Automatic differentiation between normal and disordered speech

Jan Felcyn*+

+ Adam Mickiewicz University, Institute of Acoustics
85 Umultowska Street
61-614, Poznań, Poland
janaku@amu.edu.pl

Agnieszka Będkowska Cavalcante*

* Gido Labs sp. z o. o.
1 Roman May Street
61-371, Poznań, Poland
agnieszka.betkowska@gidolabs.eu

Abstract

One of the most important elements of everyday life is communication. The most natural way of it is speech. Unfortunately, there are many diseases which disturb fluency and intelligibility of speech. Such disorders often lead to emotional and psychological problems related to social interactions. Early diagnosis is crucial to detect and minimalize results of such disorders. To make this process easier, several automatic algorithms have been proposed by scientists. In this paper, we investigated two methods: Envelope Modulation Spectra (EMS) and Multidirectional Regression (MDR). We applied both techniques to Polish language and evaluated their performance on distinguishing Polish speakers with and without speech disorders. Our experiments showed that each method is efficient in such a discrimination task. Among all 48 EMS characteristics 45 differed significantly both groups of speakers. MDR recognized disordered speech with almost 99% accuracy for several words.

Keywords: dysarthria, automatic classification, speech disorders, Polish speech

1. Introduction

Verbal communication is one of the most important aspects of daily life. Unfortunately, many people suffer from diseases that degrade their speech abilities. Inability to speak properly leads to impoverished interactions with the society and can be the reason of social exclusion (Creer et al., 2013; Iida & Campbell, 2003).

A dysarthria is one of the most common speech disorders and is characterized by the dysfunction of muscles used in speech production process. Kain et al. (2007) describe it as "an impairment in one or more of the processes of speech production: respiration, phonation, resonance, articulation, and prosody". There are several types of dysarthria: spastic, hyperkinetic, hypokinetic, ataxic, flaccid and mixed, among others (McCaffrey, n.d.) and each of them is characterized by different symptoms (Kain et al., 2007; McCaffrey, n.d.).

Dysarthric speech is not only less intelligible but also 10-17 times slower than normal speech (Rudzicz, 2013). Hence several attempts to develop techniques improving the verbal communication of affected people have been conducted. One of the existing approaches is supplementation: topic, alphabetic or combined (Hustad et al., 2003). A speaker points out the topic of its speech or first letter of the word he/she would use – or both of them. This method requires the user to use additional equipment – such as keyboard or pointer. In case person suffers also from physical disabilities, the interaction with keyboard could be even 300 times slower than for healthy people (Rudzicz, 2013), which limits the applicability of the method. Another approaches focus on enhancing the speech signal quality, which can improve communication between humans and also between human and computer (Rudzicz, 2013).

For people with reduced mobility, the second approach is of the great interest. The possibility of using voice to control devices (such as TV, mobile phones, computers, lamps) can ease the daily life significantly (Parker et al. 2006). Unfortunately, diversity of disorders in dysarthria is so wide that standard automatic speech recognition (ASR) tools fail (Caballero-Morales & Trujillo-Romero, 2014) in such tasks.

One possible solution could be to recognize the type of speech disorders and based on that choose the algorithm which gives the best results for this particular class of disorders. There are several methods that performs speech disorders differentiation (Lansford & Liss, 2014; Rosen et al., 2010; Sapir et al., 2010; Vogel et al. 2011). However, many of them require time-consuming preparation of stimuli by hand. For our purposes, we are only interested into automatic solutions, hence we investigated two techniques that do not require human assistance.

First of them is called Envelope Modulation Spectra (EMS) and was developed by Liss, LeGendre & Lotto (2010). It measures signal's features related to amplitude envelope and give 100% of recognition between normal and dysarthric speech.

The second algorithm applies multidirectional regression (MDR) and was invented by (Muhammad et al., 2012). It gives 99% accuracy in recognizing normal and disordered speech for arabic digits. Please note that patients suffer from nodules, cysts, polips etc., not from dysarthria.

In next sections we will present implementation of these two methods for Polish language.

2. Methods

2.1 Envelope Modulation Spectra (EMS)

The signal is filtered for 7 octave bands with central frequencies ranging from 125 Hz to 8 kHz. For every band (as well as for whole signal) modulation envelope is extracted (half-rectification and lowpass filter with cutoff frequency of 30 Hz, downsampling to fs=80Hz). Then, 512-point FFT is applied (with Tukey window) and the

Characteristics	Description	Abbreviations
Peak frequency	The frequency of the peak in the spectrum with the highest amplitude. The period of this frequency is the duration of the predominant repeating amplitude pattern	PF_whole_signal, PF_125Hz, PF_250Hz, PF_500Hz, PF_1000Hz, PF_2000Hz, PF_4000Hz, PF_8000Hz
Peak amplitude	The amplitude of the peak described above (divided by overall amplitude of the energy in the spectrum). This is a measure of how much the rhythm is dominated by a single frequency	PA_whole_signal, PA_125Hz, PA_250Hz, PA_500Hz, PA_1000Hz, PA_2000Hz, PA_4000Hz, PA_8000Hz
E3-6	Energy in the region of 3–6 Hz (divided by overall amplitude of spectrum). This is roughly the region of the spectrum, around 4 Hz, that has been correlated with intelligibility and inversely correlated with segmental deletions	E3_6_whole_signal, E3_6_125Hz, E3_6_250Hz, E3_6_500Hz, E3_6_1000Hz, E3_6_2000Hz, E3_6_4000Hz, E3_6_8000Hz
Below4	Energy in spectrum from 0–4 Hz (divided by overall amplitude of spectrum). The spectrum was split at 4 Hz, because pilot work demonstrated that the amount of energy below and above 4 Hz was relatively uncorrelated across a variety of speakers and sentences	B4_whole_signal, B4_125Hz, B4_250Hz, B4_500Hz, B4_1000Hz, B4_2000Hz, B4_4000Hz, B4_8000Hz
Above4	Energy in spectrum from 4–10 Hz (divided by overall amplitude of spectrum).	A4_whole_signal, A4_125Hz, A4_250Hz, A4_500Hz, A4_1000Hz, A4_2000Hz, A4_4000Hz, A4_8000Hz
Ratio4	Below4/Above4	R4_whole_signal, R4_125Hz, R4_250Hz, R4_500Hz, R4_1000Hz, R4_2000Hz, R4_4000Hz, R4_8000Hz

Table 1. EMS characteristics, their descriptions and abbreviations used in this paper. Descriptions are taken from (Liss et al., 2010)

spectrum is converted to decibels for frequencies up to 10Hz. Based on six EMS characteristics (see Table 1) calculated for each obtained spectrum and for the whole signal, 48 dimensional feature vector is created. All vectors are then divided into two groups representing speakers with and without disorders. T-Student test is launched to find out which of the features are the most significant to differentiate between these two groups.

2.2 Multidirectional Regression (MDR)

Each signal is divided into 20ms-length frames using Hamming window (with overlapping of 10ms) and FFT is applied to each frame. The obtained spectrum is then filtered with 24 triangular mel filters (according to the equation 1:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \quad m = 1,2,\dots,24, \qquad (1)$$

where *f* is the center frequency of the filter.

After filtration, all values in the frames of one signal are converted to the logarithmic scale. After this step, a matrix called TF Pattern is obtained – with 24 rows and T columns (where T is equal to number of frames for the signal). The element in i-th row and j-th column is denoted as c_{ij} and presents value obtained for i-th mel filter and j-th frame. Then, 3-point linear regression is applied to the matrix, in four directions: along time, along frequency and along TF pattern at 45° and 135° (see equations 2-5).

$$d_{t,n}^{time} = \frac{\sum_{k=1}^{3} k\left(c_{t+k,n} - c_{t-k,n}\right)}{2\sum_{k=1}^{3} k^2} \qquad (2)$$

$$d_{t,n}^{freq} = \frac{\sum_{k=1}^{3} k\left(c_{t,n+k} - c_{t,n-k}\right)}{2\sum_{k=1}^{3} k^2} \qquad (3)$$

$$d_{t,n}^{45} = \frac{\sum_{k=1}^{3} k\left(c_{t-k,n-k} - c_{t+k,n+k}\right)}{2\sum_{k=1}^{3} k^2} \qquad (4)$$

$$d_{t,n}^{135} = \frac{\sum_{k=1}^{3} k\left(c_{t+k,n-k} - c_{t-k,n+k}\right)}{2\sum_{k=1}^{3} k^2}, \qquad (5)$$

where:
t takes values from 1 to T for the analyzed signal; ***n*** takes values from 1 to 24 (equal to the number of mel filters); ***c*** is taken from TF pattern matrix (addressed with appropriate indices).

Every regression is then transformed using discrete cosine transformation (DCT) according to the equation 6:

$$dct_{t,l} = \sum_{p=1}^{24} d_{t,n} \cos\left[(p-0{,}5)\frac{l\pi}{24}\right], \qquad (6)$$

$$l = 1,2,\dots,12$$

Finally, we get a 48-elemental MDR vector as follows (equation 7):

$$MDR_t = \left[dct_t^{time},\ dct_t^{freq},\ dct_t^{45},\ dct_t^{135}\right], \qquad (7)$$

These vectors are an input for training GMMs to discriminate between normal and disordered speech.

3. Experiments

3.1 Implementation

EMS algorithm was implemented using MATLAB environment. Procedure generated a matrix consisted of 48 columns (each column represented one feature from EMS characteristics) and number of rows equal to the number of recordings. The obtained matrix was used in the statistical analysis.

MDR algorithm was implemented using Python language and features vector for each recording was calculated. Two GMM models (representing speech with and without disorders) were trained and tested using HTK Toolkit (Young et al., 1997).

3.2 Testing EMS method

3.2.1 Recordings

The database used for the evaluation contains recordings from nine speakers. Four speakers (two men and two women) are healthy subjects and five speakers (four men and one woman) have speech disorders, classified by phonologist as mild, moderate and severe dysarthria, dyslalia and articulation disorders. Every speaker recorded at least 10 times numbers from 1 to 10 and 10 polite requests. Please note, number of recordings between speakers was not equal. In total, the database consisted of 2002 recordings.

Each recording was saved in .wav format – with different sample rates (according to the devices used by speakers) but with the same bitrate of 16 kbit. However, different sample rates were not the problem because before any test recordings were downsampled to the sampling frequency of 80Hz (see Section 2.1). Due to the fact, that our subjects recorded themselves on their own, some of the recordings appeared to be overloaded. A Student t-test for independent samples was applied to verify whether there were statistically significant differences between the groups of overloaded and non-overloaded recordings

Indeed, for 40 from 48 EMS characteristics significant differences were observed. That is why overloaded signals were excluded from the statistical analysis. Finally, 1725 recordings were used.

3.2.2 Procedure

Features vectors were automatically computed and saved in Excel file using algorithm implemented in MATLAB. Results were divided into two groups – samples of speech with and without voice disorders. Student t-test for independent samples was launched to find out if any of 48 characteristics differ significantly between groups. Additionally, effect size (ES) was computed.

3.2.3 Results and discussion

Only three (from 48) characteristics do not differ groups significantly. These are: PA_125Hz (with p=0,778), R4_whole_signal (p=0,375) and R4_250Hz (p=0,134). Every other characteristics has p<0,05, so it differs groups significantly. More information is given by the ES. The largest ES value – the largest difference between groups. In the table 2 we present 10 EMS characteristics with the largest ES-s (ES>0,9) as well as all three characteristics which were not significantly different between groups.

Characteristics	t	p	Effect Size (ES)
PA_2000Hz	-23,240	,000	1,228
B4_2000Hz	-22,961	,000	1,213
PA_1000Hz	-20,191	,000	1,067
B4_whole_signal	-18,358	,000	0,970
B4_125Hz	-18,346	,000	0,969
B4_1000Hz	-18,327	,000	0,968
PA_whole_signal	-18,307	,000	0,967
E3_6_125Hz	-17,782	,000	0,939
A4_125Hz	-17,734	,000	0,937
B4_4000Hz	-17,159	,000	0,906
R4_250Hz	-1,499	,134	---
R4_whole_signal	,886	,375	---
PA_125Hz	,282	,778	---

Table 2. Ten EMS characteristics with the largest effect sizes and all three characteristics which were not statistically significant. T-test was launched between groups of normal and disordered speakers.

One can notice that seven characteristics are related to energy and three to peak amplitude – no characteristics of peak frequency. 125 Hz is the most common center frequency for the octave band while 250, 500 and 8000 Hz do not appear in table 2. We can make an assumption that the most interesting information is carried in the lowest and moderate octave bands and is related to energy and its amplitude rather than frequency. Our next algorithm will focus on measuring energy more detailed (using not octave but 1/3 octave bands) and trace changes in energy in time (using short time frames instead of full-time signal). It should give us more information about differences between speakers with and without disorders making the differentiation procedure more reliable and unambiguous.

3.3 Testing MDR method

3.3.1 Recordings

Initially, recordings were identical as in the EMS method. However, pre-test results were not satisfying. We noticed a disproportion between number of recordings of healthy and unhealthy speakers. That is why we added 3 more people with normal speech (two men and one woman), having finally 2444 recordings – from seven healthy and five unhealthy speakers.

All recordings were stored as mono .wav files with sample rate of 16 kHz and bitrate equal to 16 kbit.

3.3.2 Procedure

Recordings were divided into two groups: speech of people with and without disorders. For each frame of each recording, 48-elemental features vector was computed based on MDR algorithm. Using HTK Toolkit, two GMM models were trained – one for normal_speech and one for disordered_speech. For each class different k values (k is the number of mixtures used in GMM model) were used (k=1, 2, 4, 8, 16, 32, 64, 128, 256 and 512).

Experiment was launched using leave-one-out method. Recordings of one speaker were used as test set, and the remaining samples of other speakers were used as training material for GMM models. The procedure was repeated for each speaker, and the evaluation results were averaged.

3.3.3 Results and discussion

The discrimination procedure was not satisfactory for three speakers: speaker no. 5, 7 and 10. After analyzing the recordings, we found out that the problem is related to the quality of their recordings. Speaker 5 breathed loudly and strong ‘windy pops’ could be perceived in almost each of his recordings. Speakers 7 and 10 recorded themselves being away from a microphone – reverberation of a room is audible. Every other speaker recorded itself close to the microphone, so we find the reverberation could be the problem. Hence, we excluded those 3 speakers from further analyses.

Mean for all other speakers for all words was 96,4% (for GMM with k=4). In Table 3 we present global results of algorithm performance, including hits, misses, false alarms and correct rejections. The table answers the question ‘Did the subject have normal speech?’. The probability of hits was 98% and of correct rejections – 93,2%.

		Algorithm Performance	
		Normal	**Disordered**
Subjects	**Normal**	Hit = 98%	Miss = 2%
	Disordered	False alarm = 6,8%	Correct Rejection = 93,2%

Table 3. Results of algorithm performance regarding the question "Did the speaker have normal speech?"

We also studied every word separately. The best results were obtained for digits 4 and 6, for GMM with k=16 in both cases. For digit 4 recognition was 98,9%, for digit 6 – 99,5% (see Table 4).

Based on the results, we conclude that the choice of the spoken utterance can have influence on discrimination performance. In case of numbers four and six, we assume that good recognition rate is related to the presence of fricatives – disordered speakers distorted their pronunciation. Another thing is the presence of stop consonant ‘t’ in the Polish digit 4. People with speech disorders tend to omit stop consonants and do not speak them (Rudzicz, 2013).

Digit	**Speakers**				
	S2	**S3**	**S4**	**S8**	**S9**
4	100,0	100,0	100,0	100,0	100,0
6	100,0	100,0	95,5	100,0	100,0
	S13	**S14**	**S15**	**S16**	**Mean**
4	100,0	90,0	100,0	100,0	**98,9**
6	100,0	100,0	100,0	100,0	**99,5**

Table 4. Accuracy in % of GMM with k=16 in recognition of both (normal and disordered speakers) for Polish digits 4 and 6.

4. Conclusion and future work

Experiments showed that automatic differentiation between people with and without speech disorders is possible and effective. Only 3 EMS characteristics (from total of 48) did not differ groups significantly. In the MDR method the probability of >98% of good assignment could be achieved using specific words.

On the other hand, it was shown how important is careful preparation of recordings’ database. EMS method – because of its nature, basing on amplitude envelope – was very sensitive for overloads in recordings. MDR algorithm gave unsatisfying results for recordings with pops – as well as for those ones where people stood too far from a microphone. But not only technical aspects are important, words used by people are also a crucial factor. Several of them are very reliable and perfectly show differences between speech with and without disorders, while the others give almost the same results for both groups.

We are aware that more complex experiments with larger number of speakers, and better balanced database would be more beneficial. However, the nature of problem is making such experiments difficult to perform. People with speech disorders (also with motor impairments) have difficulties to record large amount of data and to control their way of speaking. To simplify the process we decided to allow them to record themselves at home, where they can feel more comfortable. Such approach resulted in larger amount of recorded data for each user but at the cost of the lower quality of recordings. Nevertheless, it was difficult to find representative number of users for different speech impairments. That is why we could not perform tests to find out if there were any significant differences between various types of disorders.

Our next goals are as follows:

- Improve recorded signals using automatic algorithms for noise detection and removal
- Investigate and modify algorithms of voice activity detection to work well in presence of non-stationary noise made by the speaker (such as loud breathing, noise made by wheelchair, among others). The aim is to detect when the speech is present and only then run discrimination process.

5. Acknowledgments

Our research is sponsored by the National Centre for Research and Development (NCBiR) with the grant LIDER/032/637/L-4/12/NCBR/2013 (leaded by Agnieszka Bętkowska Cavalcante PhD).

6. References

Caballero-Morales, S.-O., & Trujillo-Romero, F. (2014). Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition. *Expert Systems with Applications*, *41*(3), 841–852.

Creer, S., Cunningham, S., Green, P., & Yamagishi, J. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech & Language*, *27*(6), 1178–1193.

Hustad, K., Auker, J., Natale, N., & Carlson, R. (2003). Improving Intelligibility of Speakers with Profound Dysarthria and Cerebral Palsy. *Augmentative and Alternative Communication*, *19*(3), 187–198.

Iida, A., & Campbell, N. (2003). Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *International Journal of Speech Technology*, *6*(4), 379–392.

Kain, A. B., Hosom, J.-P., Niu, X., van Santen, J. P. H., Fried-Oken, M., & Staehely, J. (2007). Improving the intelligibility of dysarthric speech. *Speech Communication*, *49*(9), 743–759.

Lansford, K. L., & Liss, J. M. (2014). Vowel Acoustics in Dysarthria: Speech Disorder Diagnosis and Classification. *Journal of Speech, Language & Hearing Research*, *57*(1), 57–67.

Liss, J. M., LeGendre, S., & Lotto, A. J. (2010). Discriminating Dysarthria Type From Envelope Modulation Spectra. *Journal of Speech, Language & Hearing Research*, *53*(5), 1246–1255.

McCaffrey, P. (n.d.). Dysarthria Characteristics. Retrieved July 6, 2015, from http://www.csuchico.edu/~pmccaffrey//syllabi/SPPA342/342unit14.html

Muhammad, G., Mesallam, T. A., Malki, K. H., Farahat, M., Mahmood, A., & Alsulaiman, M. (2012). Multidirectional Regression (MDR)-Based Features for Automatic Voice Disorder Detection. *Journal of Voice*, *26*(6), 817.e19–817.e27.

Parker, M., Cunningham, S., Enderby, P., Hawley, M., & Green, P. (2006). Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project. *Clinical Linguistics & Phonetics*, *20*(2-3), 149–156.

Rosen, K., Murdoch, B., Folker, J., Vogel, A., Cahill, L., Delatycki, M., & Corben, L. (2010). Automatic method of pause measurement for normal and dysarthric speech. *Clinical Linguistics & Phonetics*, *24*(2), 141–154.

Rudzicz, F. (2013). Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language*, *27*(6), 1163–1177.

Sapir, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2010). Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research*, *53*(1), 114–125.

Vogel, A. P., Fletcher, J., Snyder, P. J., Fredrickson, A., & Maruff, P. (2011). Reliability, Stability, and Sensitivity to Change and Impairment in Acoustic Measures of Timing and Frequency. *Journal of Voice*, *25*(2), 137–149.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., … others. (1997). *The HTK book* (Vol. 2). Entropic Cambridge Research Laboratory Cambridge.