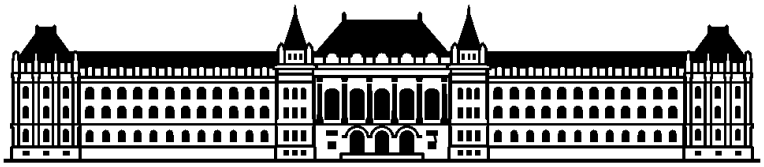

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
DEPARTMENT OF TELECOMMUNICATIONS AND MEDIA INFORMATICS

AUTOMATIC CLASSIFICATION OF DYSPHONIA

Ph.D. thesis booklet

Miklós Gábriel Tulics, MSc

Supervisor:
Klára Vicsi, DSc

BUDAPEST, HUNGARY

1 Introduction

Dysphonia (impaired voice production) generally refers to a condition where a person produces voice with an irregularity and it is affecting roughly 30% of the world's population at some point in their life [1, 2, 3, 4]. Dysphonia is not to be confused with hoarseness, as hoarseness is mostly reported by patients when they perceive an alteration in their voice quality, while dysphonia is recognized by a medical expert as hoarse, breathy, harsh or rough vocal qualities with a lower degree of phonation functionality.

Dysphonia affects patients of all ages, however research suggests that risks are higher in paediatric and elderly (>65 years of age) populations. Dysphonia is more common among professors, pedagogues, older adults and generally people who use their voice significantly more than the average in their professions [5, 6, 7, 8]. 23.4% of paediatric patients have dysphonia at some point during their childhood [9, 10, 11, 12]. The data therefore suggests that almost every fourth child produces a pathological voice. Studies agree that dysphonia is more often reported among boys than girls, the ratio being 70-30%.

People with various professions are affected by dysphonia but there is a higher likelihood of developing dysphonia among singers and entertainers, legal professionals, teachers, telemarketers etc [1, 13, 14, 15, 16, 17, 18]. Patients affected may experience an overall decrease in quality of life as it can affect a person's ability to work [19]. These people are in danger to miss work, lose wages, suffer from social isolation and develop depression. An incomprehensible speech limits a person's ability to communicate.

Dysphonia is classified as either an organic or a functional disorder of the larynx. Organic dysphonia (OD) results from some sort of physiological change in one of the subsystems of speech, while functional dysphonia (FD) refers to a voice problem in the absence of a physical condition. According to the American Speech-Language-Hearing Association, organic disorders can be subdivided into neurogenic and structural [20]. Neurogenic voice disorders include voice problems caused by abnormal control, coordination, or strength of voice box muscles due to underlying neurological diseases such as stroke, Parkinson's disease, multiple sclerosis, myasthenia gravis, and amyotrophic lateral sclerosis. Structural organic disorders include morphological alterations such as vocal cord nodules, polyps, gastroesophageal reflux disease (GERD), cyst and vocal cord paralysis (recurrent paresis, RP) [21].

2 Research Objectives

I would like to contribute with my research to the speech-based detection of dysphonia and automatic estimation of its severity in the speech of adults and children by getting a deeper understanding of the effect of functional and organic dysphonia on speech. My specific goals during the research are:

a) Examining the possibilities of automatically detecting the severity of dysphonia;

b) Attempting a binary classification to separate dysphonic and healthy speech using different machine learning approaches;

c) Analysing the possibilities of automatically separating functional and organic dysphonia;

d) Analysing the possibilities of automatically separating healthy and dysphonic voices in children.

Contribution of my theses can be summarized as follows. I used Hungarian speech samples in all of my analyses. Hungarian is a language relatively poorly researched. In fact, the topic of automatic classification of dysphonic and healthy voices using Hungarian speech samples has not been studied yet, neither in adults, nor in children.

All my analyses are done in case of continuous speech. Sustained vowels might be easier to use because they do not require a resource intensive and language-dependent segmentation. However, they lack the information (such as prosody) that could be gathered from a continuous running (context rich) speech. Continuous speech has several advantages over the analysis of sustained vowels. For example, it contains variations of fundamental frequency, pauses and phonation onsets, and provides an opportunity to examine different variations of speech sounds. Treating continuous speech constitutes a new challenge, as it requires a different approach. However, due to its many advantages I adopted this paradigm in my research.

Furthermore, I tried to automatically separate functional and organic dysphonia. To my best knowledge, there has been no research aimed at the automatic separation of functional and organic dysphonia to this date. Having such a diagnosis supporting system that can separate not only healthy from dysphonic voices, but also can predict the type of dysphonia, can greatly accelerate the process in which patients are referred to specialists. If the system detects functional dysphonia, the patient would be directed to a phoniatrist or speech therapist. However, if the system detects organic dysphonia, the patient would be directed to an otolaryngologists or oncologist. This could save a lot of time, leading the patient to care as soon as possible.

I tried to automatically separate the voices of healthy children from the ones with dysphonia. The end goal is to create a screening system that can be used by pre-school workers. If a child with dysphonic voice can be found on time, she or he has a better chance of getting professional help from an ear, nose and throat (ENT) specialist or a speech therapist.

3 Methods and materials

3.1 Methodology

During my research, I carried out statistical comparative analysis of acoustic-phonetic feature values derived from the speech of healthy people, as well as people suffering from organic and functional dysphonia. Statistical comparative analysis helps in understanding the impact of dysphonia on speech features, it can also be used as a quick characteristic selection method for machine learning processes.

I have performed classification and regression tasks, using machine learning methods, in order to examine the accuracy of automatic classification of samples of healthy subject and dysphonic voices, along with the severity of dysphonia using features extracted from speech.

3.1.1 RBH scale

The recorded voice examples were classified by a leading phoniatric according to the RBH scale [22]. The RBH scale gives the severity of dysphonia, where R stands for roughness, B for breathiness and H for overall hoarseness. The degree of the category H cannot be less than the highest rate of the other two categories. For example, if B = 3 and R = 2, H is 3, and cannot be 2 or 1. A healthy voice's code is R0B0H0; the maximum H and respectively RBH value is 3, so a voice's code with severe dysphonia is R3B3H3. Ptok and his colleagues demonstrated that the application of the RBH scale is suitable for clinical purposes [23]. This scale was used to differentiate the degree of voice disorders in the database. Speech examples of patients were labelled on the base of this numeric scale. In this study the overall hoarseness H was used.

3.1.2 Statistical methods

To inspect the relationship of the acoustic features with the severity of dysphonia Pearson product-moment correlation coefficient was calculated. To interpret the strength of the correlation, I used the guide Evans suggests for the absolute value of r [24]:

- 0.00-0.19 “very weak”;
- 0.20-0.39 “weak”;
- 0.40-0.59 “moderate”;
- 0.60-0.79 “strong”;
- 0.80-1.0 “very strong”;

To determine whether the correlation between variables is significant, one must compare the p-value to a significance level. During correlation analysis $\alpha = 0.01$ level was used.

Chi-squared test was used to compared classifier performances based on their decision tables assuming that false negatives and false positives have similar costs. The null hypothesis (H0) for a chi-square test is that the observed values and the expected values are independent, the alternative hypothesis (H1) states that they are dependent. I used a significance level of $\alpha = 0.05$.

Statistical analysis was used in order to check if there is a statistical difference in the distribution of the severity of hoarseness of the OD and FD groups. Since the RBH severity scores are ordinal the Mann-Whitney U test is used to compare severity scores in different databases. Mann-Whitney U test is a non-parametric test and it is often considered the non-parametric alternative to the independent t-test. In all Mann-Whitney U tests significance level of 95% ($\alpha = 0.05$) was used. The null hypothesis (H0) is that the distribution of the dataset is the same across the categories.

The consistency of four specialists' RBH ratings was also examined with Cronbach's Alpha and the Intra Class Correlation Coefficient (ICC). Both methods are widely used to estimate the reliability of a composite score.

3.1.3 Feature selection

In order to reduce dimensionality of the input vector the *Forward Feature Selection (FFS)* algorithm was used. Forward feature selection is an iterative algorithm, choosing the best feature that improves the performance in regard to a cost or objective function in each step and adding it to the already selected features. Here, the features were selected using maximum accuracy as an objective function.

3.1.4 Applied machine learning techniques

Binary classifications

For binary classifications an *SVM (Support Vector Machine)* classifier was used with linear and radial basis function (rbf) kernel. SVM is a supervised machine learning algorithm which is used mainly for binary classification tasks [25]. It uses the kernel trick to transform data and based on these transformations it finds an optimal boundary between the possible outputs.

The second classifier was a *Fully-Connected Deep Neural Network*, with 4 hidden layers, each of them with 25 neurons [26]. ReLU (Rectified Linear Unit) activation function was used on the hidden layers, Softmax on the output layer. Adam optimizer was used and binary crossentropy loss functions which is common for binary (two-class) classification problems. To avoid overfitting, I used dropout (with value of 0.25).

Unsupervised cluster analysis

The *k-means* is one of the simplest algorithms that uses unsupervised learning method to solve known clustering issues [27]. This method is a fast and simple approach to the problem: it is easy to implement, and it is easy to interpret the clustering results. Cluster analysis is used to classify cases into relative groups called clusters, in this case: individual assessments of severity of dysphonia. In cluster analysis, there is no prior information about the cluster membership for any of the data. If the acoustic feature set and the unsupervised learning method are fixed, it is possible to compare four cluster models for each case labelled by a specialist's judgement. In order to examine the subjective nature of RBH, k-means cluster analysis was done.

Regression analysis

Support vector regression (SVR) with linear and rbf kernel was used in order to automatically determine the severity of dysphonia [28]. Usually SVR with linear kernel less time consuming. SVR with rbf kernel has good generalization and strong tolerance to input noise.

3.1.5 Evaluation methods

To estimate and compare the performance of the machine learning algorithms *Leave-one-out cross validation (LOOCV)* was used, where the result is a large number of performance measures that can be summarized in an effort to give a more reasonable estimate of the accuracy of your model on unseen data. A downside of this approach is that it can be a computationally more expensive than a k-fold cross validation approach.

In order describe the performance of a classification or cluster model the confusion matrices are given. In my work the metrics *accuracy*, *recall* and *precision* are provided.

To describe the accuracy of regression tasks, two descriptive features are given. The performance of the regression methods is evaluated by the *root mean square error (RMSE)* value, the linear relationship between the target and the predicted H scores is described by *Pearson correlation*.

3.2 Database

3.2.1 Dysphonic and Healthy Adults Speech Database

The recordings were made using a near field microphone (Monacor ECM-100), Creative Soundblaster Audigy 2 NX external USB sound card, with good quality A/D converter and low noise level (audio coding: PCM, sampling rate: 16 kHz, quantization: 16-bit). The recordings were made in a quiet office environment (medical office). Each patient had to read out aloud one of Aesop's Fables, "The North Wind and the Sun". This folktale is frequently used in phoniatrics as an illustration of spoken language. It has been translated into several languages, Hungarian included. The text is

eight sentences long, a recording in average is 50 seconds long. The database was annotated and segmented on phone level with the help of an automatic phone segmentator which was developed in the Laboratory of Speech Acoustics [29], and was followed by manual corrections. The segmentation was done using the SAMPA phonetic alphabet [30]. In the rest of this booklet, vowels and other phones will be referred with SAMPA characters in brackets.

The collected speech database contains voices from people suffering from diseases like tumors at various places of the vocal tract, gastroesophageal reflux disease, chronic inflammation of larynx, bulbar paresis, amyotrophic lateral sclerosis, leukoplakia, spasmodic dysphonia, etc. The most frequent diseases are functional dysphonia (referred to as 'FD') and recurrent paresis (referred to as 'RP'). We refer to the recordings from patients with dysphonia as 'Dys'. Recordings from healthy control were collected as well. These recordings are used as comparison, and they were collected from people who had attended for unrelated check-ups. We refer to these recordings as 'HC'.

The distribution of the voice recordings in the database is showed in Table 1. The database contains a total of 450 recordings, 257 from patients with dysphonia (156 females and 101 males) and 193 people with a healthy voice (108 females and 85 males).

In the course of my research the database was constantly expanding with new recordings, this is why I drew my conclusions from a smaller database in some of my thesis statements. At each thesis point I present the database I used.

Table 1: Dysphonic and Healthy Adults Speech Database.

	Diagnosis		
Sex	Dysphonia	Healthy	Total
Female	156	108	264
Male	101	85	186
Total	257	193	450

3.2.2 Dysphonic and Healthy Child Speech Database

Voice samples from children were collected at several kindergartens. All the recordings were made with parental consent, mostly in the presence of the children's parents. The children recited a poem entitled "The Squirrel", written by a Erika Bartos. This poem was chosen for therapeutic reasons, speech therapists using the poem during treatment, and because children in the 5-10 year old age group are very fond of the poem and it is easy for them to learn. A recording in average is 20 seconds long. The most frequent vowel in the poem is the vowel [o], with 16 pieces followed by 14 pieces of the vowel [O] and 9 pieces of vowel [E]. The recordings were made using a near field microphone (Monacor ECM-100), Creative Soundblaster Audigy 2 NX external USB sound

card, with 44.100 Hz sampling rate and 16-bit linear quantization. The duration of the recordings is about 20 seconds each.

The segmentation was made with the help of an automatic phone segmentator (mentioned in Section 3.2.1), followed by manual corrections. A total of 59 recordings were used in this work: 25 voices from children with dysphonia (mean age: 6.52(±1.94)) (3 children had vocal nodes, the rest had functional dysphonia) and 34 recordings from healthy children (mean age: 5.35(±0.54)). Table 2 summarizes the recordings from the database used in the experiments.

Table 2: Dysphonic and Healthy Child Speech Database.

	Diagnosis		
Sex	Dysphonia	Healthy	Total
Female	5	15	20
Male	20	19	39
Total	25	34	59

3.3 Input vector

3.3.1 Input vector from acoustic features

In the case where I examined adult's voice, I created the input vector for the classifiers used from acoustic features measured on vowel [E] (being the most frequent vowel in the read text), on different phonetic classes and on the whole wave file.

On vowel [E] jitter(ddp), shimmer(ddp), HNR (Harmonics-to-Noise Ratio) and the first component (c_1) of the mel-frequency cepstral coefficients (referred to as 'mfcc01') were measured. The abbreviation 'ddp' refers to Difference of Differences of Periods. **On different phonetic classes** Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based IMF entropy frequency band ratios were measured on the voiced parts of speech, and the measured features were grouped into different phonetic classes: the vowel [E], nasal phones marked with [m], [n] and [J], on high vowels marked with [E], [e:], [i], [2] and [y], on low vowels marked with [O], [A:], [o] and [u], voiced spirants marked with [v], [z] and [Z], voiced plosives and affricates marked with [b], [d], [g], [dz], [dZ] and [d']. SPI was calculated on the **whole sample** as well.

Derived acoustic features were calculated as the means, standard deviations and ranges of acoustic features. In this way, a total of 49 acoustic features were measured per each patient's voice sample, so 49 dimensional input vector was prepared from acoustic features. Detailed description of the acoustic features can be found in [J4]. This feature set is further referred to as 'the 49 feature set'.

When dealing with children's speech acoustic features were measured on vowel [o] (being the most frequent vowel in the poem), on different phonetic classes and on the whole wave file.

On vowel [o] the following acoustic features were measured: jitter(ddp), shimmer(ddp), HNR (Harmonics-to-Noise Ratio), 12 mfccs, the fundamental frequency (F0), formant frequency (F1, F2, F3), Formant frequency bandwidth (F1BW, F2BW, F3BW). The fundamental frequency calculation was done by an autocorrelation method described in [31]. Formant frequency tracking was realized by applying Gaussian window for a 150 ms long signal at a 10 ms rate. For each frame LPC coefficients were measured.

On different phonetic classes SPI and IMF entropy frequency band ratios were measured on the voiced parts of speech, and the measured features were grouped into different phonetic classes: on vowel [o], on nasal phones, on high vowels, on low vowels, on voiced spirants, on voiced plosives and affricates. SPI was calculated on the **whole sample** as well. **Derived acoustic features** were calculated as the means, standard deviations and ranges of acoustic features. A total of 103 acoustic features were calculated per each children's voice sample.

3.3.2 Input vector from phone level posterior probability values of an ASR

The acoustic models of Automatic Speech Recognizers (ASR) can also be used to extract features for dysphonia detection and classification. Today‘s state-of-the-art hybrid ASR acoustic models are composed of a transition model (a Hidden Markov Model) and a phone classifier (DNN) [32]. The phone classifier can also be used to classify frames in standalone mode (without adding the recognition network and the ASR decoder) by using a forward pass for the speech frames one-by-one. In this way we obtain posterior probabilities of phones every 10 ms time frame from the DNN softmax layer of the phone classifier. Hence, only the phone classifier component of the acoustic model is used for prediction.

The acoustic model used for my experiments was trained on Hungarian data mixed from BABEL [33], the Hungarian Reference Speech Database (MRBA) [34] and the Hungarian Broadcast News Database [35] with the Kaldi toolkit [36], following the ‘nnet2‘ WSJ recipe. Its phone classifier is based on spliced and LDA+MLLT transformed MFCC features input into a feed-forward DNN with 4, 1024 dimensional hidden layers with p-norm nonlinearity (p=2) and a softmax output for up to 2500 senones (context sensitive logical phone entities). After the forward pass in inference time, the senones are collapsed to phones of the 39 element SAMPA Hungarian phone set [33].

When the phoneme in question takes the highest probability value of all the other phonemes in the frame (i.e., the phoneme ‘wins its frame’), I stored the values in a list, then I calculated the mean, standard deviation and the range. I did the calculations for vowels [E], for nasals, high vowels, low vowels, voiced spirants and plosives and affricates, for each recording. This resulted in

a 21 dimensional vector per recording. I refer to this input vector as "ASR posterior features" to maintain coherence of terms with international literature, although as we have seen, these features are not ASR features in the strict sense, as they are generated by the phone classifier of a small acoustic model.

4 Results

4.1 The examination of the automatic assessment of the severity of dysphonia

4.1.1 Phonetic-class based correlation analysis for the severity of dysphonia

In the diagnosis and management of dysphonic speech, a voice clinician typically assesses the quality of a patient's voice personally. The assessment is subjective by nature. The target severity of a voice is usually defined as one clinician's assessment or as the median or average severity rating determined by a group of experienced raters assessing the voice [37, 38]. If multiple raters are recruited for the subjective assessment of severity of dysphonia, the assessment is done by listening to the previously recorded voice samples. The assessment can vary among raters; thus, analysis of rating consistency is advisable. In the work of Law and his colleagues [39], it was found that higher intra-rater reliability was achieved with continuous speech than with sustained vowel samples. In most voice clinics, acoustic measures are derived from sustained vowel samples; however, continuous speech has several advantages over analysis of sustained vowels. It contains a variation of fundamental frequency, pauses and phonation onsets, and there is the opportunity to examine different combinations of speech sounds.

An important task is to identify relevant acoustic features to predict the severity of the dysphonic voices automatically. The following theses address this issue.

Using a small speech database, it is very important to optimize the speech features as much as possible, rather than to use a lot of acoustic features, with the risk of bringing unwanted noise into the system. My hypotheses are the followings: speech defect severity determined by a clinician (RBH) is correlated (coincides) with the distortion degree of the characteristic acoustic features.

In my first thesis I performed correlation analysis between acoustic features (presented in subsection 3.3.1) and the severity of hoarseness given by a specialist. The specialist treated the patient and determined the diagnosis. The specialist directly listened to and evaluated the quality of the patient's speech during the consultations. The Pearson correlation coefficient was calculated in every case where correlation was significant at the 0.01 level (2-tailed) between the acoustic feature and the subjective rating.

The analysis was carried out on a subset of the database presented in subsection 3.2.1. The distribution of the voice recordings by H used in this experiment is shown in Table 3. Note that the recordings with the value H equal to 0 are all recordings from healthy patients. Thus, a total of 136 records from healthy people and 206 records from patients suffering from dysphonia were used.

The results presented on Figure 1 indicate that features such as jitter(ddp), shimmer(dda), Harmonics-to-Noise Ratio (HNR) and "mfcc01" correlate with the severity of dysphonia. In the

Table 3: Distribution of healthy and dysphonic speakers in the database, depending on the value of H.

Count	**Value of H**				**Total**
	0	1	2	3	
Male	67	34	20	32	153
Female	69	72	27	21	189
Total	136	106	47	53	342

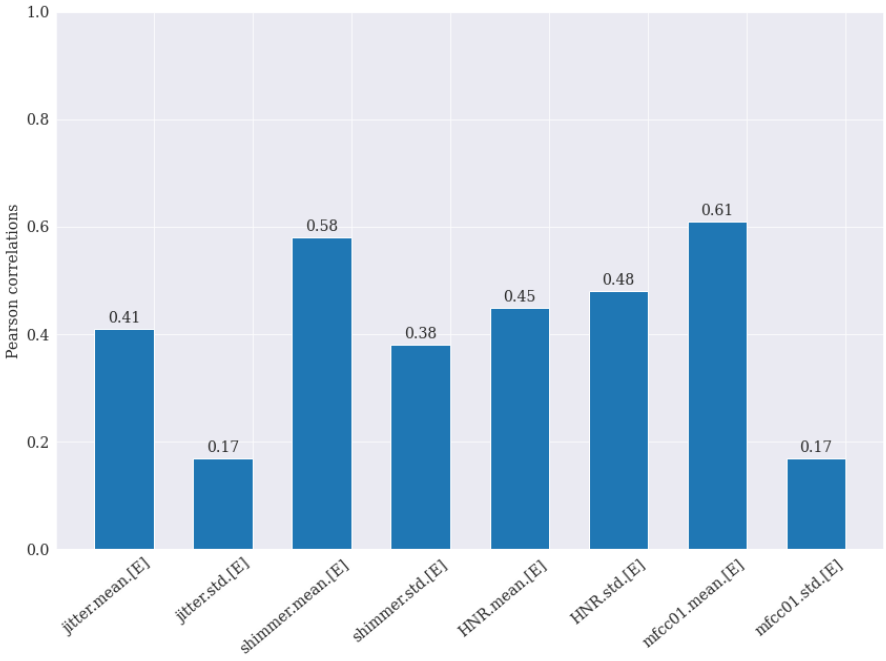

Figure 1: Pearson correlation with commonly used acoustic features.

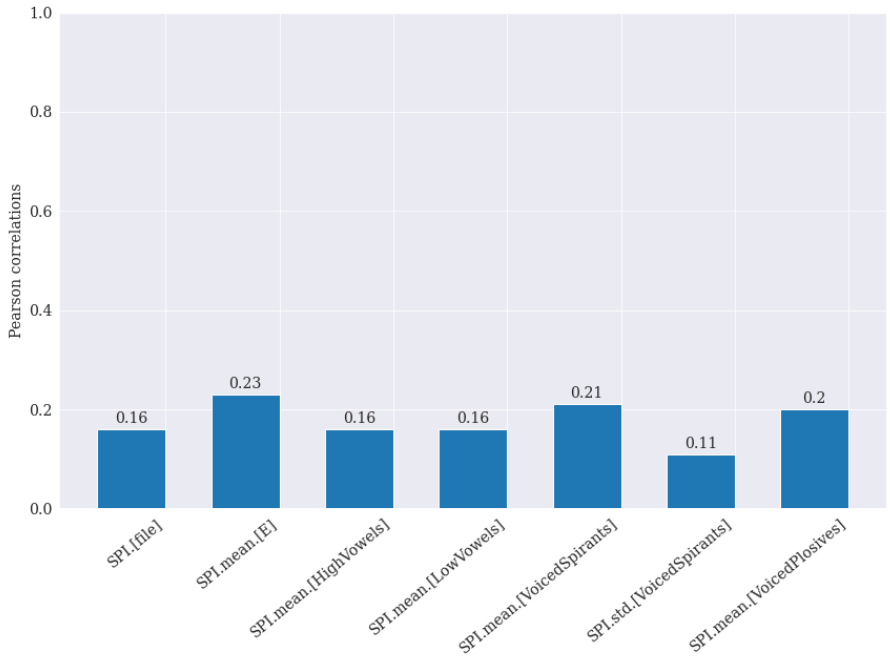

Figure 2: Pearson correlation with SPI measured on phonetic classes.

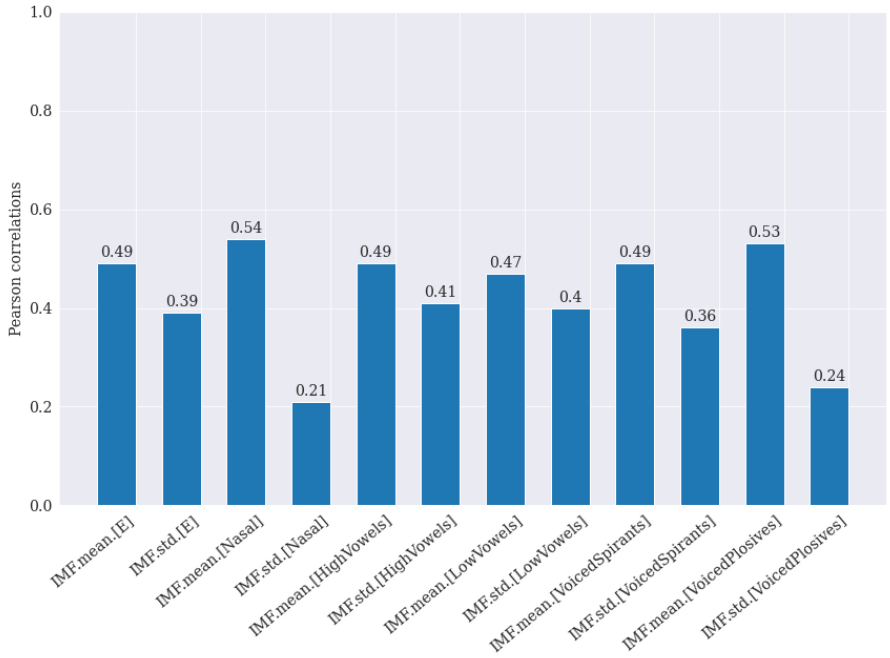

Figure 3: Pearson correlation with IMF entropy-based frequency band ratios measured on phonetic classes.

figures, the absolute values of the Pearson correlations are shown where correlation is significant at the 0.01 level (2-tailed). The greater the absolute value of the correlation coefficient, the stronger the relationship between the acoustic features and the severity of hoarseness. According to Evans suggestions, the correlation of jitter.std.[E] and mfcc01.std.[E] with the severity of hoarseness is "very weak", the correlation of shimmer.std.[E] is "weak", the correlation of jitter.mean.[E], shimmer.mean.[E], HNR.mean.[E] and HNR.std.[E] is "moderate", while the correlation of mfcc01.mean.[E] with the severity of hoarseness can be considered "strong".

When SPI was measured on phonetic classes, the Pearson correlation coefficients ranged from 0.11 and 0.23, indicates "very weak" and "weak", but significant correlation. The results are shown in Figure 2.

EMD-based IMF entropy frequency band ratios correlate with the severity of dysphonia, as Figure 3 suggests. IMF.std.[E], IMF.std.[Nasal], IMF.std.[VoicedSpirants], IMF.std.[VoicedPlosives] show "weak" correlation, while IMF.mean.[E], IMF.mean.[Nasal], IMF.mean.[HighVowels], IMF.std.[HighVowels], IMF.mean.[LowVowels], IMF.std.[LowVowels], IMF.mean.[VoicedSpirants] and IMF.mean.[VoicedPlosives] show "moderate" correlation with the severity of dysphonia.

Thesis I. A. [C4] *I showed that jitter(ddp), shimmer(dda), Harmonics-to-Noise Ratio (HNR), mfcc01, Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based IMF entropy frequency band ratios measured at specific phones show significant correlation at the 0.01 level with the severity of dysphonia when measured on the Hungarian Dysphonic and Healthy Adult Speech Database.*

4.1.2 Unsupervised and supervised learning methods for the modeling of the four grade assessments of the specialists

It is also an important question whether the acoustic features selected by the correlation analysis are suitable for modelling the four grade assessments of the specialists (RBH subjective scale). In this investigation two datasets were used, the Initial Dysphonic and Healthy Database and the Selected Dysphonic and Healthy Database. An unsupervised learning method, the k-means algorithm was used on the Selected Dysphonic and Healthy Database. K-means clustering is a type of unsupervised learning where we have unlabelled data. The goal of the algorithm is to find groups in the data called clusters, with the number of groups represented by the variable k. Before I performed the unsupervised learning method a two-class classification was performed to find out whether the chosen acoustic features are rich enough in information to differentiate between healthy and dysphonic voices, even after reducing the dimensionality of the input vector.

The Initial Dysphonic and Healthy Database contains a total of 263 speech recordings, 127 recordings from healthy subjects (62 male and 65 female) and 136 recordings from patients suffering from functional or organic dysphonia (66 male and 70 female), thus each recording is from a separate subject. The specialist who treated the patient determined the diagnosis. The specialist evaluated the quality of the patient's speech during consultation time. This database was used for the two-class classification experiment.

Four specialists were asked to evaluate the voice recordings of the Selected Dysphonic and Healthy Database with respect to the severity of the dysphonia. The Selected Dysphonic and Healthy Database contains a total of 148 recordings, and it was used for the unsupervised cluster and regression analysis. One of the four specialists set up the diagnosis and evaluated the quality of the patient's speech during the consultations; the other three specialists did not know the patient and only listened to the previously recorded voice files and determined the severity of dysphonia. Every rater is experienced in working with patients with voice disorders and dysphonia.

A two-class classification was performed on the Initial Dysphonic and Healthy Database using leave-one-out cross validation, with SVM classifier. Classification experiments were made using several combinations. Liner and rbf kernels were also tried out. The default value of C of support vector machine is 1, while γ is 1/number of features. In order to choose the optimal hyperparameters for the SVM classifier grid search was used. Leave-one-out cross validation was used in all cases.

The highest accuracy of 89% was reached by using an rbf kernel. The FFS feature selection algorithm reduced the input dimensionality to 18 acoustic features. The acoustic features selected by the FFS algorithm are the following: jitter.mean.[E], shimmer.mean.[E], HNR.mean.[E], mfcc01.mean.[E], jitter.std.[E], shimmer.std.[E], HNR.std.[E], mfcc01.std.[E], SPI.std.[E], SPI.mean.[Nasal], SPI.std.[Nasal], SPI.std.[LowVowels], SPI.mean.[VoicedSpirants], SPI.std.[VoicedSpirants], IMF.std.[E], IMF.mean.[Nasal], IMF.mean.[VoicedPlosives], IMF.std.[VoicedPlosives].

It is an interesting question whether the chosen acoustic features can model the individual assessments. Cluster analysis tries to identify structures, homogeneous groups of cases not previously known within the data. In my case, the hidden structure is the "true label" of severity score for each recording given by an ideal examiner. The cluster analyses is used to mimic this ideal examiner in order to get the true label for each recording. Of course, we rely more on a specialist's rating than on a clustering process, but if the found clusters are really close to a specialist's rating we can consider calling that assessment as true. Hence, I need to compare four cluster models labelled by a specialist's judgement. If the acoustic feature set and the unsupervised learning method are fixed, it is possible to compare four cluster models, for each case, labelled by a specialist's judgement. To examine the subjective nature of RBH, k-means cluster analysis was performed.

k-means clustering has the objective of putting the observations into k clusters, where k is the number of clusters determined by the user as an input. I set the number of clusters to four. The cluster analysis classified the observations into clusters A, B, C, and D.

The clusters where assigned to the severity value by the minimum mean of absolute errors (minimum of MAE): A to H = 0, B to H = 1, C to H = 2, and D to H = 3. The confusion matrices for each specialist are shown separately in Table 4, 5, 6 and 7. The accuracies for the decision in case of each specialist in order is: 49%, 44%, 45%, 47%, the mean accuracy is 46.25% with 2.22% standard deviation. In the case of a balanced distribution of 4 classes, the baseline classification would be 25%. From this experiment I can conclude that the acoustic feature set is suitable for modelling the individual assessments of dysphonia severity.

Table 4: Confusion matrix based on the assessment of Specialist 1.

		Specialist 1 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**12**	1	2	1	75%
	1	13	**33**	5	3	61%
	2	9	25	**10**	5	20%
	3	2	4	6	**17**	59%
Class recall		33%	52%	43%	65%	

Table 5: Confusion matrix based on the assessment of Specialist 2.

		Specialist 2 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**11**	3	2	0	69%
	1	5	**26**	23	0	48%
	2	6	24	**16**	3	33%
	3	0	2	15	**12**	41%
Class recall		50%	47%	29%	80%	

Table 6: Confusion matrix based on the assessment of Specialist 3.

		Specialist 3 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**11**	2	2	1	69%
	1	2	**20**	25	7	37%
	2	3	15	**16**	15	33%
	3	0	1	9	**19**	66%
Class recall		69%	53%	31%	45%	

Table 7: Confusion matrix based on the assessment of Specialist 4.

		Specialist 4 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**12**	2	2	0	75%
	1	7	**24**	18	5	44%
	2	6	18	**17**	8	35%
	3	0	6	6	**17**	59%
Class recall		48%	48%	40%	57%	

I calculated the Pearson correlation between the cluster defined severity scores and the individual specialists' ratings and I also evaluated this using the mean RBH perceptual evaluation of the four specialists. Values are shown in Table 8. All correlations show "moderate" relations. The mean correlation is 0.52 with 0.01 standard deviation. The highest value was measured between the cluster defined severity scores and the mean of the ratings, giving a value of 0.59. Since the found clusters correlate the best with the mean of the four specialists, this is the true label I use for the regression analyses.

Table 8: Pearson correlation between the cluster defined severity scores and the specialists' ratings.

	Specialist 1	Specialist 2	Specialist 3	Specialist 4	The mean of the ratings
Pearson correlation	0.51	0.54	0.53	0.51	0.59

Thesis I. B. [C4, J4] *I showed that when clustering the data, with the selected acoustic features using k-means clustering, the found clusters correlate well with the severity of dysphonia. A 0.59 Pearson correlation was achieved between the cluster defined values and the mean of the four specialists' ratings.*

4.1.3 The automatic assessment of the severity of dysphonia with regression analysis

This analysis was performed on the Selected Dysphonic and Healthy Database. As a result of subsection 4.1.2 the mean RBH perceptual evaluation of specialists was used as the target for my regression models.

It is important to analyse whether the rater reliability of the 4 experts is consistent enough. The value of the internal consistency of the specialists gives us an idea of what is the maximum correlation value we can expect from our regression model. We do not expect the regression model to achieve better results than a well trained specialist. For measuring internal consistency ("reliability") of the raters' evaluations Cronbach's Alpha and the Intra Class Correlation Coefficient (ICC) methods were used. Despite the interesting differences among the decision of the specialists, a high degree of reliability (Cronbach's Alpha = 0.89, ICC = 0.89) was measured between their severity judgements when measuring internal consistency.

Regression has a significant advantage compared to cluster analysis, since it's prediction is not an ordinal, but a continuous variable. This property can significantly improve the quality of the model. Due to the small sample size, leave-one-out cross validation was used. The performance of the regression methods is evaluated by the RMSE value, the linear relationship between the target and the predicted H scores is described by Pearson correlation. To find the optimal hyperparameters grid search was used.

In this analysis, support vector regression with linear and radial basis function kernel were used. To reach the best performance the 18-feature set (described in 4.1.2) and the result of the FFS algorithm was used, for SVR with linear and rbf kernel separately. As previously mentioned the mean of the four specialists' ratings was used as target. Table 9 summarizes the results.

Table 9: Regression analysis results – the mean of the four specialist's ratings as target.

Acoustic feature set	**Type of regression**	**Correlation**	**RMSE of H**	**hyperparameters**
18 feature set	linear kernel	0.83	0.50	C = 1
Result of FFS, 8 feature set	linear kernel	**0.85**	**0.46**	C = 1
18 feature set	rbf kernel	0.81	0.51	C = 2, $\gamma = 0.125$
Result of FFS, 14 feature set	rbf kernel	**0.85**	**0.45**	C = 4, $\gamma = 0.25$

The FFS algorithm reduced the original 33-dimension input to only eight features using linear kernel. The following features were selected mfcc01.mean.[E], shimmer.mean.[E], SPI.std.[LowVowels], HNR.std.[E], SPI.mean.[HighVowels], IMF.mean.[Nasal], SPI.std.[VoicedPlosives],

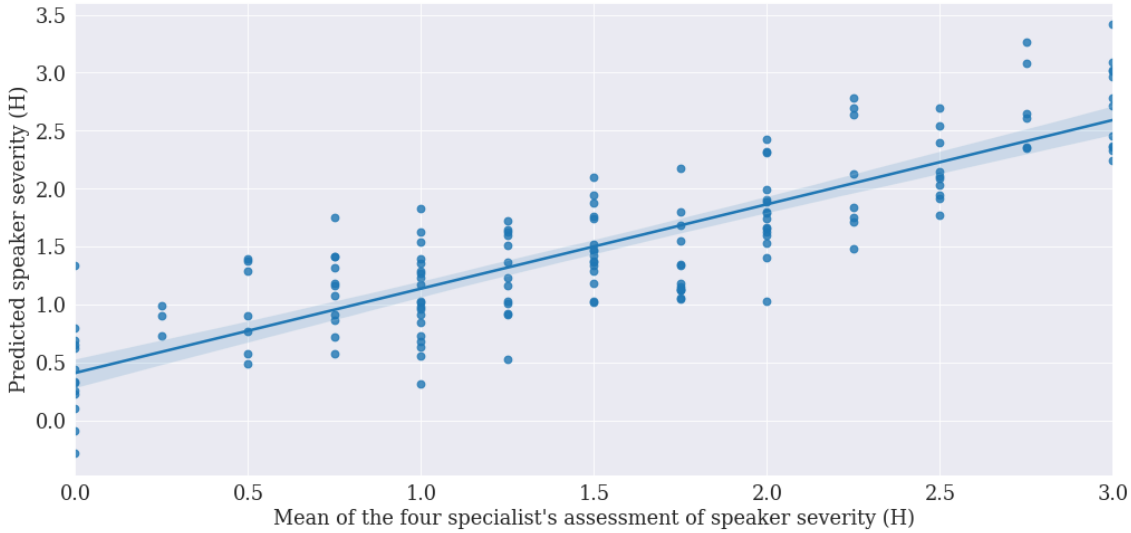

Figure 4: Automatically predicted dysphonia severity degree according to perceptual assessment of H, using SVR with linear kernel regression with 8 parameters.

IMF.std.[LowVowels]. This configuration gave the highest 0.85 correlation.

When rbf kernel was used, the FSS algorithm selected 14 features, these were the following: shimmer.mean.[E], HNR.mean.[E], mfcc01.mean.[E], HNR.std.[E], SPI.mean.[E], SPI.std.[E], SPI.std.[Nasal], SPI.mean.[HighVowels], SPI.mean.[LowVowels], SPI.std.[LowVowels], SPI.mean.[VoicedPlosives], IMF.mean.[Nasal], IMF.mean.[VoicedPlosives], IMF.std.[VoicedPlosives]. The lowest RMSE value of 0.454 was obtained here. Furthermore, the FFS models gave only slightly better results than the models with the 18-feature set.

This illustrates the capacity of the proposed approach in predicting the severity of dysphonia regardless of the speaker's pathology or severity degree. Since the ICC value between the 4 specialists resulted in 0.89 it can be considered as a theoretical goal we want to achieve. In light of this, the correlation value of 0.85 obtained by the regression model is considered almost perfect.

Figure 4 depicts the automatically predicted severity of the dysphonia compared to the reference perceptual assessment of speaker severity. The figure shows the SVR linear kernel regression model created by the result of the FFS algorithm. The figure illustrates once again the capacity of the proposed approach in predicting the severity of dysphonia regardless of the speaker's pathology or severity degree. It can be observed that the model gives good prediction of severity of H1.

Thesis I. C. [C4, J4] *I showed that an automatic estimation of the severity of dysphonia is possible using only eight acoustic features as input vector with a SVR with linear kernel reaching 0.85 Pearson correlation and 0.46 RMSE on the Selected Dysphonic and Healthy Database.*

4.2 The automatic classification of dysphonic and healthy speech

4.2.1 The comparison of SVM and DNN classifiers in case of acoustic features as an input vector

In order to do a binary classification of dysphonic and healthy speech, researchers generally use a wide variety of acoustic features, derived from speech and used as input vectors with machine learning algorithms [40, 41].

For classification tasks, a common machine learning algorithm is based on SVMs [42, 43], as they are good at dealing with small data samples, but Deep Learning technics are also exploited [44, 45, 46, 47, 48, 49]. Deep neural networks (DNNs) are used on a variety of tasks, usually on big datasets.

In this experiment I used two classification approaches. The first classifier used was SVM, the second classifier was a Fully-Connected Deep Neural Network as described in Section 3.1.4. FFS algorithm was used in order to reduce dimensionality of the input vector in the case when SVM was used as a classifier. More on the FFS algorithm in Section 3.1.3. In order to choose the optimal hyperparameters for the SVM classifier grid search was used.

The database used in this experiment is the same database described in Section 3.2.1 and in Table 1. The database contains a total of 450 recordings, 257 from patients with dysphonia (156 females and 101 males) and 193 people with a healthy voice (108 females and 85 males).

I created the input vector from acoustic features described in Section 3.3.1, thus 49 acoustic features were used as input vector.

The results of the binary classification with LOOVC between healthy and dysphonic voices using acoustic features as input vector are shown in Table 10.

Table 10: Two-class classification results between HC and Dys in case of leave-one-out cross validation.

Input vector	**FFS**	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
Acoustic features	Yes	11	SVM	linear kernel	C = 1	85%
Acoustic features	Yes	9	SVM	rbf kernel	C = 256; $\gamma = 0.0625$	85%
Acoustic features	No	49	SVM	linear kernel	C = 4	83%
Acoustic features	No	49	SVM	rbf kernel	C = 1024; $\gamma = 0.00098$	84%
Acoustic features	No	49	DNN	dropout	0.25	**88%**

The first column of the table shows the type of the input vector, the next whether FFS was performed on the input vector or not, then the classifier and the configuration used, followed by the accuracy. In case of SVM, grid search was used in every case. In case of DNN dropout value of 0.25 was used.

Using DNN as a classifier yields higher accuracy than the SVM approach with 3.53% relative accuracy increase, resulting in the highest accuracy of 88%. Also, there is no considerable difference in accuracy between linear and rbf kernel in case of SVM (85% and 85%).

The confusion matrix using FFS and SVM with linear kernel is shown in Table 11, while the confusion matrix of the DNN scenario in Table 12. When using SVM the class precision of the HC class is 84.83%, while the precision of the Dys class is 84.56%. The recall for class HC is 78.24% and 89.49% for class Dys.

Table 11: Confusion matrix using FFS and SVM with linear kernel.

	true HC	**true Dys**	class precision
pred. HC	151	27	84.83%
pred. Dys	42	230	84.56%
class recall	78.24%	89.49%	

Table 12: Confusion matrix using a Fully-Connected Deep Neural Network.

	true HC	**true Dys**	class precision
pred. HC	189	48	79.75%
pred. Dys	4	209	98.12%
class recall	97.93%	81.32%	

The confusion matrix of the 88% accuracy setting is shown in Table 12. As the table suggests, the class precision of the HC class is 79.75%, while the precision of the Dys class is 98.12%. This means that the number of cases where the two classes were predicted correctly is not balanced. The recall for class HC is 97.93% and 81.32% for class Dys.

Accuracy is not the only absolute measure by which we characterize our classifier. It obscures a lot of important information, so it should be handled with care. We like confusion matrices if they are symmetric, if the mismatch weights of the classes are even. In medical applications, in general, the confusion matrix is not a symmetric matrix. Classifying a sick person as healthy is a more serious mistake than classifying a healthy person as sick. This means that the recall (also known as sensitivity) of class Dys is also a very important aspect of the classifier. The lower the risk that a person with dysphonia is miss-classified as healthy the better. We rather have some healthy people labelled dysphonic over predicting a dysphonic person healthy. In this sense using FFS and SVM with linear kernel seems to be a better approach since the recall of Dys is 89.49%,

while when using DNN the recall of Dys is 81.32%.

If false negatives and false positives have similar costs two decision tables can be constructed by the number of good predictions and the number of miss-predictions and chi-square test be performed. The p-value of the test was 0.09, so there is no statistical difference found between the system with 85% accuracy (provided by the SVM with linear and rbf kernel) and the 88% accuracy Fully-Connected Deep Neural Network.

Thesis II. A. [C2, C9] *I showed that the binary classification of dysphonic and healthy voices is possible for Hungarian. When applying a Fully-Connected Deep Neural Network, an accuracy of 88% can be achieved with LOOCV, using acoustic features as input on the Hungarian Dysphonic and Healthy Adult Speech Database.*

4.2.2 Using ASR posterior probability features as input vectors for the DNN classifier

Automatic speech recognition (ASR) is traditionally decomposed into creating an acoustic and a language model with a vocabulary [50]. The acoustic model is most often a hybrid of a Hidden Markov Model (HMM) to facilitate dynamic warping for alignment, and a set of phone or phone alike (obtained through a decision tree to group acoustically similar entities) models responsible for providing similarity measures between the actual frame(s) to be classified and the phone (senone) set. Although this set of models rarely corresponds to pure phone models, for simplicity I will refer to this as a 'phone model', especially as the output of the phone model can be collapsed to phone posteriors. It is obvious that these phone posteriors can be used individually to classify frames, or better, to derive a Goodness of Pronunciation (GOP) score [51] which can be used in speech assessment to automatically evaluate pronunciation.

Using the GOP or pure phone posteriors as additional or standalone features to detect or classify voice disorders – although not particularly dysphonia – has been addressed by many researchers. For example, in [52] and [53] researchers use ASR posteriors to predict severity of 'general' voice disorders, that is the type and characteristics of the disorders are not classified, but their severity is known. The frame level posteriors produced by a DNN phone model are a good measure of the acoustic mismatch caused by voice quality change and thus can be exploited for classification and assessment of voice disorders.

I argue that this method should be treated with caution in dysphonia. It is questionable whether an acoustic model trained for normal speech recognition can be used to distinguish dysphonic speech at all. The training of an Automatic Speech Recognition system require large amount of training data recorded from speakers with different genders, voice characteristics, regional accents,

education backgrounds, etc. The training data may include different accents and dialects, voices from smokers, from the young and from the elderly. Dysphonic speech, even by accident, can be included in larger numbers in the database. The general goal of speech recognition is to recognize hoarse, nasal, sad, cheerful, old and young speech equally. I wanted to verify this hypothesis if using or adding phone posteriors might produce a better classification system for dysphonia than using acoustic features as input vector. To the best of my knowledge, this issue has never been evaluated for dysphonic speech.

I created the input vector from posterior probabilities of phones as described in Section 3.3.2, thus 21 phone posterior features were used as input, then the input vector was fed into a dysphonia classifier (DNN) and compared to the result presented in section 4.2.1. Results show that using acoustic features as the input vector of the classifier outperforms the ASR posterior features using DNN as a classifier. An accuracy of 88% were reached when acoustic features was used as input and 60% when ASR posterior features were used as input vector.

Table 13: Two-class classification results between HC and Dys using DNN and comparing input vectors.

Input vector	**FFS**	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
Acoustic features	No	49	DNN	dropout	0.25	88%
ASR posterior features	No	21	DNN	dropout	0.25	60%
Joint features	No	70	DNN	dropout	0.25	**89%**

Table 14: Confusion matrix using a Fully-Connected Deep Neural Network with the joint features vector.

	true HC	**true Dys**	class precision
pred. HC	167	25	86.98%
pred. Dys	26	232	89.92%
class recall	86.53%	90.27%	

Since the ASR posterior features fall short behind the results obtained by the acoustic features, I examined whether the combination of the two input vectors (called "joint feature vector") increases the result of the classification accuracy. When the joint feature vector was used at the input of the neural network, the classification accuracy increased to 89%. Results of the classifications are shown in Table 13. While this is better than just using acoustic features, there is no significant impact of using ASR posterior probability values.

The confusion matrix of the DNN with the joint vector feature input can be seen in Table 14. The class precision of the HC class is 86.98%, while the precision of the Dys class is 89.92%. The recall for class HC is 86.53% and 90.27% for class Dys.

Although the class recall of Dys when using the joint feature vector is higher than using only acoustic features (90.27% in the first case and 81.32% in the second), comparing the joint feature vector's result with the case when acoustic features where used with FFS and SVM with linear kernel (presented in Table 11) the increase in the recall of Dys is not significant.

If false negatives and false positives have similar costs and chi-square test are performed there is no significant difference between the classifications with accuracy of 85%, 88% and 89%. The p-value between the acoustic features with SVM and the joint feature vector with DNN is 0.07, while between acoustic features with DNN and the joint feature vector with DNN is 0.91. Based on these, it can be concluded that it is not worthwhile to calculate ASR phone posterior, as it has no significant impact, but it can greatly complicate and slow down the current proposed system.

To verify why the ASR posterior probability features failed to improve the classification accuracy the distributions of a specific phone posteriors for the four severity categories (H) were calculated. In my first approach I calculated the highest posterior (where the phoneme 'wins its frame') of the frames across the four severity categories, in the second approach I calculated all the posteriors where the specific phone appeared. The results are shown in Figure 5 and 6 for phone [E]. It is seen that different severity categories do not separate well by the phone posteriors. Other phones have similar trend.

I argue that the different objective criteria and uncontrolled data with respect to dysphonia that are used when training ASR phone models justify why phone posteriors could not help improving my results. The way to move forward in this research is to get more data and search for further acoustic features.

From these results the following theses can be formulated.

Thesis II. B. [C2, C9] *I have shown, that using ASR phone posterior derived features, that were trained for general ASR purpose, is less effective in the automatic classification of healthy and dysphonic voices, than using the acoustic feature set directly. Deeper analyses showed weak relation between phone posterior distributions and dysphonia severity scores.*

Thesis II. C. [C2, C9] *I have shown that adding ASR phone posterior derived features to the acoustic features does not significantly improve the automatic classification accuracy of healthy and dysphonic voices.*

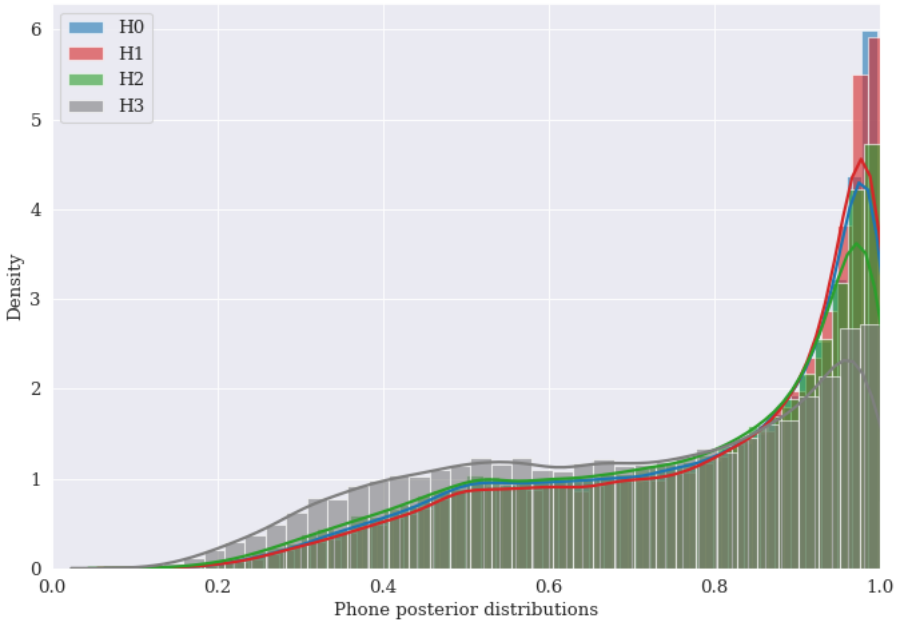

Figure 5: Phone posterior distributions of highest probability [E] phones.

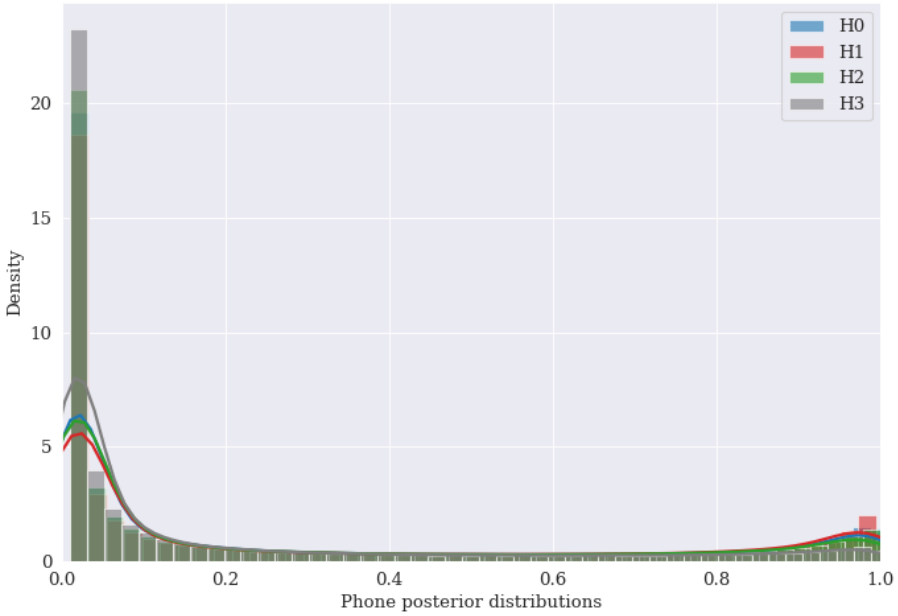

Figure 6: Phone posterior distributions of all [E] phones.

4.3 The automatic classification of functional and organic dysphonia

As mentioned in the Introduction (Section 1) dysphonia can be classified as either an organic or a functional disorder of the larynx.

According to Barth [54] and Stern [55] we are talking about a functional phonation disorder, if the diagnostic tools available to us do not detect organic lesions. The voice organs are healthy, yet the interplay of the temporal and dynamic systems of the factors necessary for voice production is disturbed. Weiss states that the "functional" indicator is temporary and valid only until the means of science are able to reveal the real organ causes of the illness [56]. Gundermann argues against this view and states it is not appropriate to use the term "functional" instead of "lack of organic" [57]. Organ abnormality can be the starting point of a functional disorder and vice versa and it can lead to organic alteration. From the literature presented above, the two categories do not seem to be always mutually exclusive.

It is an interesting question whether it is possible to automatically separate functional from organic dysphonia. If functional dysphonia can be determined with high probability, with the help of a diagnosis support system, the patient would be directed to a phoniatrician or speech therapist. If the system detects organic dysphonia the patient would be directed to an otolaryngologists or oncologist. This would save a lot of time and would lead the patient to care as soon as possible.

There are disputes in the definition and separation of FD and OD, and that the two categories may not be always mutually exclusive. It is natural that the two groups could better classified on a database where the distributions of the severity of hoarseness for the two groups are statistically different, for example if the OD group has a statistically significantly higher degree of severity than the FD group. In this way, the classifier may divide the severity of hoarseness (and not the disease types) into two groups: low and high. What we really want to achieve instead is to classify the two disease types in two. To investigate this phenomenon, the Filtered Dysphonic Database was created such way that the distribution of the severity of hoarseness was not significantly different in the OD and FD groups.

In this Section I make an attempt to automatically separate functional from organic dysphonia with the help of SVM algorithm and try to identify acoustic features that are best for classification purposes.

The database used in this experiment is the filtered version of the database described in Section 3.2.1. The Filtered Dysphonic Database was created is such a way that the distribution of sexes and hoarseness levels in the OD and FD groups are equal. The Filtered Dysphonic Database contains a total number of 164 recordings, 82 from patients suffering from organic and 82 suffering from functional dysphonia. The database contains 122 females (61 with OD and 61 with FD) and 42

males (21 with OD and 21 with FD) recordings. The mean hoarseness score (H parameter from the RBH subjective scale) given for the OD group was 1.5 with 0.7 standard deviation, while for the FD group the mean was 1.4 with 0.7 standard deviation. The description of the Filtered Dysphonic Database is shown in Table 15.

Table 15: The Filtered Dysphonic Database

	Number of female recordings	**Number of male recordings**	**H severity**	**Female H severity**	**Male H severity**
OD	61	21	1.5 (±0.7)	1.5 (±0.6)	1.5 (±0.8)
FD	61	21	1.4 (±0.7)	1.3 (±0.7)	1.5 (±0.8)

The same 49 acoustic features were used for the classifier input vector as described in Section 3.3.1.

The Filtered Dysphonic Database is constructed with the purpose that it should not have significant difference in the distribution of the severity in the OD and FD dataset. Since the severity scores are ordinal the Mann-Whitney U test is used check the statistical difference. Section 3.1.2 describes the Mann-Whitney U test in more detail. Significance level of 95% ($\alpha = 0.05$) was used.

When performing the Mann-Whitney U test, the calculated p-value equals to 0.65 for the total sample, 0.34 in case of females, 0.65 in case of males (p-value $> \alpha$). The distribution of hoarseness of the OD group is considered to be the same to the distribution of hoarseness of the FD group. In other words, the difference between the distributions of hoarseness of the OD and FD populations is not big enough to be statistically significant. Thus, the Filtered Dysphonic Database is suitable for further classification investigation as it excludes the possibility that the classifier is classifying the degree of hoarseness.

In the experiment SVM with FFS was used with LOOCV data split technique. More on the SVM classifier and FFS algorithm in Section 3.1.4 and 3.1.3. For input the 49 feature set was used described in Section 3.3.1.

Table 16 shows the classification results when the Filtered Dysphonic Database was used. The highest accuracy was 71% using SVM with linear kernel. The FFS algorithm selected 5 acoustic features: the std of SPI on nasals, the std and range of IMF entropy measured on high vowels and the mean and std of IMF entropy on spirants. The scenarios where FFS was used outperformed the scenarios where the original 49 acoustic feature set was used in case of SVM with linear kernel, but not with rbf kernel.

The confusion matrix of the 71% classification result can be seen in Table 17. The class precision of the OD class is 72% and 69% in case of FD. The recall for class FD was 74% and 67%

Table 16: Two-class classification results between OD and FD on the Filtered Dysphonic Database.

FFS	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
No	49	SVM	linear kernel	C = 0.125	66%
Yes	5	SVM	linear kernel	C = 1	**71%**
No	49	SVM	rbf kernel	C = 128; $\gamma = 0.0005$	67%
Yes	10	SVM	rbf kernel	C = 2; $\gamma = 0.00781$	66%

Table 17: Confusion matrix using FFS and SVM with linear kernel in cases of females and males together.

	true FD	**true OD**	class precision
pred. FD	61	27	69%
pred. OD	21	55	72%
class recall	74%	67%	

for class OD. This classification accuracy value is more reliable, than if I had done the classification on the dysphonic recordings from the database presented in Section 3.2.1, since it is unaffected by the difference in severity of hoarseness between the two groups.

The results clearly indicate that the separation between the two diseases can be done.

Thesis III. A. [C3] *I showed that the automatic separation between organic and functional dysphonia based on acoustic features is possible with 71% accuracy using SVM with linear kernel on the Hungarian Filtered Dysphonic Database.*

4.4 The automatic classification of the voices of children with dysphonia

The goal of this section is to make an attempt to automatically distinguish healthy voices of children from ones with dysphonia with SVM. The acoustic parameters used in this experiment are presented in Section 3.3.1.

For the binary classification an SVM classifier was used with linear and radial basis function (rbf) kernel. First, all 103 features calculated were used as input, then the FFS algorithm was used to reduce the dimensionality of the input vector. Usually in the case of rbf kernel the hyperparameter C is set to the number of parameters, while γ is set to 1/number of parameters. Leave-one-out cross validation was used in all cases. Classification results are summarized in Table 18.

Table 18: Two-class classification results on the Dysphonic and Healthy Child Speech Database.

FFS	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
No	103	SVM	linear kernel	C = 1	88%
No	103	SVM	rbf kernel	C = 124; γ = 0.008	86%
Yes	8	SVM	linear kernel	C= 1	**93%**
Yes	8	SVM	rbf kernel	C = 10; γ = 0.1	**93%**

Table 19: Confusion matrix using FFS and SVM with linear kernel.

	true HC	**true Dys**	class precision
pred. HC	32	2	94%
pred. Dys	2	23	92%
class recall	94%	92%	

As the table shows that the highest accuracy of 93% was reached using linear and rbf kernel. The features selection algorithm reduced the input dimensionality to 8 acoustic features, while achieving higher accuracy than the case when the starting features were used.

The confusion matrix can be seen in Table 19 when FFS and linear kernel was used. The class precision of the Healthy class is 94% and 92% in case of Dysphonia. The recall for the HC class was 94% and 92% for Dys.

A successful classification should have a symmetric confusion matrix if the weights of the mismatch of the two classes are even. Otherwise, an asymmetric confusion matrix might indicate a biased classifier. The confusion matrix presented with 93% accuracy is as symmetric as it gets. However, as I mentioned earlier, confusion matrix is not symmetric in a medical system. Predicting a healthy child as dysphonic is less bad than predicting a child with dysphonia as healthy.

Furthermore, in my research it is essential that the number of true Dys cases misclassified as HC should be minimized. This happens only twice, resulting in a high 92% recall of class Dys.

We can conclude that input vectors used have great power to distinguish healthy from dysphonic voices of children. From this result, it seems that dysphonia can be better screened at an early stage, but much more data need to be collected to make such statements.

Thesis IV. A. [J2] *I showed that the automatic separation of the voices of healthy children and children with dysphonia is possible by 93% classification accuracy using SVM with linear and rbf kernel on the Dysphonic and Healthy Child Speech Database.*

5 Applicability of my results

The results demonstrate that developing a diagnosis support system which can differentiate dysphonic speech from healthy one is practically feasible. It is important to note, that while the system could be used for pre-screening, giving an exact diagnosis remains the responsibility of the physician.

The system proposed for adults comprises several steps: the speech recordings of the patients are arranged into speech databases (Dysphonic and Healthy Adult Speech Database). The recordings are normalized and segmented on phone level. After selecting the phones to be analysed, acoustic features are extracted and arranged into a feature vector. The feature vector is given to a classifier to perform the binary classification (healthy or unhealthy) in possession of prior knowledge. If the recording is classified as healthy the process stops. If it is classified as unhealthy this practical diagnosis support system would recognize the type of dysphonia, namely: functional or organic dysphonia, whilst performing the estimation of the severity of dysphonia based on a regression module.

Prior knowledge is gained by the procession of a carefully built speech database and optimal classification and regression models described in Sections 3.2.1, 4.1, 4.2, 4.3 and 4.4.

The class (healthy / unhealthy) or the severity of dysphonia is unknown for new speech samples. The preprocessing of the speech record is the same and after the acoustic features are measured on phone level a testing feature vector is constructed that enters a comparative unit, thus the classifications or regression are performed. This process is summarized in Fig 7.

If functional dysphonia can be determined with high probability, with the help of a diagnosis support system, the patient would be directed to a phoniatrist or speech therapist. If the system detected organic dysphonia, the patient would be directed to an otolaryngologists or oncologist. This would save a lot of time and would lead the patient to care as soon as possible. The end system proposed in this study can help young physicians or general practitioners filter out patients with dysphonia more efficiently and determine the severity of dysphonia automatically.

A diagnostic support system for the early recognition of dysphonia in the voices of children would follow the same logic described above, with the difference that the separation of organic from functional causes of dysphonia is not yet possible. Since the classification results in case of children's voice are promising, collecting further speech records to generalize the classification model on a larger dataset is advised. In the long term it is worth developing a tool for the automatic detection of dysphonic voices among children. Mobile devices are suitable for implementing this method and using it in practice. Mobile health applications are usually designed for smart-phones or tablets, on some occasions smart-watches. They allow users to access information when and where they need

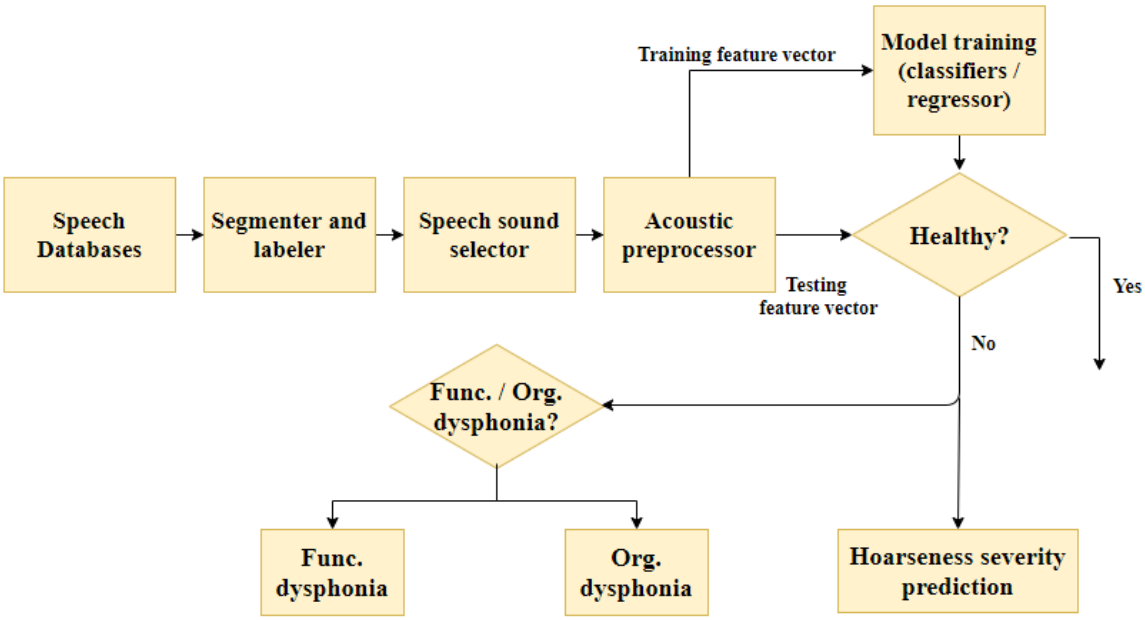

Figure 7: Proposed framework of a practical diagnosis support system for adults.

it; reducing the time wasted with searching for specific data. These devices are cheap, easy-to-use and lightweight. Voice samples, metadata, acoustic feature values and the classifier output can be collected and uploaded to a cloud server. In this way, we can monitor the quality of the children's voice over the long term. The goal is to build a screening system that can be used by pre-school workers. If a child with dysphonic voice can be filtered in time, they will have a better chance of getting a professional help from an ear, nose and throat (ENT) specialist or a speech therapist.

References

[1] R. J. Stachler, D. O. Francis, S. R. Schwartz, C. C. Damask, G. P. Digoy, H. J. Krouse, S. J. McCoy, D. R. Ouellette, R. R. Patel, C. C. W. Reavis *et al.*, "Clinical practice guideline: hoarseness (dysphonia)(update)," *Otolaryngology–Head and Neck Surgery*, vol. 158, no. 1_suppl, pp. S1–S42, 2018.

[2] S. M. Cohen, J. Kim, N. Roy, C. Asche, and M. Courey, "Prevalence and causes of dysphonia in a large treatment-seeking population," *The Laryngoscope*, vol. 122, no. 2, pp. 343–348, 2012.

[3] S. M. Cohen, "Self-reported impact of dysphonia in a primary care population: An epidemiological study," *The Laryngoscope*, vol. 120, no. 10, pp. 2022–2032, 2010.

[4] R. Reiter, T. K. Hoffmann, A. Pickhard, and S. Brosch, "Hoarseness—causes and treatments," *Deutsches Ärzteblatt International*, vol. 112, no. 19, p. 329, 2015.

[5] K. Jones, J. Sigmon, L. Hock, E. Nelson, M. Sullivan, and F. Ogren, "Prevalence and risk factors for voice problems among telemarketers," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 128, no. 5, pp. 571–577, 2002.

[6] J. Long, H. N. Williford, M. S. Olson, and V. Wolfe, "Voice problems and risk factors among aerobics instructors," *Journal of Voice*, vol. 12, no. 2, pp. 197–207, 1998.

[7] E. Smith, H. L. Kirchner, M. Taylor, H. Hoffman, and J. H. Lemke, "Voice problems among teachers: differences by gender and teaching characteristics," *Journal of Voice*, vol. 12, no. 3, pp. 328–334, 1998.

[8] T. Davids, A. M. Klein, and M. M. Johns III, "Current dysphonia trends in patients over the age of 65: is vocal atrophy becoming more prevalent?" *The Laryngoscope*, vol. 122, no. 2, pp. 332–335, 2012.

[9] N. Bhattacharyya, "The prevalence of pediatric voice and swallowing problems in the united states," *The Laryngoscope*, vol. 125, no. 3, pp. 746–750, 2015.

[10] M. C. Duff, A. Proctor, and E. Yairi, "Prevalence of voice disorders in african american and european american preschoolers," *Journal of Voice*, vol. 18, no. 3, pp. 348–353, 2004.

[11] P. N. Carding, S. Roulstone, K. Northstone, A. S. Team *et al.*, "The prevalence of childhood dysphonia: a cross-sectional study," *Journal of Voice*, vol. 20, no. 4, pp. 623–630, 2006.

[12] E.-M. Silverman and C. H. Zimmer, "Incidence of chronic hoarseness among school-age children," *Journal of Speech and Hearing Disorders*, vol. 40, no. 2, pp. 211–215, 1975.

[13] M. Rosa and M. Behlau, "Mapping of vocal risk in amateur choir," *Journal of Voice*, vol. 31, no. 1, pp. 118–e1, 2017.

[14] J. Guss, B. Sadoughi, B. Benson, and L. Sulica, "Dysphonia in performers: toward a clinical definition of laryngology of the performing voice," *Journal of Voice*, vol. 28, no. 3, pp. 349–355, 2014.

[15] K. Verdolini and L. O. Ramig, "Occupational risks for voice problems," *Logopedics Phoniatrics Vocology*, vol. 26, no. 1, pp. 37–46, 2001.

[16] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *Journal of Speech, Language, and Hearing Research*, 2004.

[17] E. Smith, J. Lemke, M. Taylor, H. L. Kirchner, and H. Hoffman, "Frequency of voice problems among teachers and other occupations," *Journal of voice*, vol. 12, no. 4, pp. 480–488, 1998.

[18] F. S. G. Fortes, R. Imamura, D. H. Tsuji, and L. U. Sennes, "Profile of voice professionals seen in a tertiary health center," *Brazilian journal of otorhinolaryngology*, vol. 73, no. 1, pp. 27–31, 2007.

[19] D. Isetti and T. Meyer, "Workplace productivity and voice disorders: A cognitive interviewing study on presenteeism in individuals with spasmodic dysphonia," *Journal of Voice*, vol. 28, no. 6, pp. 700–710, 2014.

[20] "American speech-language-hearing association - voice disorders," https://www.asha.org/practice-portal/clinical-topics/voice-disorders/, accessed: 2020-04-03.

[21] R. Jani, S. Jaana, L. Laura, and V. Jos, "Systematic review of the treatment of functional dysphonia and prevention of voice disorders," *Otolaryngology—Head and Neck Surgery*, vol. 138, no. 5, pp. 557–565, 2008.

[22] J. Wendler, A. Rauhut, and H. Kruger, "Classification of voice qualities," *Journal of Phonetics*, vol. 14, no. 3-4, pp. 483–488, 1986.

[23] M. Ptok, C. Schwemmle, C. Iven, M. Jessen, and T. Nawka, "On the auditory evaluation of voice quality," *HNO*, vol. 54, no. 10, pp. 793–802, 2006.

[24] J. D. Evans, *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co, 1996.

[25] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] G. Horváth, "Neurális hálózatok és műszaki alkalmazásaik," *Műszaki Kiadó*, 1998.

[27] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[28] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in neural information processing systems*, 1997, pp. 155–161.

[29] G. Kiss and K. Vicsi, "Mono-and multi-lingual depression prediction based on speech processing," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 919–935, 2017.

[30] V. Klára, "Sampa computer readable phonetic alphabet," 2008.

[31] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[32] P. Dighe, A. Asaei, and H. Bourlard, "On quantifying the quality of acoustic models in hybrid *DNN-HMM ASR*," *Speech Communication*, 2020.

[33] P. Roach, S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel *et al.*, "Babel: An eastern european multi-language database," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1892–1893.

[34] K. Vicsi, A. Kocsor, C. Teleki, and L. Tóth, "Hungarian speech database for computer-using environments in offices," in *Proc. 2nd Hungarian Conf. on Computational Linguistics*, 2004, pp. 315–318.

[35] C. Teleki, V. Szabolcs, T. S. Levente, and V. Klára, "Development and evaluation of a hungarian broadcast news database," in *Forum Acusticum*, 2005.

[36] 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. CONF. IEEE Signal Processing Society, 2011.

[37] Y.-R. Chien, M. Borskỳ, and J. Guðnason, "Objective severity assessment from disordered voice using estimated glottal airflow." in *Interspeech*, 2017, pp. 304–308.

[38] I. Laaridh, W. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Interspeech*, 2017.

[39] T. Law, J. H. Kim, K. Y. Lee, E. C. Tang, J. H. Lam, A. C. van Hasselt, and M. C. Tong, "Comparison of rater's reliability on perceptual evaluation of different types of voice sample," *Journal of Voice*, vol. 26, no. 5, pp. 666–e13, 2012.

[40] N. Adiga, C. Vikram, K. Pullela, and S. M. Prasanna, "Zero frequency filter based analysis of voice disorders." in *Interspeech*, 2017, pp. 1824–1828.

[41] M. Markaki, Y. Stylianou, J. D. Arias-Londoño, and J. I. Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5162–5165.

[42] F. Kazinczi, K. Mészáros, and K. Vicsi, "Automatic detection of voice disorders," in *International Conference on Statistical Language and Speech Processing*. Springer, 2015, pp. 143–152.

[43] V. Klára, I. Viktor, and M. Krisztina, "Voice disorder detection on the basis of continuous speech," in *5th European Conference of the International Federation for Medical and Biological Engineering*. Springer, 2011, pp. 86–89.

[44] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: a preliminary study," in *2017 international conference and workshop on bioinspired intelligence (IWOBI)*. IEEE, 2017, pp. 1–4.

[45] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, 2018.

[46] H. Wu, J. J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks." in *Interspeech*, vol. 2018, 2018.

[47] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[48] L. Salhi, M. Talbi, and A. Cherif, "Voice disorders identification using hybrid approach: Wavelet analysis and multilayer neural networks," *World Academy of Science, Engineering and Technology*, vol. 45, no. 21, pp. 330–339, 2008.

[49] V. Srinivasan, V. Ramalingam, and P. Arulmozhi, "Artificial neural network based pathological voice classification using mfcc features," *International Journal of Science, Environment and Technology*, vol. 3, no. 1, pp. 291–302, 2014.

[50] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech & Language*, vol. 4, no. 2, pp. 127–165, 1990.

[51] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[52] T. Lee, Y. Liu, Y. T. Yeung, T. K. Law, and K. Y. Lee, "Predicting severity of voice disorder from *DNN-HMM* acoustic posteriors." in *Interspeech*, 2016, pp. 97–101.

[53] Y. Liu, T. Lee, P. Ching, T. K. Law, and K. Y. Lee, "Acoustic assessment of disordered voice with continuous speech based on utterance-level *ASR* posterior features." in *Interspeech*, 2017, pp. 2680–2684.

[54] E. Barth, *Einführung in die Physiologie, Pathologie und Hygiene der menschlichen Stimme.* G. Thieme, 1911.

[55] H. Stern, "Klinik und therapie der krankheiten der stimme," *Mschr Ohrenheilk*, vol. 58, pp. 1–53, 1924.

[56] D. Weiss, "Der begriff des funktionellen mit besonderer berücksichtigung der sprach-und stimmheilkunde," *Mschr Ohrenheilk*, vol. 68, pp. 830–832, 1934.

[57] H. Gundermann, *Die Berufsdysphonie: Nosologie der Stimmstörungen in Sprechberufen unter besonderer Berücksichtigung der sogenannten Lehrerkrankheit.* Thieme, 1970.

Publications

International journals

[J1] Szaszák, Gy., **Tulics, M. G.**, & Tündik, M. Á., "Analyzing FO discontinuity for speech prosody enhancement," *Acta Univ. Sapientiae Elect. Mech. Eng*, vol. 6, no. 1, pp. 59–67, 2014. (6/3=2 points)

[J2] **Tulics, M. G.**, & Vicsi, K., "Automatic classification possibilities of the voices of children with dysphonia," *Infocommunications Journal* Vol. X. No.3. pp. 30-36., 7 p. 2018. (4 points)

[J3] Kovács, A., **Tulics, M. G.**, Tündik, M. Á., Moró, A., Gróf, A., "Magmanet: Ensemble of 1d convolutional deep neural networks for speaker recognition in hungarian," *Phonetician*, vol. 115, pp. 72–86, 2018. (6/5=1.2 points)

[J4] **Tulics, M. G.**, & Vicsi, K. (2019). "The automatic assessment of the severity of dysphonia," International Journal of Speech Technology, 1-10. (6 points)

[J5] Szántó, D., Jenei, A. Z., **Tulics, M. G.**, & Vicsi, K., "Developing a Noise Awareness Rising Web Application within the "Protect your Ears" project," *Infocommunications Journal* 2020. Accepted

Hungarian journals

[J6] Sztahó, D, Kiss, G, **Tulics, M G**, Czap, L, Vicsi, K, "Számítógéppel támogatott prozódiaoktató program," *Alkalmazott Nyelvészeti Közlemények* 9 : 1 pp. 144-153. , 10 p. (2014) (2/4=0.5 points)

International conferences

[C1] **Tulics, M. G.**, Kazinczi, F., & Vicsi, K., "Statistical analysis of acoustical parameters in the voice of children with juvenile dysphonia," In *International Conference on Speech and Computer* (pp. 667-674). Springer, Cham. 2016. (3/2=1.5 points)

[C2] **Tulics, M. G.**, Szaszák, Gy., Mészáros, K. & Vicsi, K., "Artificial Neural Network and SVM based Voice Disorder Classification," In *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE. 2019. (3/3=1 points)

[C3] **Tulics, M. G.**, Lavati, L. J., Mészáros, K. & Vicsi, K., "Possibilities for the automatic classification of functional and organic dysphonia," In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019. (3/3=1 points)

[C4] **Tulics, M. G.**, & Vicsi, K., "Phonetic-class based correlation analysis for severity of dysphonia," In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 000021-000026). IEEE. 2017. (3 points)

[C5] Kiss, G., **Tulics, M. G.**, Sztahó, D., Esposito, A., & Vicsi, K., "Language independent detection possibilities of depression by speech," In *Recent advances in nonlinear speech processing* (pp. 103-114). Springer, Cham. 2016. (3/4=0.75 points)

[C6] Sztahó, D., **Tulics, M. G.**, Vicsi, K., & Valálik, I., "Automatic estimation of severity of Parkinson's disease based on speech rhythm related features," In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 000011-000016). IEEE. 2017. (3/3=1 points)

[C7] Sztahó, D., Kiss, G., **Tulics, M. G.**, & Vicsi, K., "Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features," In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)* (pp. 1-4). IEEE. 2018. (3/3=1 points)

[C8] Sztahó, D., Kiss, G., **Tulics, M. G.**, Dér-Hajduska,B. & Vicsi, K., "Automatic discrimination of several types of speech pathologies," In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019. (3/4=0.75 points)

[C9] **Tulics, M. G.**, Szaszák, Gy., Mészáros, K. & Vicsi, K., "Using ASR Posterior Probability and Acoustic Features for Voice Disorder Classification," In *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE. 2020. Accepted

Hungarian conferences

[C10] **Tulics, M. G.**, Jászai, H., & Vicsi, K., "A diszfónia súlyosságának automatikus becslése, a szakértői értékelések szubjektív jellegének figyelembevételével," In: *Vincze, Veronika (szerk.) XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)* Szeged, Magyarország : Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 206-218. , 13 p. 2018. (1/2=0.5 points)

Total publication score: 24.2 points

Independent citations: 13