INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-taper Spectral Estimation

Chitralekha Bhat, Bhavik Vachhani, Sunil Kopparapu

TCS Innovation Labs, Mumbai

bhat.chitralekha@tcs.com, bhavik.vachhani@tcs.com, sunilkumar.kopparapu@tcs.com

Abstract

Dysarthria is a motor speech disorder resulting from impairment in muscles responsible for speech production, often characterized by slurred or slow speech resulting in low intelligibility. With speech based applications such as voice biometrics and personal assistants gaining popularity, automatic recognition of dysarthric speech becomes imperative as a step towards including people with dysarthria into mainstream. In this paper we examine the applicability of voice parameters that are traditionally used for pathological voice classification such as jitter, shimmer, F0 and Noise Harmonic Ratio (NHR) contour in addition to Mel Frequency Cepstral Coefficients (MFCC) for dysarthric speech recognition. Additionally, we show that multi-taper spectral estimation for computing MFCC improves the unseen dysarthric speech recognition. A Deep neural network (DNN) - hidden Markov model (HMM) recognition system fared better than a Gaussian Mixture Model (GMM) - HMM based system for dysarthric speech recognition. We propose a method to optimally use incremental dysarthric data to improve dysarthric speech recognition for an ASR with DNN-HMM. All evaluations were done on Universal Access Speech Corpus.

Index Terms: dysarthria, speaker adaptation, fMLLR, jitter, shimmer, deep neural network

1. Introduction

Dysarthria is a motor speech disorder resulting from impairment in muscles responsible for speech production. Neurological injury to the nervous system may result in 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. For dysarthric speakers, speech is a more efficient/convenient mode of communication with electronic devices as compared to keyboard input [1]. Voice or speech as a computer interface for dysarthric speakers, was implemented as early as 1985 [2]. Authors designed an assistive device to bypass the keyboard and activate the computer using voice control. Despite the early start, automatic recognition of dysarthric speech is poorer as compared to that of normal speech, owing to the inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data. As per the literature, the work so far can be broadly classified into two types of research - (1) improving intelligibility by modifying or enhancing the dysarthric speech and (2) ASR based speech recognition by speaker adaptation. In [3], authors study the effect that certain modifications have on the intelligibility of dysarthric speech and report that by transforming the dysarthric speech at the short-term spectral levels, an increase in intelligibility was attained. In the study [4], authors have achieved increased intelligibility by transforming the vowels of a dysarthric speaker to more closely match the vowel space of a normal speaker. Features that provided optimum performance were vowel duration and F1 - F3 (formant 1 - formant 3) stable points that were computed using shape-constrained isotonic regression. In another study [5], the author transforms various aspects of speech such as the correction of pronunciation errors, adjustment of the tempo and the frequency characteristics of speech to obtain increased intelligibility. Yet another technique to increase both the perceptual quality of the speech as well as intelligibility is transformations to formant trajectories of dysarthric speech, to closely match that of a normal speaker [6].

In [7], one of the earlier works in ASR based dysarthric speech recognition, authors stress on the data insufficiency challenge and define confusability and consistency measures to predict recognizer performance. Several works [8, 9] discuss the merits of selection of ASR type namely - speaker independent (SI) , speaker dependent (SD) or speaker adapted (SA) by analysing the correlation between the severity of dysarthria and best performing ASR type (one of SA or SD). In [10], authors have used a method of measuring similarity between dysarthric speakers and select only the most similar speaker data for training rather the SI acoustic models, followed by maximum a posteriori (MAP) adaptation. Studies [11] also suggest an improvement in recognition by using more suitable prior model or background model, for adaptation based on the dysarthric speaker's acoustic characteristics. Work pertaining to speaker based lexical or pronunciation model adaptation in addition to acoustic model adaptation, [12, 13, 14] have shown improvement in the ASR performance. An understanding of the speech production process through the articulatory models for speech has proven beneficial in improved accuracy of the ASR, both conventional GMM-HMM and DNN-HMM [1, 15, 16]. More recently application of neural network topologies [17, 18], feature space maximum likelihood linear regression (fMLLR) transformation [16] and a hybrid adaptation using maximum likelihood linear regression (MLLR) and MAP [19] have been used to improve dysarthric speech recognition.

We believe that speech based applications such as voice biometric, personal assistants can immensely benefit dysarthric speakers if designed well. Given the challenges in collecting dysarthric data, the thrust is now on recognition of unseen speech utterances, i.e. recognition of dysarthric speech that is not a part of the training set. In this paper, we propose a method and examine a set of features to improve speech recognition of unseen dysarthric speech. We incorporate multi-taper MFCC (MT-MFCC) which has been proven to be effective in speaker verification and speech recognition [20, 21] as well as voice disorder classification [22]. Additionally, we examine the voice

http://dx.doi.org/10.21437/Interspeech.2016-1085

parameters (VP) such as jitter, shimmer, F0 features and noise-to-harmonics ratio (NHR), that have traditionally been used for voice disorder classification [23, 24]. Some of these parameters have been used to automatically assess the severity level of dysarthria [25]. The main contribution of this paper is a framework for unseen dysarthric speech recognition, using a DNN-HMM SA-ASR system along with the use of a combination of speaker specific features. To the best of our knowledge no other work has examined the usefulness of voice parameters such as jitter, shimmer F0 features and noise-to-harmonics ratio (NHR) in the context of dysarthric speech recognition.

The rest of the paper is organized as follows. Section 2 describes the features and their role in dysarthric speech recognition, Section 3 discusses the various experimental setups and a description of the data used, Section 4 describes the results and analysis and we conclude in Section 5.

2. Features for dysarthric speech recognition

2.1. Multi-taper spectral estimation

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 [26]. 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.

2.2. Jitter and Shimmer

Jitter and shimmer are characteristic to the speech of an individual and have been beneficial in speaker recognition tasks [27]. *Jitter* represents the perturbations that occur in the fundamental frequency F0 and can be interpreted as a modulation of the periodicity of the voice signal. Reduced control of vocal fold vibration, as is the case in dysarthria manifests as jitter. Pathological voices are generally characterized by a high degree of jitter and perceived as hoarse. Hence, an estimation of jitter has been used in classification of pathological speech. Absolute jitter is computed as per Equation 2.

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \quad (2)$$

where $T_i = 1\ /\ F0$ and N is the number of $F0$ periods.

Shimmer pertains to the amplitude variation of the sound wave and varies with the glottal resistance and mass lesions in the vocal folds manifesting as presence of noise emission and breathiness in the voice [24]. Absolute shimmer is computed as per Equation 3 and is expressed in decibels (dB).

$$Shimmer(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * log\left(\frac{A_{i+1}}{A_i}\right)\right| \quad (3)$$

where A_i is the extracted peak-to-peak amplitude and N is the number of $F0$ periods.

2.3. F0 features

The role of fundamental frequency $F0$ in the intelligibility of speech has been studied for both normal and dysarthric speech [28]. These studies suggest that a higher variation in $F0$ contributes significantly to increased intelligibility. However, for dysarthric speakers, the precision and flexibility of the vocal folds, articulators and other speech subsystems are lower, leading to reduced prosodic control, reflecting as a reduction in intelligibility. Additionally, studies show that the slower articulatory rate tends to be associated with low values of mean, maximum and variations of $F0$ [29]. $F0$ measurements such as mean and variation are also indicative of the vocal loudness of speech, which has a bearing on speech intelligibility.

2.4. Noise to Harmonic ratio (NHR)

Noise-to-Harmonics ratio (NHR) is indicative of the abnormal vibratory characteristics of the vocal folds, manifesting as hoarseness in dysarthric speech. NHR is measured in dB, calculated by the ratio of noise energy or the aperiodic part of a sustained vowel to the energy of the periodic part. NHR can be used as a measure of voice quality and is defined as below.

$$NHR(dB) = 10 * log\left(\frac{E_n}{E_p}\right) \quad (4)$$

where E_p is the energy of the periodic part and E_n is the energy of the noise . NHR has been used as one of discriminative features to evaluate the degree or severity of dysarthria in [25].

3. Speech recognition methodology

3.1. Data

Data from Universal Access (UA) speech corpus [30] was used for both training and testing of the two ASR systems discussed in this section. 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 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. Data from all channels was used for the purpose of this work. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners is also included in the corpus. We use this information to analyse the performance of our recognition systems at dysarthria severity level.

The objective of this work is to recognize unseen dysarthric data and explore the applicability of voice parameters in recognition of dysarthric speech. The training and testing corpus as

described in Table 1 allows us to compare and contrast the performance of our recognition systems for seen and unseen testing data, i.e. DYS-computer command words (DYS-CC words).

Purpose	Data	Number of Utterances
Training	HC-digits	800
	HC-computer command words	1500
	DYS - digits	800
Testing	HC-digits	110
	HC - computer command words	229
	DYS - digits	169
	DYS - computer command words	361

Table 1: Training and testing corpus

3.2. ASR systems and experimental setup

3.2.1. Feature extraction and normalization

Multi-taper spectral estimation was done using Discrete Prolate Spheroidal sequences (DPSS) or Thomson or Slepian tapers [31] with 6 orthonormal tapers.

$$w_p(j) = \frac{\sin[\omega_c T(p-j)]}{(p-j)}, \qquad j = 0, 1, \ldots, N-1 \quad (5)$$

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 MFCC features were computed using Thomson multi-taper spectral estimation with a $30\,ms$ window and a $10\,ms$ shift rate.

All the *voice parameters* such as the jitter, shimmer, $F0$, and NHR measures were computed using the voice analysis software, 'Praat'[32], wherein a cross-correlation (cc) method was used for acoustic periodicity estimation, using a $30\,ms$ window and a $10\,ms$ shift rate. 'Praat'gives various measurements for each of the above voice parameter. Based on experimental evidence and literature [27], features as shown in Table 2 were chosen for speech recognition.

Feature	Praat Measurement
Jitter	Jitter(local, relative)
Shimmer	Shimmer(local, dB)
Fundamental Frequency F0	Standard Deviation Range (Maximum - Minimum)
Noise to Harmonic ratio	Standard Deviation Mean

Table 2: Voice parameters extracted from Praat

We have three sets of features namely, MFCC, multi-taper MFCC (MT-MFCC) and voice parameters (VP).

3.2.2. Speech recognition

We use Kaldi toolkit [33] for both GMM-HMM based and DNN-HMM based dysarthric speech recognition. A 3-state HMM with a monophone or a triphone context model is used. GMM-HMM system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean normalization (CMN) was applied on each of the above sets of features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use a context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 ($40 \times 9 frames$) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the data). We used 2 hidden layers with 512 nodes in each layer. Trigram language model was used and performance of each of the recognition systems is reported in terms of word error rate (WER).

We explore the use of our feature sets - MFCC, MT-MFCC and MT-MFCC-VP for speech recognition with speaker adaptation(SA).

3.2.3. Speaker Adaptation

Traditionally speaker adaptation techniques such as MLLR, MAP are applied on SI acoustic models at the time of decoding. We use Maximum Likelihood Linear Transform (MLLT) for speaker normalization. MLLT derives a unique transformation for each speaker using the reduced feature space from the LDA.

An inter-speaker feature space normalization technique known as feature space maximum likelihood linear regression (fMLLR) [34] is performed for each speaker, wherein the acoustically transformed feature vector $\hat{o}(t)$ is estimated using a transformation matrix A and a bias vector b as $\hat{o}(t) = Ao(t) + b$, where $\hat{o}(t)$ is obtained by transforming the the input feature vector $o(t)$ at frame t.

Speaker adaptive training (SAT) [35] is applied at the time of training the acoustic models and aims at eliminating the inter-speaker variation. fMLLR based SAT was applied to create speaker adapted (SA) acoustic models; further, fMLLR was applied on the features of the input utterances at the time of decoding. SAT using fMLLR remain common to both GMM-HMM and DNN-HMM based systems.

3.2.4. Incremental training of DNN

Considering the application of DNN-HMM based speech recognizer for unseen dysarthric speech, it is expected that there will be incremental data as a dysarthric user uses the system. This data can be used to improve upon the existing acoustic models and thereby improve the performance of the recognition engine. Two mechanisms of training the DNN-HMM were considered - (1) DNN weights built using original corpus, are updated by re-training, using the incremental data alone. (2) System is trained on the entire data (original + incremental).

4. Results and Discussion

Speech recognition using the GMM-HMM system as well as the DNN-HMM system was carried out using a set of features, namely MFCC, MT-MFCC and VP, individually as well as in fusion. Training and testing data set up was designed so as to understand the speech recognition performance for a set of words for which no training has been done on dysarthric data such as dysarthric computer command words (DYS-CC). The train and test for all other cases such as healthy control (HC) and DYS-digits are disjoint or mutually exclusive. The word error rates (WER) for Triphone GMM-HMM and DNN-HMM

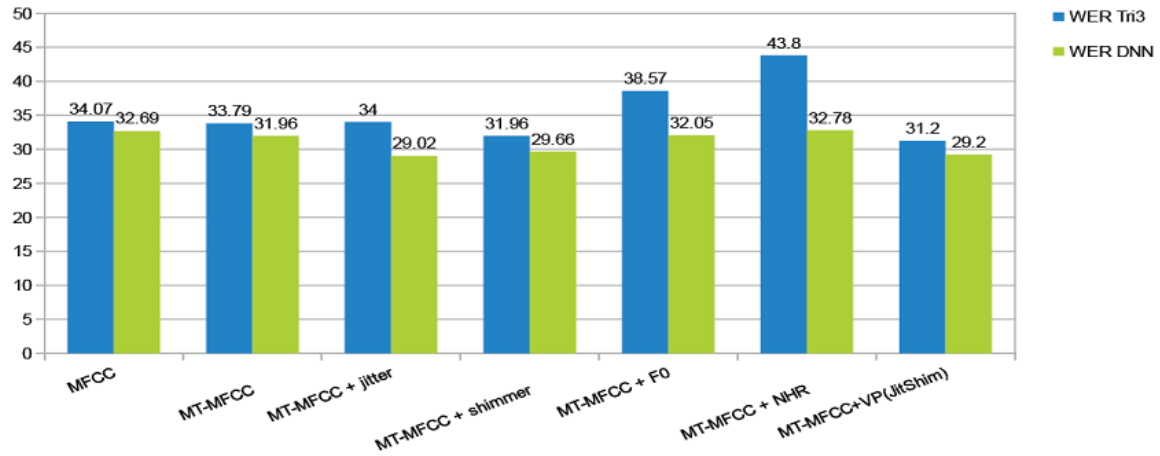

Figure 1: WER for GMM-HMM based and DNN-HMM based recognition using speaker adaptation

systems are as shown in Figure 1.

It can be seen that using MT-MFCC and VP - jitter and shimmer with speaker adaptation, showed a reduction in WER for the DNN-HMM system, whereas adding F0 features and NHR features had an adverse impact. It has been observed that jitter and shimmer are not discernible perceptually by human listeners [36], whereas any difference in fundamental frequency F0 or NHR are perceptually apparent [28]. Using CMN and SAT improved the speech recognition using MT-MFCC-F0 features. However, using F0 features in addition to MT-MFCC did not improve the overall speech recognition for any of the SA or SI systems. It was seen that, features that have a clear bearing on speech perception adversely impacted the performance of the recognizers.

Fusion of MT-MFCC, jitter and shimmer (VPJitShim) features shows a relative improvement of 8.4% in GMM-HMM based system and 10.7% in DNN-HMM based system over the MFCC features alone.

Table 3 shows the recognition results based on speaker type for the DNN-HMM using MT-MFCC-VPJitShim feature set. This indicates a correlation between severity of dysarthria and the accuracy of the recognition system. Similar trend was seen for MT-MFCC, MT-MFCC-Jitter and MT-MFCC-Shimmer feature based recognition systems, wherein the WER increased with the increase in severity of dysarthria.

Speaker type	%Accuracy-DNN-HMM **Initial**		%Accuracy-DNN-HMM **Incremental**	
	Digits	**CC words**	**Digits**	**CC words**
Healthy control	98.93	99.4	94.9	99
DYS Very Low	94.66	91.47	94.7	98.66
DYS Low	92.68	36.84	83.1	95.35
DYS Medium	88.24	31.51	82.34	90.09
DYS High	52.38	7.22	51.56	93.65

Table 3: Dysarthria severity wise accuracy for DNN-HMM system with original training data and incremental training data for MT-MFCC-VPJitShim

Experiments pertaining to incremental training were conducted for SA based DNN-HMM recognizer, using fusion MT-MFCC and VPJitShim features. The DNN-HMM system was retrained, using the initial weights from the training data mentioned in Table 1 and a 10% additional DYS-CC word data. This system performed poorly in comparison to the system trained with original training data. This could be attributed to the updating of the neural network to a specific type of data, namely dysarthric CC word data. As expected, training the DNN-HMM system using the entire data (original data + incremental data) provided a significant improvement, especially in the recognition of DYS-CC words for dysarthric speakers, as shown in Table 3. Recognition of digits deteriorated for both healthy control and dysarthric data, owing to higher number of digits being incorrectly recognised as CC words, especially the confusable pairs like the digit 'nine' and the CC word 'line'.

5. Conclusions

In this paper, we propose a method and examine a set of features to improve speech recognition of unseen dysarthric speech. We incorporate multi-taper MFCC (MT-MFCC) and examine the applicability of voice parameters (VP) such as jitter, shimmer, F0 features and noise-to-harmonics ratio (NHR) in two types of recognition systems namely - GMM-HMM and DNN-HMM using speaker adaptation approach. For the MT-MFCC-VP(JitShim) fused feature set, a relative improvement of 8.4% in GMM-HMM based system and 10.7% in DNN-HMM based system was seen over the MFCC features alone. This indicates that while using jitter and shimmer voice parameters was beneficial in speaker adaptation based speech recognition, using F0 and NHR features added no advantage. This difference in behaviour of both the recognition systems could be understood from the perspective of human listener perception of dysarthric speech. It has been observed that jitter and shimmer are not discernible perceptually by human listeners, whereas any difference in fundamental frequency F0 or NHR are perceptually apparent. An increment in the training data clearly increased the recognition accuracy of the DNN-HMM based system using MT-MFCC-VPJitShim features, for DYS-CC words. Our future work would involve further improving the accuracy of the dysarthric speech recognition under the DNN-HMM architecture, exploring different topologies and network types that would suit best for dysarthric speech recognition.

6. References

[1] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. NIPS 2010, – Workshop on Machine Learning for Assistive Technologies at the 24th annual conference on Neural Information Processing Systems*, 2010, pp. 70–78.

[2] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people," in *Archives of physical medicine and rehabilitation*, vol. 66, no. 10, 1985, pp. 678–81.

[3] J. P. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *In Proc. ICASSP*, April 2003, pp. I–924–I–927 vol.1.

[4] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743 – 759, 2007.

[5] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163 – 1177, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[6] A. Kain, X. Niu, J.-P. Hosom, Q. Miao, and J. P. H. van Santen, "Formant re-synthesis of dysarthric speech," in *ISCA Speech Synthesis Workshop*, 2004.

[7] P. D. Green, J. Carmichael, A. Hatzis, P. Enderby, M. S. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers," in *In Proc. INTERSPEECH*, 2003, pp. 1189–1192.

[8] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[9] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models." in *In Proc. INTERSPEECH*, 2013, pp. 3622–3626.

[10] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 254–259.

[11] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, vol. 27, no. 6, pp. 1147 – 1162, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[12] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–14, 2009.

[13] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *In Proc. ICASSP*, May 2011, pp. 4924–4927.

[14] C.-H. Wu, H.-Y. Su, and H.-P. Shen, "Articulation-disordered speech recognition using speaker-adaptive acoustic models and personalized articulation patterns," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 10, no. 2, p. 7, 2011.

[15] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 947–960, 2011.

[16] S. Hahm, D. Heitzman, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, p. 47.

[17] S. R. Shahamiri and S. S. B. Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Advanced Engineering Informatics*, vol. 28, no. 1, pp. 102 – 110, 2014.

[18] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *12th International Conference on Signal Processing (ICSP)*, Oct 2014, pp. 505–509.

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

[20] M. J. Alam, P. Kenny, and T. Stafylakis, "Combining amplitude and phase-based features for speaker verification with short duration utterances," in *In Proc. INTERSPEECH*, 2015, pp. 249–253.

[21] M. J. Alam, P. Kenny, and D. O'Shaughnessy, *A Study of Low-variance Multi-taper Features for Distributed Speech Recognition*, ser. NOLISP'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 239–245.

[22] Ö. Eskidere and A. Gürhanlı, "Voice disorder classification based on multitaper mel frequency cepstral coefficients features," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[23] 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.

[24] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112 – 1122, 2013, CENTERIS 2013 / ProjMAN 2013 / HCIST 2013.

[25] K. Kadi, S. Selouani, B. Boudraa, and M. Boudraa, "Discriminative prosodic features to assess the dysarthria severity levels," in *Proceedings of the World Congress on Engineering*, vol. 3, 2013.

[26] 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.

[27] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition." in *In Proc. INTERSPEECH*, 2007, pp. 778–781.

[28] R. Patel and P. Campellone, "Acoustic and perceptual cues to contrastive stress in dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 1, pp. 206–222, 2009.

[29] K. Tjaden and G. Wilding, "The impact of rate reduction and increased loudness on fundamental frequency characteristics in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 178–186, 2011.

[30] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in *In Proc. INTERSPEECH*, 2008, pp. 1741–1744.

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

[32] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[34] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[35] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *in Proc. ICSLP*, 1996, pp. 1137–1140.

[36] J. Kreiman and B. R. Gerratt, "Jitter, shimmer, and noise in pathological voice quality perception," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.