each dysarthric speaker is measured based on listening tests done with native listeners. The intelligibility ratings are in the range of 2% to 95%. Based on the speech intelligibility, dysarthric speakers are grouped into four categories namely, very low (0-25%), low (25-50%), medium (50-75%) and high (75-100%) [23]. The speech data from these dysarthric speaker groups used for the automatic dysarthric speech intelligibility assessment. In this work, dysarthric speech utterances correspond to uncommon words are used for testing and the rest of the data used for training. Leave one speaker out evaluation scheme is considered in the experimentation, so that test speaker data is not exposed to the trained model.

3.2. Features used in dysarthric speech intelligibility assessment

Dysarthric speech intelligibility assessment has been carried out using perceptual linear prediction (PLP) features in joint factor space [24]. In our work, we considered these features as one of the primary baseline features. As discussed in Section 2, the proposed PE-SFCC feature representation with SFF frequency step of 20Hz, values of a = 0.98 and sampling rate $fs = 8000Hz$ is explored for the dysarthric classification. The performance of PE-SFCC features is also compared with the following state-of-the-art features:

- Mel-frequency cepstral coefficients (MFCC) computed from a 20 ms window having an overlap of 5 ms using 512 point FFT are considered.

- Multi-taper based spectrum is computed using 7 Thomson orthonormal tapers [25] from which Mel frequency cepstral coefficients are estimated. These features are considered as multi-taper MFCC features.
- Constant Q transform (CQT) [26] is one of the perceptually enhanced time-frequency representation of speech. The constant Q cepstral coefficients (CQCCs) are computed by using an open source Matlab implementation of CQT[1]. In computing the CQT, the number of bins per octave is set to 48, f_{min} = 100 Hz and $f_{max} = fs/2$ Hz. A 39-dimensional CQCC feature representation is obtained from the 13-dimensional static features of CQT spectrum.

3.3. The i-vector based classification system

In this work, i-vector with PLDA scoring mechanism is investigated for dysarthric speech detection and intelligibility assessment. In this system, a Gaussian mixture model with a universal background model (GMM-UBM) of 512 mixtures is trained with 10 expectation-maximization iterations using the TIMIT database. The 100 dimensional T matrix is considered in extracting the i-vectors. The i-vector dimensions are reduced to 10 by using linear discriminant analysis. Then, PLDA scoring mechanism is used for classification. This work uses MSR identity toolbox[2] to train UBM and T matrix.

4. RESULTS AND DISCUSSION

Dysarthric speech detection and intelligibility assessment in speaker-independent conditions is a challenging problem [24]. This work aims to improve the performance of dysarthric speech detection and intelligibility assessment in speaker-independent conditions on UASPEECH database. Firstly, the model is trained to discriminate dysarthric speech from the healthy speech which is considered as dysarthric speech detection (DSD) system. Additionally, dysarthric speech is further classified into different intelligibility groups to assess the severity. This classification process of severity levels is considered as dysarthric speech intelligibility assessment (DSIA), where the model is trained to discriminate different intelligibility levels of dysarthric speech. It is observed that most of the utterances from UASPEECH database contain long silence regions and they are trimmed to 50ms by using SOX toolkit.

The confusion matrix for the DSIA system trained with PE-SFCC feature is depicted in Fig. 4. From the confusion matrix, it is concluded that the system is found to be optimally trained without over-fitting. It is observed that there is a high correlation between neighboring classes. Out of all the classes, the classification accuracy is less for the class "Medium" may be due to the less amount of training data. The DSD and DSIA systems implemented by using the experimental setup outlined in Section 3. The results are shown in Table. 1. It is observed that the proposed feature outperformed other state-of-the-art features in terms of detection accuracy. However, it is observed that the CQCC and multi-taper MFCC features are nearly comparable to the proposed features in terms of dysarthric speech detection. The detection accuracy of the DSD system is found to be higher than the DSIA system, and this could be connected to the fact that the dysarthric speech is highly contrasting with healthy speech. Compared to MFCC and PLP features, the multi-taper MFCC, CQCC, and PE-SFCC features performed well in classifying the dysarthric speech from healthy speech. The better performance of the proposed feature can be attributed to the perceptual enhancement done on the time-frequency spectrogram. It is hypothesized that the vocal-tract irregularities of dysarthric speech are adequately captured in feature representation. Moreover, SFF spectrum highlights the gross level speech information with good time-frequency trade-off and avoiding block processing.

[1] http://audio.eurecom.fr/content/software

[2] The MSR identity toolbox has matlab implementations of the GMM-UBM and state-of-the-art i-vector-PLDA based classification methods. https://www.microsoft.com/en-us/research/publication/msr-identity-toolbox-v1-0-a-matlab-toolbox-for-speaker-recognition-research-2/

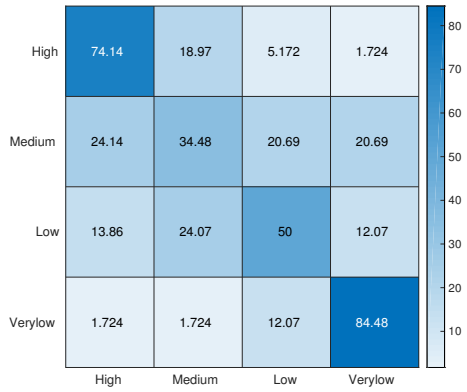

Fig. 4. Confusion matrix of DSIA system trained with PE-SFCC feature set.

Table 1. Comparison between of PE-SFCC and other state-of-the-art features in DSIA and DSD on UASPEECH database

	DSIA system Accuracy in %	DSD system Accuracy in %
PLP features	45.55	80.01
MFCC	42.07	78.70
Multi-taper MFCC	50.43	88.79
CQCC	49.14	91.38
PE-SFCC	**60.78**	**93.64**

In this work, the intelligibility assessment of dysarthria is done by using the vocal tract dynamics, derived from SFF time-frequency representation. In general, dysarthric speech assessment incorporates the knowledge of vocal tract dynamics, excitation information, and speech prosody. The combination of PE-SFCC features with glottal and prosody features may give the balanced intelligibility assessment of dysarthria.

5. CONCLUSIONS

In this paper, proposed a new feature representation named perceptually enhanced single frequency filtering based cepstral coefficients (PE-SFCC) for dysarthric speech classification. This feature set is formulated from the perceptually enhanced time-frequency distribution of the speech signal obtained using single frequency filtering technique. The proposed features representation exploits the auditory perceptual enhancement and good time-frequency resolution of SFF technique. On experimental evaluation, the objective measures reveal that the proposed feature set outperformed other state-of-the-art-features such as MFCC, PLP, multi-taper MFCC, and CQCC features for dysarthric speech intelligibility assessment. The PE-SFCC features achieved 60% and 93% accuracy in dysarthric speech severity assessment and dysarthric speech detection respectively.

6. REFERENCES

[1] Joseph R Duffy, *Motor Speech Disorders-E-Book: Substrates, Differential Diagnosis, and Management*, Elsevier Health Sciences, 2013.

[2] Cynthia Marie Fox and Carol Ann Boliek, "Intensive voice treatment (LSVT LOUD) for children with spastic cerebral palsy and dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 3, pp. 930–945, 2012.

[3] Marie Klopfenstein, "Interaction between prosody and intelligibility," *International Journal of Speech-Language Pathology*, vol. 11, no. 4, pp. 326–331, 2009.

[4] Gwen Van Nuffelen, Catherine Middag, Marc De Bodt, and Jean-Pierre Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International journal of language & communication disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[5] Gabriella Constantinescu, Deborah Theodoros, Trevor Russell, Elizabeth Ward, Stephen Wilson, and Richard Wootton, "Assessing disordered speech and voice in parkinson's disease: A telerehabilitation application," *International Journal of Language & Communication Disorders*, vol. 45, no. 6, pp. 630–644, 2010.

[6] Andreas Maier, Tino Haderlein, Ulrich Eysholdt, Frank Rosanowski, Anton Batliner, Maria Schuster, and Elmar Nöth, "PEAKS–A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[7] Tiago H Falk, Wai-Yip Chan, and Fraser Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, 2012.

[8] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer speech & language*, vol. 29, no. 1, pp. 132–144, 2015.

[9] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kumar Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *Proc. ICASSP*, 2017, pp. 5070–5074.

[10] Achraf Benba, Abdelilah Jilbab, and Ahmed Hammouch, "Discriminating between patients with Parkinsons and neurological diseases using cepstral analysis," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 24, no. 10, pp. 1100–1108, 2016.

[11] Frank Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. ICASSP*, 2009, pp. 4605–4608.

[12] Stephanie Gillespie, Yash-Yee Logan, Elliot Moore, Jacqueline Laures-Gore, Scott Russell, and Rupal Patel, "Cross-database models for the classification of dysarthria presence," in *Proc. Interspeech*, 2017, pp. 3127–3131.

[13] Narendra Nonavinakere Prabhakera, Paavo Alku, et al., "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. Interspeech*, 2018.

[14] Susan J LeGendre, Julie M Liss, and Andrew J Lotto, "Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra.," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2530–2530, 2009.

[15] Julie M Liss, Sue LeGendre, and Andrew J Lotto, "Discriminating dysarthria type from envelope modulation spectra," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1246–1255, 2010.

[16] Visar Berisha, Julie Liss, Steven Sandoval, Rene Utianski, and Andreas Spanias, "Modeling pathological speech perception from data with similarity labels," in *Proc. ICASSP*, 2014, pp. 915–919.

[17] G Aneeja and B Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[18] Sudarsana Reddy Kadiri and B Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[19] Anamitra Makur and Sanjit K Mitra, "Warped discrete-Fourier transform: Theory and applications," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 9, pp. 1086–1093, 2001.

[20] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[21] Chanwoo Kim and Richard M Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[22] K.N.R.K. Raju Alluri, Sivanand Achanta, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty, and Anil Kumar Vuppala, "Detection of replay attacks using single frequency filtering cepstral coefficients," in *Proc. Interspeech*, 2017, pp. 2596–2600.

[23] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[24] David Martínez, Eduardo Lleida, Phil Green, Heidi Christensen, Alfonso Ortega, and Antonio Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, pp. 10, 2015.

[25] David J Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[26] Judith C Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.