AUTOMATIC ASSESSMENT OF DYSARTHRIA SEVERITY LEVEL USING AUDIO DESCRIPTORS

Chitralekha Bhat, Bhavik Vachhani, Sunil Kumar Kopparapu

TCS Innovation Labs, Mumbai, India
Email: {bhat.chitralekha, bhavik.vachhani, sunilkumar.kopparapu}@tcs.com

ABSTRACT

Dysarthria is a motor speech impairment, often characterized by speech that is generally indiscernible by human listeners. Assessment of the severity level of dysarthria provides an understanding of the patient's progression in the underlying cause and is essential for planning therapy, as well as improving automatic dysarthric speech recognition. In this paper, we propose a non-linguistic manner of automatic assessment of severity levels using audio descriptors or a set of features traditionally used to define timbre of musical instruments and have been modified to suit this purpose. Multi-tapered spectral estimation based features were computed and used for classification, in addition to the audio descriptors for timbre. An Artificial Neural Network (ANN) was trained to classify speech into various severity levels within Universal Access dysarthric speech corpus and the TORGO database. An average classification accuracy of 96.44% and 98.7% was obtained for UA speech corpus and TORGO database respectively.

Index Terms— Dysarthria, Severity level, Automatic assessment, Audio descriptors, Multi-taper

1. INTRODUCTION

Dysarthria is a motor speech impairment, often characterized by speech that is indiscernible by human listeners. Dysarthria is generally caused by neurological diseases such as amyotropic lateral sclerosis (ALS), Parkinsons disease (PD), cerebral palsy or neurological trauma, manifesting as weakness, paralysis, or a lack of co-ordination of the motor-speech system, resulting in reduction in intelligibility, audibility, naturalness, and efficiency of vocal communication. Assessment of the severity level of dysarthria could be treated as a diagnostic step and is crucial to understand the patients progression in the underlying cause, to take clinical decision regarding the course of therapy or medication as well as to plan speech therapy sessions whenever applicable. Severity assessment is undertaken by a trained speech language pathologist, which turns out to be expensive and inconsistent. On the other hand, an objective severity assessment has the advantages of being cost effective, repeatable and paves way for further automations such as improved speech recognition of dysarthric speech. An understanding of severity has contributed to improved speech recognition of dysarthric speech as seen in [1, 2, 3]. In general, speech intelligibility has been used as an indicator of severity of speech disorders [4]. Automatic intelligibility assessment has been carried out broadly by either (a) Automatic Speech Recognition (ASR) based methods that require reference data as well as linguistic know how [4, 5, 6] or (b) blind intelligibility assessment [7, 8, 9]. In [10], authors discuss the applicability of acoustic and phonological ASR-free features for intelligibility assessment. Authors discuss classification of pathological speech as intelligible or non-intelligible using scores from the fusion of multiple subsystems addressing various aspects of speech such as phonological, intonation etc. in [11]. Literature indicates that research is trending towards moving away from language-specific ASR based methods to language independent automatic intelligibility assessment. While speech quality and intelligibility are closely related, their relationship is not trivial. Frenchay Dysarthria Assessment (FDA) [12] defines several parameters that need to be considered for automatic assessment of the severity level of dysarthria, of which intelligibility is but one. For Parkinson's disease, voice quality symptoms are visible earlier than intelligibility symptoms. Hence it is desirable to assess dysarthria severity level using the speech utterance at the voice quality level in addition to the granular level of articulatory accuracy. In this paper, we propose the applicability of a set of acoustic descriptors that have been used to characterize the timbre of a musical instrument [13]. Timbre is the quality of music or voice that renders each one as distinct. We investigate the use of features suggested in [13] for dysarthria severity classification. Additionally, we compute the acoustic descriptors using a multi-taper based spectral estimation [14] for improved spectral resolution. Significant improvement in severity level assessment was seen using the multi-taper based timbre acoustic descriptors as compared to the work in literature, wherein authors reported 95% classification accuracy using feature fusion on Universal Access (UA) Dysarthric Speech Corpus [9] and in [15] authors reported 93.2% correct classification rate of dysarthria severity levels on TORGO and Nemours database.

The rest of the paper is organized as follows. Section 2 describes the audio descriptors and their role in dysarthric speech severity classification, Section 3 discusses the severity classification methodology and a description of the data used, Section 4 discusses the experimental setup used, Section 5 describes the results and analysis and we conclude in Section 6.

2. AUDIO DESCRIPTORS

In this paper, we use audio descriptors that have been designed for timbre characterization of a musical instrument as a set of features for dysarthric speech severity classification. Timbre is a multidimensional attribute, encompassing a set of auditory descriptors in addition to pitch, loudness, duration, and spatial position [13]. In [13], authors define a set of audio descriptors that can be categorized into *global descriptors*, that are computed across the utterance and *time-varying descriptors*, that are extracted within each frame of the utterance. Audio descriptors are computed various representations of the speech utterance such as (1) Temporal Energy Envelope (2) Short term Fourier transform (STFT) (3) Equivalent rectangular Bandwidth (ERB) based auditory model and (4) Harmonics. For each audio descriptor as shown in Table 1, median and interquartile range have been considered.

Table 1. Audio descriptors used for severity classification

Serial Number	Audio Descriptor	Serial Number	Audio Descriptor
1	Attack	17	Spectral Slope
2	Decay	18	Spectral Decrease
3	Log-Attack time	19	Spectral Rolloff
4	Attack-slope	20	Specto temporal variation
5	Decrease slope	21	Frame energy
6	Temporal Centroid	22	Spectral Flatness
7	Effective Duration	23	Spectral Crest
8	Frequency of Energy Modulation	24	Harmonic Energy
9	Amplitude of Energy Modulation	25	Noise Energy
10	RMS-Energy Envelope	26	Noisiness
11	Autocorrelation-12 coefficients	27	Fundamental Frequency
12	Zero Crossing Rate	28	Inharmonicity
13	Spectral Centroid	29	Tristimulus (3 coefficients)
14	Spectral Spread	30	Harmonic Spectral Deviation
15	Spectral Skewness	31	Odd to Even Harmonic Ratio
16	Spectral Kurtosis		

2.1. Multi-taper spectral estimation

In our work we investigate the usage of multi-taper spectral estimation to compute STFT and Harmonic based features. Conventional spectral estimation of speech uses a Hamming-window or a single taper. Using a single taper windowing results in a significant portion of the signal being discarded and the data points at the extremes being down-weighted, giving a high variance for the direct spectral estimate [16]. Hence, a multi-taper method is used so that the statistical information lost by using just one taper is partially recovered by using multiple windows for the same duration. The multi-taper spectrum is thus a weighted sum of the several tapered periodograms. Spectral estimation of a signal S using multi-taper method is as follows,

$$S(m,k) = \frac{1}{M}\sum_{p=0}^{M-1}\lambda(p)\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{-i2\pi\frac{k}{N}j} \quad (1)$$

where $w_p(j)$ is the p^{th} data taper function, M is the number of tapers and $\lambda(p)$ is the weight corresponding to the p^{th} taper, N is the speech frame length and k is the FFT points. In practice, weights are designed so as to compensate for increased energy loss at higher order tapers.

10 feature sets have been used for dysarthria severity classification is as shown in Table 2.

Table 2. Feature sets used for severity classification

Feature Set	Dimension	Input Representation	Acoustic Descriptors
F1	22	Temporal Energy Envelope	1-10
F2	26	Audio Signal	11-12
F3	22	STFT Magnitude	13-23
F4	22	STFT Power	13-23
F5	22	ERB FFT	13-23
F6	22	ERB Gammatone	13-23
F7	38	Harmonic	15-31
F8	22	Multi-taper Magnitude	13-23
F9	22	Multi-taper Power	13-23
F10	38	Multi-taper Harmonic	15-31

3. SEVERITY CLASSIFICATION

In this paper, Artificial Neural Network (ANN) has been used as a classifier for dysarthria severity classification. The ANN consists of three layers, an input layer, a hidden layer and an output layer. The input layer comprises I nodes equivalent to the dimension of the input feature set being used and the output layer comprises K nodes, the number of classes into which dysarthria severity is being classified. The number of nodes in the hidden layer J is varied based on the dimension of the input feature set being used. ANN configuration is as shown in the Figure 1.

3.1. Data

The proposed technique was validated using two different dysarthric databases i.e., (a) Universal Access (UA) Dysarthric Speech Corpus [17] and (b) TORGO database [18].

3.1.1. UA Dysarthric Speech Corpus

UA speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. The recording material consisted of 455 distinct words with 10 digits, 26 international radio alphabets, 19

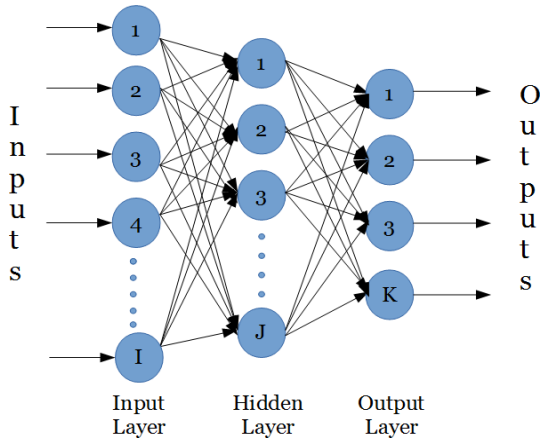

Fig. 1. ANN configuration for severity classification

computer commands, 100 common words and 300 uncommon words that were distributed into three blocks. Three blocks of data were collected for each speaker such that in each block speaker recorded the digits, radio alphabets, computer commands, common words and 100 of the uncommon words. Thus each speaker recorded 765 isolated words. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners is also included in the corpus. Speakers were divided into four different categories based on the intelligibility, namely high, mid, low and very low. We use this information to classify dysarthria severity level.

3.1.2. TORGO

The TORGO database of dysarthric articulation consists of aligned acoustics and measured 3D articulatory features from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). Torgo database consists of 8 dysarthric (DYS) speakers (3 females and 5 males) and 7 non-dysarthric or healthy control (HC) speakers (3 females and 4 males) as a control group. The acoustic data were recorded through two different microphones; an array microphone with 8 recording elements placed at a distance of 61 cm facing the speaker, and a head-mounted microphone. The corpus consists of (1) non-words, (2) Short words such as digits, international radio alphabets, (3) Restricted sentences, (4) Unrestricted sentences. The motor functions of each subject were assessed according to the standardized Frenchay Dysarthria Assessment (FDA) [12] by a speech-language pathologist. FDA measures 28 relevant perceptual dimensions of speech grouped into 8 categories, namely reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility.

The speaker wise severity classification for both UA Speech and TORGO database is as shown in the Table 3. The severity classification for UA speech database is based on intelligibility whereas for TORGO database the overall FDA score for the dysarthric speakers as per [15] is used.

Table 3. Speaker-wise severity distribution for UASPEECH and TORGO database (F** for female speakers, M** for male speakers)

Severity	UA Speech	TORGO
Very Low	F05, M08, M09, M10, M14	F03, F04, M03
Low	F04, M05, M11	F01, M05
Medium	F02,M07, M16	M01, M02, M04
High	F03, M04, M12, M01	—

4. EXPERIMENTAL SETUP

4.1. Data

For the UA Speech corpus, a total of 2812 dysarthric utterances with utterances corresponding to 10 digits and 19 computer commands from block B1 and B2 for training and testing of the classifier has been used.

For the TORGO database, we have used total of 1540 dysarthric utterances for experimentation.

4.2. Multi-taper Spectral Estimation

Multi-taper spectral estimation was done using Discrete Prolate Spheroidal sequences (DPSS) or Thomson or Slepian tapers [14] with 6 orthonormal tapers.

$$w_p(j) = \frac{\sin[\omega_c T(p-j)]}{(p-j)}, \qquad j = 0, 1, \ldots, N-1 \quad (2)$$

where N denotes the desired window length in samples, ω_c is the desired main-lobe cut-off frequency in radians per second, and T is the sampling period in seconds. Twelve dimensional Mel Frequency Cepstral Coefficients (MFCC) features were computed using Thomson multi-taper spectral estimation with a $30\,ms$ window and a $10\,ms$ shift rate.

4.3. ANN configuration

Classification was carried out for 8 different settings of hidden layer neurons. For the hidden layer, number of neurons J or nodes is varied based on the dimension I of the input feature set, and is given as $J = I \star m$, where $m \in \{0.5, 0.66, 0.75, 0.8, 0.83, 1, 1.25, 1.5\}$. The number of output nodes K = 4 and 3 for UA speech and TORGO database respectively. For both UA Speech and TORGO data, 70% of the data was used for training the network, 15% was used for validation and 15% was used for testing.

5. RESULTS AND DISCUSSION

Severity classification was carried out using the experimental setup described in Section 4. It was observed that feature set F1 corresponding to Temporal Energy Envelope performed poorly as compared to the other feature sets. Feature sets STFT magnitude (F3), ERB FFT (F4), ERB Gammatone (F5), Multi-taper Harmonics (F10) performed well for all settings. This could be attributed to the fact that this is a global

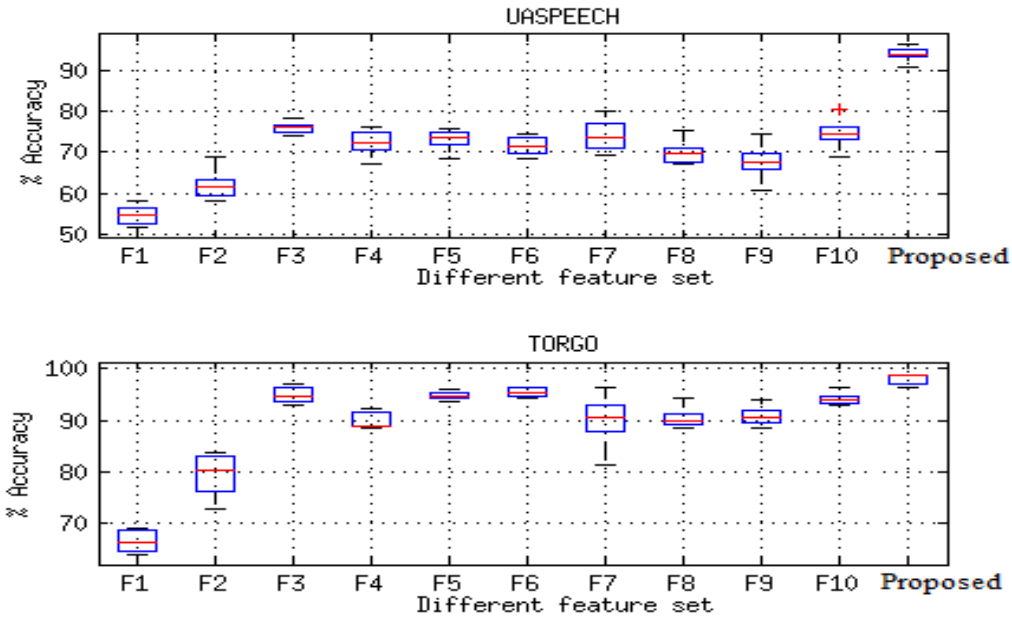

Fig. 2. Feature-wise classification accuracy for varying hidden layer nodes

measure and hence is unable to characterize the severity adequately. Also, for each of the feature sets, similar accuracies were observed across validation and training set, indicating that there is no overfitting or underfitting. The classification accuracy for individual feature sets F1-F10 across different numbers of hidden nodes (varied as discussed in Section 4), is as seen in the Figure 2. Multi-taper spectral estimation out performed the Hamming window based Harmonics audio descriptors (F10) in the severity classification accuracy. This could be attributed to the inherent noise robustness of the multi-taper spectral estimation [19]. We obtained the best classification accuracy when the fusion of all the features from F1-F6 and F10 (Proposed) were used together to give a comprehensive feature of dimension 164. Here we replace the Harmonic timbre feature set F7 with multi-taper based Harmonic feature set F10. Severity wise classification accuracy for the above fusion set is given as in Table 4. For both

Table 4. Severity-wise classification acuracy for UA Speech and TORGO database

Severity	UA Speech	TORGO
Very Low	96.1	99.1
Low	95.1	98.4
Medium	96.7	97.0
High	95.7	—

UA Speech and TORGO database, the overall classification accuracy as well the classification accuracy at feature level outperforms the accuracies cited in recent works [9][15].

6. CONCLUSION

Dysarthria is a motor speech impairment, often characterized by speech that is generally indiscernible by human listeners. Assessment of the severity level of dysarthria is essential for planning therapy, as well as improving automatic dysarthric speech recognition. Objective assessment of severity level or intelligibility of dysarthric speech is essential with reliability, speed and consistency in view. Literature suggests that automatic speech recognition of dysarthric speech can be improved if prior knowledge of severity of dysarthria is available. In this paper, we propose a non-linguistic technique of automatic assessment of severity levels using audio descriptors or a set of features traditionally used to define timbre of musical instruments. Additionally, we use multi-taper based spectral estimation to compute the spectral and harmonic features. An Artificial Neural Network (ANN) was trained to classify speech into various severity levels within Universal Access dysarthric speech corpus and the TORGO database. It was observed that classification accuracies using multi-taper based harmonics was higher than the Hamming window based harmonic features. A fusion of feature sets F1-F6 and F10 (proposed) to give a comprehensive feature set of dimension 164 provided an average classification accuracy of 96.44% for UA speech corpus 98.7% for TORGO database respectively. For both UA Speech and TORGO database, the overall classification accuracy as well the classification accuracy at feature level outperforms the accuracies cited in one of most recent works [9, 15] for these dysarthric speech corpora.

7. REFERENCES

[1] Myung Jong Kim, Joohong Yoo, and Hoirin Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models.," *In Proc. INTERSPEECH 2013*, pp. 3622–3626, 2013.

[2] Siddharth Sehgal and Stuart Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, p. 65, 2015.

[3] Mumtaz Begum Mustafa, Siti Salwah Salim, Noraini Mohamed, Bassam Al-Qatab, and Chng Siong, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," */PLoS/ One. 2014 Jan 23;9(1):e86285 2014.*, 2014.

[4] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425 – 437, 2009.

[5] Doyle P.C., Leeper HA, Kotler AL, Thomas-Stonell N, O'Neill C, Dylke MC, and Rolls K., "Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility," *Journal of rehabilitation research and development*, vol. 34, no. 3, pp. 309–16, 1997.

[6] Philip C Doyle, Herbert A Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, et al., "Dysarthric speech: A comparison of computerized speech recognition and listener intelligility," *Journal of Rehabilitation Research and Development*, vol. 34, no. 3, pp. 309, 1997.

[7] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2825–2828, May 2013.

[8] Myung Jong Kim and Hoirin Kim, "Automatic assessment of dysarthric speech intelligibility based on selected phonetic quality features," *Proceedings of the 13th International Conference on Computers Helping People with Special Needs - Volume Part II*, pp. 447–450, 2012.

[9] Richard Hummel, Wai-Yip Chan, and Tiago H. Falk, "Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech," *In Proc. INTERSPEECH 2011*, pp. 3017–3020, 2011.

[10] Catherine Middag, Tobias Bocklet, Jean-Pierre Martens, and Elmar Nöth, "Combining phonological and acoustic asr-free features for pathological speech intelligibility assessment," *In Proc. INTERSPEECH 2011*, pp. 3005–3008, 2011.

[11] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and S Narayanan, "Intelligibility classification of pathological speech using fusion of multiple subsystems," *In Proc. INTERSPEECH 2012*, pp. 534–537, 2012.

[12] P. Enderby, "Frenchay Dysarthria Assessment," *Int J Lang Commun Disord*, vol. 15, no. 3, pp. 165–173, Dec. 2010.

[13] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting acoustic descriptors from musical signals," *Journal of The Acoustical Society Of America*, vol. 130, pp. 2902–2916, 2011.

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

[15] Kamil Lahcene Kadi, Sid Ahmed Selouani, Bachir Boudraa, and Malika Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233 – 247, 2016.

[16] G. A. Prieto, R. L. Parker, D. J. Thomson, F. L. Vernon, and R. L. Graham, "Reducing the bias of multitaper spectrum estimates," *Geophysical Journal International*, vol. 171, no. 3, pp. 1269–1281, 2007.

[17] 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 Proc. INTERSPEECH 2008*, pp. 1741–1744, 2008.

[18] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, Dec. 2012.

[19] Tomi Kinnunen, Rahim Saeidi, Johan Sandberg, and Maria Hansson Sandsten, "What else is new than the hamming window? Robust MFCCs for speaker recognition via multitapering," *In Proc. INTERSPEECH 2010*, pp. 2734–2737, September 2010.