Characterisation of voice quality of Parkinson's disease using differential phonological posterior features

Milos Cernak[a,*], Juan Rafael Orozco-Arroyave[b,e], Frank Rudzicz[c], Heidi Christensen[d], Juan Camilo Vásquez[b,e], Elmar Nöth[e]

[a] *Idiap Research Institute, Martigny, Switzerland*
[b] *Universidad de Antioquia Medellín, Colombia*
[c] *University of Toronto, Canada*
[d] *University of Sheffield, UK*
[e] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

Abstract

Change in voice quality (VQ) is one of the first precursors of Parkinson's disease (PD). Specifically, impacted phonation and articulation causes the patient to have a breathy, husky-semiwhisper and hoarse voice.

A goal of this paper is to characterize a VQ spectrum – the composition of non-modal phonations – of voice in PD. The paper relates non-modal healthy phonations: breathy, creaky, tense, falsetto and harsh, with disordered phonation in PD. First, statistics are learned to differentiate the modal and non-modal phonations. Statistics are computed using phonological posteriors, the probabilities of phonological features inferred from the speech signal using a deep learning approach. Second, statistics of disordered speech are learned from PD speech data comprising 50 patients and 50 healthy controls. Third, Euclidean distance is used to calculate similarity of non-modal and disordered statistics, and the inverse of the distances is used to obtain the composition of non-modal phonation in PD. Thus, pathological voice quality is characterised using healthy non-modal voice quality "base/eigenspace". The obtained results are interpreted as the voice of an average patient with PD and can be characterised by the voice quality spectrum composed of 30% breathy voice, 23% creaky voice, 20% tense voice, 15% falsetto voice and 12% harsh voice. In addition, the proposed features were applied for prediction

[*] Corresponding author
Email address: milos.cernak@idiap.ch (Milos Cernak)

of the dysarthria level according to the Frenchay assessment score related to the larynx, and significant improvement is obtained for reading speech task.

The proposed characterisation of VQ might also be applied to other kinds of pathological speech.

Keywords: Phonological features, non-modal phonation, Parkinson's disease

1. Introduction

Speech of hypokinetic dysarthria in Parkinson's disease (PD) is characterised by hypokinesia (rigid, less motion describing decreased range and frequency of movement) of the vocal folds and articulators. Besides of impacted prosody and articulation, phonation is impacted by incomplete vocal fold adduction. Clinicians, otolaryngologists and speech-language pathologists, consider hoarseness – a rough quality of voice – as a basic symptom of a voice disorder in PD. When hoarse, the voice may sound breathy, raspy, or strained, and if this abnormal/pathological voice quality is accompanied with relatively constant loudness and pitch deviations, it is diagnosed as Parkinsonian dysphonia (Aronson and Bless, 2011).

Healthy subjects may also produce speech sounds of different voice quality based on different modes of vibration of the vocal folds. Laver (1980) defines the term of voice quality in a broad sense as the characteristic auditory colouring of an individual speaker's voice, and not just in a narrow sense coming from the laryngeal activity. The neutral mode phonation, often used in **modal voice**, is one against which the other modes can be contrastively described, also called non-modal phonations. Ladefoged and Johnson (2014) describe four basic states of the glottis (which is defined as the space between the vocal folds). The position of the vocal folds is adjusted by the arytenoid cartilages placed toward the back. In (i) a voiced sound, the vocal folds are close together (adducted) and vibrating, whereas in (ii) a voiceless sound, they are pulled apart (abducted). If there is considerable airflow, the abducted vocal folds will be set vibrating – flapping in the airstream – producing what is called (iii) **breathy voice**, or murmur. Alternatively, breathy voice is produced with the vocal folds apart only between the arytenoid cartilages in the lower (posterior) part. If the arytenoid cartilages are tightly together, so that the vocal folds can vibrate only at the anterior end, (iv) **creaky voice** is produced. Creaky-voiced sounds may also be called laryngealized. Besides these basic non-modal phonation, Laver (1980) defines **tense**, **harsh** and **falsetto** phonations. Such voice qualities impact the production of the speech sounds, and we hypothesise that these changes might be captured by changes in phonological features.

The goal of this paper is to present a study on the production of speech sounds with healthy non-modal phonation, and project its non-modal statistics to analyse disordered production of speech sounds with pathological phonation. This approach might help to aleviate a problem of missing data in research of pathological speech. Voice quality of the speech sounds can be characterised by phonological features (Cernak et al., 2017b), and the current work proposes to use differential phonological posterior features (between modal and non-modal, and between healthy and disordered phonations) for characterisation of both healthy non-modal and parkinsionian phonations. Comparing to the work of Cernak et al. (2017b), the novel aspects of this paper is in using pathological speech and characterization of pathological voice quality using healthy non-modal voice quality "base/eigenspace". An Euclidean distance between the non-modal and disordered phonation characterisations quantifies the composition of non-modal voice qualities in PD. This characterisation of non-modal phonation in PD is novel, and shows objective quantification of voice quality using phonological features not investigated in previous approaches.

For studying speech with non-modal phonation, the read-VQ database of Kane (2012) is used, the recording of which was inspired by prototype voice quality examples produced by Laver (1980). Laver's recordings are considered as recordings of non-modal phonation with excellent quality, however only one utterance per phonation type is available, and thus they are speaker-specific. The read-VQ database contains recordings from four speakers. The database covers five different non-modal phonations: falsetto, creaky, harshness, tense and breathiness. For studying speech with pathological phonation, the Colombian-Spanish database (Orozco-Arroyave et al., 2014) is used, which contains speech recordings of 50 patients with PD and 50 healthy controls (HC).

The structure of the paper is as follows: Section 2.1 gives an overview of the non-modal (healthy) and pathological (Parkinsonian) phonation types considered in this work. Section 3 introduces differential phonological posterior (DPP) features used in further characterisation of VQ. Section 4 describes experimental setup and evaluation databases, and Section 5 presents results and their validation. Finally, Section 6 concludes the paper.

2. Voice quality of Parkinson's disease

2.1. Non-modal (healthy) phonation

Different modes of vibration of the vocal folds contribute significantly to VQ. The modal (periodic) phonation can be contrastively described against the other modes, also called non-modal (aperiodic) phonations.

Recent work on non-modal phonation focuses on detection (Drugman et al., 2014), analysis (Malyska, 2008; Malyska et al., 2011) and synthesis (Bangayan et al., 1997) of speech with non-modal phonation. Modern computational paralinguistics tries to 1) get rid of non-modal phonation, or 2) model it, for example, for classification purposes (Schuller and Batliner, 2013). Non-modal phonation is also studied in sociolinguistics. For example, creaky and falsetto phonations are used more commonly by women (Anderson et al., 2014; Podesva, 2007).

Breathy and creaky voices belong to the most studied non-modal phonation types. In breathy phonation, the vibration of the vocal folds is accompanied by aspiration noise, which causes a higher first formant bandwidth and a missing third formant (Klatt and Klatt, 1990) due to steeper spectral tilt (Hanson, 1997). In creaky phonation (also referred to as vocal fry, laryngealisation), secondary vibrations occur with lower fundamental frequencies.

Tense voice is produced with higher degree of overall muscular tension involved in the whole vocal tract. The higher tension of the vocal folds does not result in irregularities that are seen in harsh voice. It is characterised by richer harmonics in higher frequencies due to a less steep spectral tilt. Harsh voice is a result of very high muscular tension at the laryngeal level. Pitch is irregular and low, and the speech spectrum contains more noise.

Falsetto voice is the most different with respect to modal voice (Laver, 1980). The voice is produced with thin vocal folds, that results in a higher pitch voice with a steeper spectral slope.

2.2. Pathological (Parkinsonian) phonation

Auditory-perceptual evaluation of disordered VQ is the most commonly used clinical assessment method, and is considered by clinicians as the “gold standard” for documenting voice impairment severity (Kreiman et al., 1993). Describing a particular voice as breathy and rough, for example, is likely to be more easily interpreted by a wide range of people than a description that specifies the noise-to-harmonic ratio associated with that voice (Oates, 2009). Moreover auditory-perceptual evaluation is cheap and practical. Perceptual analysis is used with the human auditory perceptual system, often in combination with an external rating system, such as the GRBAS protocol (Hirano, 1981) developed by the Japanese Society of Logopedics and Phoniatrics. The GRBAS protocol contains 4-point scales for grade (overall severity), roughness, breathiness, asthenia (lack of vocal power), and strain.

On the other hand, the perceptual evaluation has been characterized by questionable validity and poor reliability, adding further analysis error via measurement and scaling issues (Aronson and Bless, 2011), and missing consensus on stimulus categories (Barsties and De Bodt, 2015). At present, the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V)[1], containing six primary perceptual parameters (overall severity, roughness, breathiness, strain, pitch, and loudness), is undergoing field testing, and experimental data on its validity and reliability are forthcoming.

Acoustic analysis is widely employed in clinical and research settings, and focuses on analysis of parkinsonian speech that provides objective measures of vocal function, such as fundamental frequency, signal amplitude, jitter, shimmer, noise-to-harmonic ratios, voice onset time and glottal leakage, and last but not least the spectral features such as spectral tilt (J Holmes et al., 2000; Little et al., 2009; Rusz et al., 2011; Bauer et al., 2011). Parkinsonian speech is characterised by higher jitter (more roughness), higher shimmer, descreased pitch range, shorter maximum phonation time and slower diadochokinetic (articulation) rate (Darley et al., 1969). However, acoustic measures cannot be applied to more severe disordered voices due to their nonlinear and non-Gaussian random properties (Little et al., 2007), that limits their clinical usefulness.

There is a considerable amount of literature on objective perceptual evaluation based on acoustic and aerodynamic speech production characteristics. For example, Wuyts et al. (2000) propose a Dysphonia Severity Index, constructed from highest frequency, lowest intensity, maximum phonation time and jitter. Bhuta et al. (2004); Maryn et al. (2009) provide detailed studies of correlation of acoustic measurements with perceived voice quality. Recent methods include in objective perceptual evaluation also spectral/cepstral features, such as spectrum slope and tilt (Maryn et al., 2010), and cepstral peak prominence (Awan et al., 2009).

[1] `http://www.asha.org/uploadedFiles/members/divs/D3CAPEVprocedures.pdf`

3. Differential phonological posteriors

The probabilities of phonological features inferred from the speech signal – phonological posteriors – can be reliably estimated using a deep learning approach (Cernak et al., 2015). This extraction processes is further called phonological analysis. In this work, the Sound Patterns of English (SPE) feature set of Chomsky and Halle (1968) is used. Motivation to this older phonological system was that (i) it takes the articulatory production mechanism as the underlying principle of phoneme organisation (and thus allows easier interpretation of obtained results), and (ii) SPE assumes that the flat, unstructured binary feature specifications are language independent and characterise the set of possible phonemes in languages of the world (and thus is more suitable for studies with more languages like described in this paper). The mapping from phonemes to SPE phonological classes is taken from Cernak et al. (2017a). The distribution of the phonological labels is non-uniform, driven by mapping different numbers of phonemes to the phonological classes.

Phonological analysis starts with a short-term analysis of speech, which consists of converting the speech signal into a sequence of acoustic feature vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\}$. Each $\mathbf{x}_n$ is also known as an acoustic frame or just frame, and can be composed by the conventional Mel frequency cepstral coefficients (MFCC). N is the number of frames and frames are equally spaced in time.

Then, K phonological probabilities z_n^k are estimated for each frame. Each probability is computed independently by using a binary classifier based on deep neural network (DNN) and trained with one class versus the rest. Finally, the acoustic feature observation sequence X into a sequence of phonological vectors $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_n, \ldots, \mathbf{z}_N\}$. Each vector $\mathbf{z}_n = [z_n^1, \ldots, z_n^k, \ldots, z_n^K]^\top$ consists of phonological class posterior probabilities $z_n^k = p(c_k|x_n)$ of K phonological features (classes) c_k. The a posteriori estimates $p(c_k|x_n)$ are $0 \leq p(c_k|x_n) \leq 1, \forall k$, and $\max \sum_{k=1}^{K} p(c_k|x_n) = K$.

The matrix of posteriors Z consists of N rows, indexed by the processed speech frames, and K columns. The following analysis is done on non-silence speech frames of the evaluation data:

$$\mu_k = \frac{1}{N_s} \sum_{n=1}^{N_s} p(c_k|x_n), \forall n \Longleftrightarrow p(c_{\mathrm{SIL}}|x_n) < 0.5, \tag{1}$$

where c_{SIL} is a posterior probability of silence class being observed, and N_S is the number of non-silence frames. The probability of c_{SIL} is computed as for the other phonological classes (i.e., the silence versus the rest) but it is only taken into account when computing each μ_k. The statistics μ_k is calculated for different "contrastive" data groups, such as data with modal vs. data with non-modal phonations, and data from healthy speakers vs. data from pathological speakers.

Differential phonological posterior (DPP) features are obtained by mean normalization of contrastive data:

$$\begin{aligned}\Delta\mu_{kl}^{NM} &= \mu_{kl}^{\text{non-modal}} - \mu_{kl}^{\text{modal}},\\ \Delta\mu_{k}^{P} &= \mu_{k}^{\text{PD}} - \mu_{k}^{\text{HC}}.\end{aligned} \tag{2}$$

Thus, the non-modal mean posteriors are normalized by modal means that yields the normalized statistics $\boldsymbol{\Delta\mu_l^{NM}} = [\Delta\mu_{1l}^{NM}, \ldots, \Delta\mu_{kl}^{NM}, \ldots, \Delta\mu_{Kl}^{NM}]^\top$ for $l \in L$ non-modal phonations, and PD posteriors are normalized by means from healthy speakers that yields pathological (Parkinsonian) statistics $\boldsymbol{\Delta\mu^P} = [\Delta\mu_1^P, \ldots, \Delta\mu_k^{NM}, \ldots, \Delta\mu_K^{NM}]^\top$.

Finally, similarity of non-modal phonation and pathological speech is calculated as the Euclidean distance:

$$q_l = \|\boldsymbol{\Delta\mu^P} - \boldsymbol{\Delta\mu_l^{NM}}\|, \tag{3}$$

for $l \in L$ non-modal phonations, where q_l represents a similarity of the l-th non-modal phonation with VQ in PD. The Euclidean distance was already successfully used as a similarity measure between VQ characterisations in forensic speaker comparison (San Segundo et al., 2017).

The normalization of the mean posteriors by the posterior features from the modal or healthy speakers is conceptually similar to likelihood ratio test in speaker recognition (Hansen and Hasan, 2015), where likelihoods from the speaker model are subtracted by likelihoods obtained from the background/world model. In the DPP features, the background models represent the modal phonation and healthy speakers.

4. Experimental setup

4.1. Training data

The phonological analyser is trained on the Wall Street Journal (WSJ0 and WSJ1) continuous speech recognition corpora (Paul and Baker, 1992). This training database consists primarily of read speech using a close-talking

Sennheiser HMD414. The *si_tr_s_284* set of 37 514 utterances was used, split into 90% training and 10% cross-validation sets. Titze (1995) recommends the WSJ database to be used in acoustic analysis research of pathological speech. In addition, Cernak et al. (2015) introduced a deep learning approach using WSJ data to achieve high classification accuracy of phonological features.

4.2. Evaluation data

Prototype voice quality examples produced by Laver (1980) and the read-VQ database of Kane (2012) were used to obtain characterisation of modal and non-modal phonation. Audio of the read-VQ database was recorded at 44.1 kHz using high quality recording equipment: a B&K 4191 free-field microphone and a B&K 7749 pre-amplifier. The microphone was placed at a distance of approximately 30 cm from the speaker and participants were asked to keep this distance as constant as possible throughout the recording session. Recordings were subsequently downsampled to 16 kHz.

The read-VQ database contains 4 speakers (2 males and 2 females) who were asked to read 17 sentences in six different phonation types: modal, breathy, tense, harsh, creaky, and falsetto. Participants were given prototype voice quality examples, produced by John Laver and John Kane, and were asked to practise producing them before coming to the recording session. For the recordings, participants were asked to produce the strong versions of each phonation type and to maintain it throughout the utterance. During the recording session, participants were asked to repeat the sentence if it was deemed necessary. The sentences were chosen from the phonetically balanced sentences in the TIMIT corpus (Garofolo et al., 1993), four of which contained all-voiced sounds. 451 sentences were chosen to obtain a wide phonetic coverage, as it is likely that it can be very difficult for speakers to maintain a constant type of phonation over a long utterance. The recordings with modal phonation were 2.2 minutes long, and the remaining recordings with non-modal phonation were 2.0 minutes long each. The read-VQ data was used for estimation of non-modal DPP features $\boldsymbol{\Delta\mu_l^{NM}}$.

Speech recordings from the HC and PD patients were obtained from the database provided by Orozco-Arroyave et al. (2014). This database contains speech recordings of 50 patients with PD and 50 HCs sampled at 44.1 kHz with 16 resolution-bits. The recordings were captured in noise controlled conditions, in a sound proof booth. All of the speakers are balanced by gender and age. All of the patients were diagnosed and labeled by neurologist experts. The speech samples were recorded with the patients in the ON-state, i.e., no more than 3 hours after the morning medication. None of the people in the HC group has a history of symptoms related to PD or any other kind of neurological disorder. The HC and PD data was used for estimation of parkinsonian DPP features $\boldsymbol{\Delta\mu_l^P}$. It is worth to note that the training data of phonological analyser contains English recordings, whereas the HC and PD data contain Columbian-Spanish recordings. It is assumed that phonological features are language independent, and in addition, Cernak et al. (2016) showed effective cross-language usage of phonological posteriors, for English training data and French evaluation data, and vice versa.

PD data contains several different speech tasks comprising of isolated and running speech: 24 isolated words, the 'pataka' speech task consisting of repeating /pataka,petaka,pakata/ speech production, 10 sentences, one read text with 36 words, and a monologue with an average duration of 44.86 s. All data was used for experiments described in Section 5.

4.3. Training of phonological analysis

The open-source phonological vocoding platform developed by Cernak and Garner (2016) was used to perform phonological analysis and synthesis. Briefly, the platform is based on cascaded speech analysis and synthesis that works internally with the phonological speech representation. In the phonological analysis part, phonological posteriors are estimated directly from the speech signal by DNNs. Binary (Yu et al., 2012) or multi-valued classification (Stouten and Martens, 2006; Rasipuram and Magimai.-Doss, 2011) may be used. In the latter case, the phonological classes are grouped together based on place or manner of articulation. The binary classification approach is used in this work, and thus each DNN determines the probability of a particular phonological class.

To train the DNNs for phonological analysis, a phoneme-based automatic speech recognition system is first trained using Mel frequency cepstral coefficients (MFCC) as acoustic features. The phoneme set comprises 40 phonemes (including silence) defined by the CMU pronunciation dictionary. The three-state, cross-word triphone models were trained with the HMM-based speech synthesis system (HTS) variant (Zen et al., 2007) of the Hidden Markov Model Toolkit (HTK) on the 90% subset of the WSJ data. The remaining 10% subset was used for cross-validation. The triphone models are tied with decision tree state clustering based on the minimum description length (MDL) criterion (Shinoda and Watanabe, 1997). The MDL criterion allows an unsupervised determination of the number of states. The trained model had 12 685 tied states, and each is modeled with a Gaussian mixture model consisting of 16 Gaussians.

The acoustic models were used to get phonetic alignment. Each phoneme was mapped to the 13 SPE phonological classes or the one silence class, and thus 14 DNNs were trained as phonological/silence analysers using the frame alignment with a particular phonological/silence label scheme that took two binary values: the phonological/silence class exists for the aligned phoneme or not. In other words, the two DNN outputs correspond to the target class vs. the rest.

Some classes might seem to have unbalanced training data; for example, the two labels for the nasal class are associated with the speech samples from just 3 nasal phonemes /m/, /n/, and /ŋ/, and with the remaining 36 (non-nasal) phonemes. However, this split is necessary to appropriately train a discriminative classifier, as all the remaining phonemes convey information about all different phonological classes. Each DNN was trained on the whole training set. Several DNN sizes were tested, from 3 to 6 hidden layers with 500 to 2000 neurons. Finally, the selected size of the DNNs was $351 \times 1024 \times 1024 \times 1024 \times 2$ neurons, is a balance between the training time and the performance. Sigmoid activation functions were used in the hidden layers. The input feature vectors consisted of Energy plus 12 MFCC (13 parameters) with the first and second time derivatives. The temporal context from 7 to 11 successive frames was tested with no particular performance increase, so the temporal context of 9 frames was used for the training.

The parameters were initialized using deep belief network pre-training following the single-step contrastive divergence (CD-1) procedure of Hinton et al. (2006). The DNNs with the softmax output function were then trained using a mini-batch based stochastic gradient descent algorithm with the cross-entropy cost function of the KALDI toolkit (Povey et al., 2011).

5. Results

5.1. Analysis of non-modal phonation

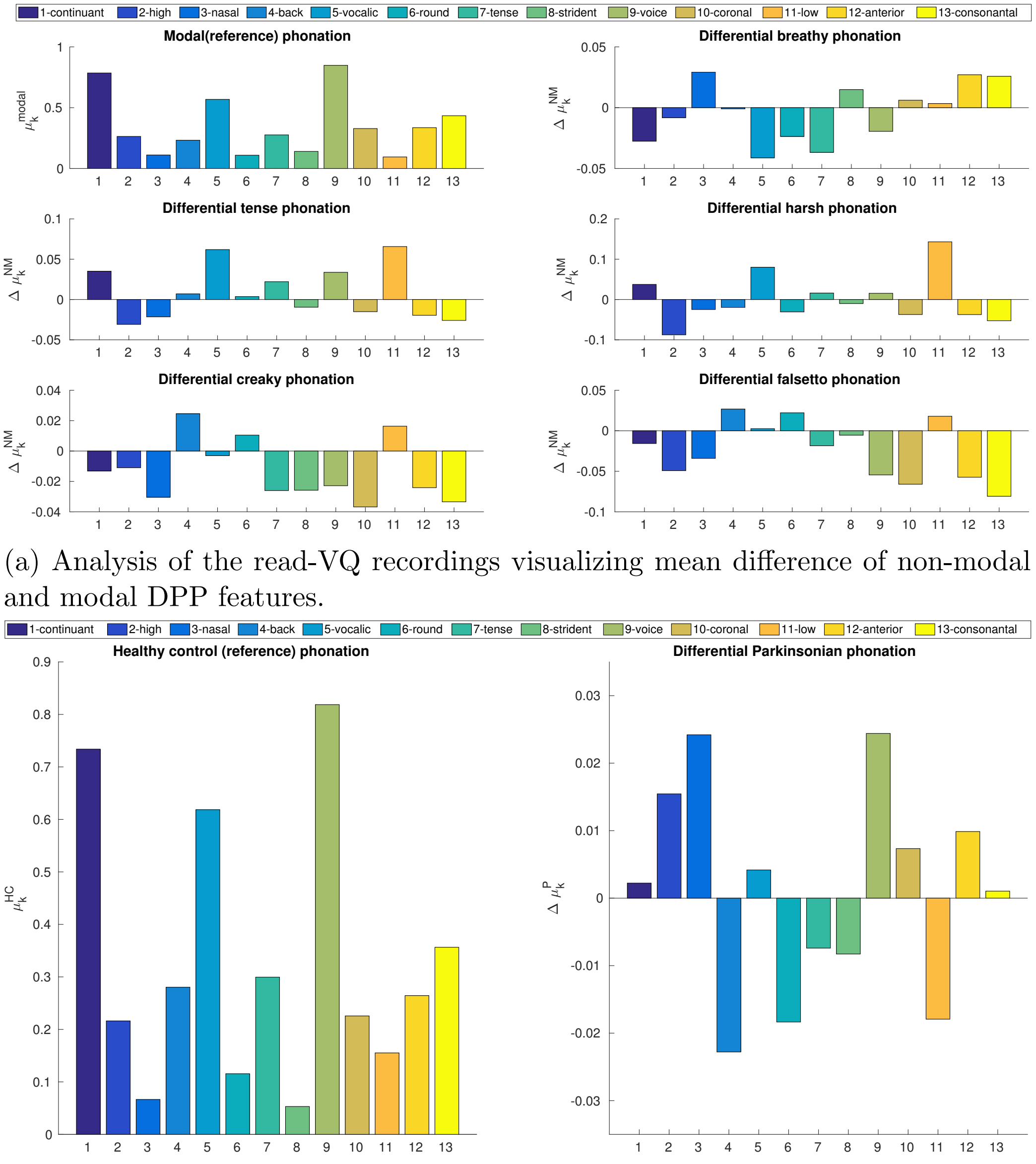
Fig. 1a shows the analysis of the read-VQ evaluation data. Table 2 shows the results of further statistical analysis performed by using the two-sample t-test without assuming equal variance, that was carried out to study the differences between speech with modal and non-modal phonations. The significance of the test also allows for the determination of invariant phonological features, listed in Table 2.

Table 1: *Statistical significance (p -values) of difference between* μ_{kl}^{modal} *and* $\mu_{kl}^{non\text{-}modal}$ *for* $k \in K = 13$ *SPE features and* $l \in L = 5$ *non-modal phonations. For of the level of significance* $\alpha = 0.001$*, the bold items represent the invariance of a particular pair of the SPE feature and non-modal phonation, i.e., the SPE features unaffected by non-modal phonations, where statistical significance of differences is* $p > \alpha$*. The other items shown by '–' represent the SPE features affected by non-modal phonation, with significance* $p < \alpha$*.*

SPE/Phonation	Breathy	Creaky	Tense	Harsh	Falsetto
Continuant	–	**0.0042**	–	–	–
High	**0.0958**	**0.0267**	–	–	–
Nasal	–	–	–	–	–
Back	**0.8261**	–	**0.1308**	–	–
Vocalic	–	**0.5948**	–	–	**0.6657**
Round	–	**0.0031**	**0.3114**	–	–
Tense	–	–	–	**0.0012**	–
Strident	–	–	**0.0251**	**0.0208**	**0.2198**
Voice	–	–	–	–	–
Coronal	**0.2413**	–	**0.0041**	–	–
Low	**0.2902**	–	–	–	–
Anterior	–	–	–	–	–
Consonantal	–	–	–	–	–

According to Table 2, the [Strident] phonological feature is more invariant – "resistant" – to non-modal phonation, whereas the [Nasal], [Voice], [Anterior] and [Consonantal] features are heavily impacted (they are not invariant for any phonation type). The [Strident] feature is significantly different only in creaky phonation, which indicates its usefulness, for example, in creaky voice detection. On the contrary, the invariant [Tense] feature might indicate harsh, and the invariant [Low] feature may indicate breathy phonation.

The number of invariant features also indicates the impact of non-modal phonation on phonological features. While breathy, creaky and tense phonations keep 4 invariant features, harsh and falsetto phonation keep only 2 invariant features.



(a) Analysis of the read-VQ recordings visualizing mean difference of non-modal and modal DPP features.

(b) Analysis of the HC and PD recordings visualizing mean difference of PD and HC DPP features.

Figure 1: *Mean modal/HC SPE posteriors μ_k (top-left figures) and differentials $\Delta\mu_k$ of non-modal/PD phonations with respect to the modal/HC voice.*

Table 2: *The impact of non-modal phonation on phonological features, measured as a positive (+) or negative (−) difference between the mean phonological posteriors of speech with modal phonation, and the mean phonological posteriors with non-modal phonation. The three features with the greatest differences are listed. Invariance is concluded based on statistics in Table 1.*

Phonation	Invariant features	Most different features
Breathy	High, Back, Coronal, Low	−Vocalic, −Tense, +Nasal
Creaky	Continuant, High, Vocalic, Round	−Coronal, −Consonantal, −Nasal
Tense	Back, Round, Strident, Coronal	+Low, +Vocalic, +Continuant
Harsh	Strident, Tense	+Low, −High, +Vocalic
Falsetto	Strident, Vocalic	−Consonantal, −Coronal, −Anterior

5.2. Analysis of Parkinsonian speech

Fig. 1b shows the analysis of the HC and PD non-silence speech data: 10 ms framed 805 511 phonological posterior vectors of the HC group, and 10 ms framed 784 128 vectors of the PD group.

Statistical analysis using the two-sample *t*-test, without assuming equal variances, of the differences between HC and PD speech, resulted into the only invariant [Consonantal] feature with $p = 0.1029$, which is in contradiction with non-modal analysis above, where the [Consonantal] feature was significantly different between modal and non-modal phonations. PD speech exhibited higher values of the [Nasal], [Voice] and [High] features, and lower values of the [Back], [Low], and [Round] features. Validation of these findings is discussed further in Section 5.3.

Having the statistics of mean DPP features, we calculated Euclidean distances using Equation 3 between parkinsonian DPP $\boldsymbol{\Delta\mu^P}$ (visualized at right of Figure 1b), and L non-modal DPP $\boldsymbol{\Delta\mu_l^{NM}}$. Table 3 lists obtained Euclidean distances. As said in Section 1, non-modal phonation modes are contrastive against modal phonation modes, in other words, they are di-similar. The q_l quantities represent the similarity measures, so to be used for characterisation of Parkinsonian non-modal phonation, they are turned into di-similarity measures by calculating their inverse, $1/q_l$. Finally, we assume that each of the non-modal phonation partially (relatively) contributes to the perceived overall non-modal phonation.

Table 3: *Euclidean distances q_l between non-modal and Parkinsonian DPP features. As Euclidean distance is a similarity measure, whereby smaller is more similar, we calculate an inverse of the Euclidean distance to plot composition of non-modal voice quality in Parkinsonian speech in Figure 2.*

Voice quality	q_l	$1/q_l$
Breathy	0.0935	10.69
Creaky	0.1240	8.06
Tense	0.1417	7.06
Falsetto	0.1904	5.25
Harsh	0.2321	4.31

Figure 2 shows composition of voice quality in parkinsonian speech. It might be interpreted as: a voice of an average patient with Parkinson's disease contains "a voice quality spectrum" composed of 30% breathy voice, 23% creaky voice, 20% tense voice, 15% falsetto voice and 12% harsh voice, where about 75% of overall voice quality on average is composed of breathy, creaky and tense phonations.

5.3. Validity

Oates (2009) describes basic pathological phonations as a breathy voice that arises from incomplete glottal closure and/or the presence of a posterior glottal chink, a rough voice that arises from irregular vocal fold vibration patterns, and a strained or pressed voice that is due to excess laryngeal muscle tension. Barsties and De Bodt (2015) review three ratings schemes that are the most frequently reported and accepted: (i) the GRBAS scale that includes R for roughness, B for breathiness and S for strain; (ii) the CAPE-V that includes in the standard analysis the same parameters as the GRBAS; and (iii) the RBH scale that focuse on only three dimensions: roughness, breathiness, and hoarseness.

We objectively estimated that the majority of the VQ spectrum of PD is composed of 30% breathy voice, 23% creaky voice, 20% tense voice; all the three most-important VQs expected/evaluated by perceptual assessment of hypokinetic dysarthria in PD. Breathy phonation causes breathiness, creaky phonation contributes significantly to roughness, and tense phonation results into vocal strain (known also as muscle tension dysphonia).

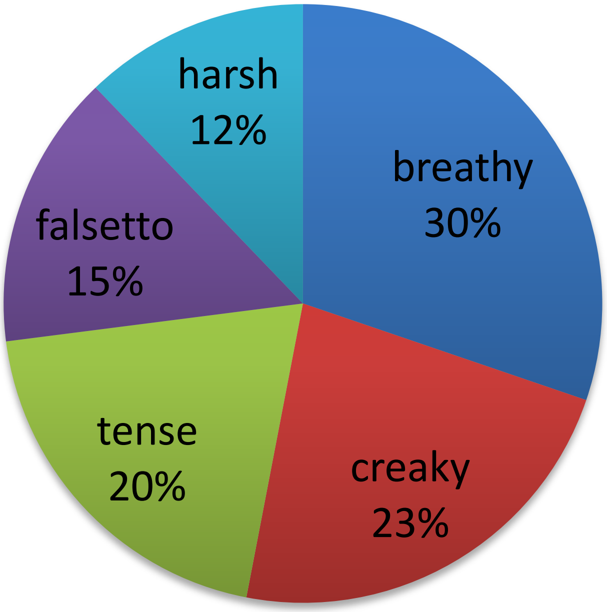


Figure 2: Composition of voice quality in Parkinson's speech.

Severity of dysarthria in PD is also rated by the the Frenchay Dysarthria Assessment (FDA-2) score (Enderby, 1983; Enderby and Palmer, 2008). The assessment includes 28 relevant perceptual dimensions of speech, namely related to the following dimensions:

- Respiration: noting running out of breath when speaking, and breathy voice.
- Laryngeal: noting weather the patient has clear phonation with the vocal folds, without huskiness.
- Tongue: noting accurate tongue movements (positions) with correct articulation.
- Palate: noting nasal resonance in spontaneous conversation, without hypernasality or nasal emission.
- Lips: observing the movements of lips in conversation, noting correct shape of lips.

While the first dimension is similar to the perceptual assessment of the three rating schemes described above, further dimensions are more related to articulation. According to Figure 1b, PD speech data exhibits:

1. Greater values of the [Voice] and [Nasal] phonological features that

might be related to the Laryngeal and Palate dimensions. It can be interpreted as more analysed speech frames having higher values of these phonological features, as compared to HC speech data. Thus, patients produced more nasal or voiced sounds compared to HCs.

2. Lower values of the [Round] that might be related to the Lips dimension (i.e., patients produced less rounded sounds).

3. Lower [Back] and [Low], and greater [High] vales that might be related to Tongue dimension (i.e., the patients articulated more central speech sounds, that might indicate weaker articulation of PD patients).

5.3.1. Prediction of laryngeal FDA scores

To validate usefulness of the proposed characterisation of the VQ of Parkinson's disease, we investigated using the q_l features for the prediction of the dysarthria level according to a modified version of the Frenchay assessment score. This perceptual evaluations includes the following aspects of speech: respiration, lips movement, palate/velum movement, larynx, tongue, and intelligibility. We hypothesized that the DPP features should be useful for prediction of the FDA scores related particularly to the larynx, which impacts the VQ the most.

The baseline features include articulation and prosody-based features, which are concatenated to form a 724-dimensional feature vector per utterance (Orozco-Arroyave, 2016; Vasquez-Correa et al., 2017). The articulation-based features includes 86 descriptors such as the energy content distributed in 22 Bark bands in the transition from voiced to unvoiced segments (22 descriptors), and from unvoiced to voiced segments (22 descriptors) Orozco-Arroyave et al. (2016). The feature set is augmented with the first and second formant frequencies, and 12 MFCC with their derivatives. The extracted features are grouped and four functionals are computed (mean, standard deviation, skewness, and kurtosis), forming a 344-dimensional feature vector per utterance. The second feature set contains prosody-based features computed with the Erlangen prosody module (Zeißler et al., 2006), using voiced segments as speech unit. The set of features comprises a total of 95 features. 19 of them are based on duration and include among others the number and the length of voiced frames, and duration of pauses. 36 of the features are based on the F_0 contour, including the mean, standard deviation, jitter, and others. The energy–based features include measures of the energy within the voiced frames, shimmer, position of the maximum energy, and others. The features are grouped into one feature vector and four functionals are also computed: mean, standard deviation, maximum, and minimum, forming a 380-dimensional feature vector per utterance.

The evaluated features consisted of the concatenated baseline and q_l features calculated per speaker. All 50 PD speakers were considered in this evaluation. For the prediction task, we used the same Super Vector Regression as described by Vasquez-Correa et al. (2017), using a leave-one-subject-out (LOSO) cross-validation. The performance is evaluated using the Spearman's correlation coefficient between the predicted scores and the real scores. The real scores were obtained by three professional phoniatricians, with the inter-rater reliability of 0.86 measured as the average Spearman's correlation coefficient obtained between all the evaluators.

Table 4: *The Spearman's correlation coefficients between the real and predicted modified FDA scores related to the larynx. Median values are calculated for the correlations with the three evaluators. Results obtained for the three sub-sets of the PD data (see Section 4.2) are reported.*

Speech task	Baseline features	Proposed features
Pataka	0.56	0.56
Read text	0.39	**0.47**
Monologue	0.55	**0.57**

Table 4 shows the correlation achieved with the baseline and the q_l features. Improvements are obtained for the monologue and reading speech tasks, of 3% and 16% respectively, whereas no improvement is obtained with the pataka speech task. The results imply that the proposed q_l features depend on statistics (μ_k as the mean values of phonological probabilities), and better results are obtained with more observed (recorded) data. For example, while the pataka tasks contain speech samples with repeated single word, the read text task includes speech samples of 36 spoken words.

6. Conclusions

The paper has proposed the characterisation of voice quality (VQ) applied to pathological speech in PD. Often, the analysis of pathological speech is limited by available data, and advanced deep machine learning techniques cannot be fully applied. The lack of proper perceptual labels of pathological speech adds further complication. Therefore, the proposed characterisation learns statistics from healthy speech data that is more widely available, and calculates similarity with disordered speech by using the Euclidean distance.

The results obtained by DPP features have been validated by matching the obtained most significant, non-modal phonation types with evaluating parameters of the perceptual assessments. In addition, DPP features of PD have been interpreted by the Frenchay assessment. This interpretation ability can be directly used in clinical assessment.

A drawback of the presented experimental study was in missing VQ perceptual labels of PD data. To the authors' knowledge, the used PD database is the biggest open-source database available, containing both isolated and connected speech, and was selected primarily for its size. By missing perceptual labels, validation of the proposed VQ characterisation thus has been done on all speakers focusing on differentiating HC and PD speech, and its direct application in diagnosis and therapy is limited. In future, we plan to obtain PD data with labeled VQ, and validate the VQ characterisation on individual patients, looking for example for regression of the perceptual scores.

7. Acknowledgements

The work reported here was partially carried out during the 2016 Jelinek Memorial Summer Workshop on Speech and Language Technologies, which was supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, and Facebook. It was also partially supported by the Swiss National Science Foundation project "MoSpeeDi. Motor Speech Disorders: characterizing phonetic speech planning and motor speech programming/execution and their impairments", and by CODI from Universidad de Antioquia, project 2015-7683, and COLCIENCIAS project # 111556933858.

8. References

Anderson, R. C., Klofstad, C. A., Mayew, W. J., Venkatachalam, M., May 2014. Vocal fry may undermine the success of young women in the labor market. PloS one 9 (5), e97506+.

Aronson, A. E., Bless, D., 2011. Clinical voice disorders. New York: Thieme.

Awan, S. N., Roy, N., Dromey, C., 2009. Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model. Clinical Linguistics & Phonetics 23 (11), 825–841.

Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., Gerratt, B. R., Sep. 1997. Analysis by synthesis of pathological voices using the Klatt synthesizer. Speech Communication 22 (4), 343–368.

Barsties, B., De Bodt, M., 2015. Assessment of voice quality: current state-of-the-art. Auris Nasus Larynx 42 (3), 183–188.

Bauer, V., Alerić, Z., Jančić, E., Miholović, V., 2011. Voice quality in Parkinsons disease in the Croatian language speakers. Collegium antropologicum 35 (2), 209–212.

Bhuta, T., Patrick, L., Garnett, J. D., 2004. Perceptual evaluation of voice quality and its correlation with acoustic measurements. Journal of Voice 18 (3), 299–304.

Cernak, M., Asaei, A., Honnet, P.-E., Garner, P. N., Bourlard, H., 2016. Sound Pattern Matching for Automatic Prosodic Event Detection. In: Proc. of Interspeech. pp. 170–174.

Cernak, M., Benus, S., Lazaridis, A., 2017a. Speech vocoding for laboratory phonology. Computer, Speech and Language 42, 100–121.

Cernak, M., Garner, P. N., 2016. PhonVoc: A Phonetic and Phonological Vocoding Toolkit. In: Proc. of Interspeech. San Francisco, CA, USA, pp. 988–992.

Cernak, M., Nöth, E., Rudzicz, F., Christensen, H., Orozco-Arroyave, J. R., Arora, R., Bocklet, T., Chinaei, H., Hannink, J., Nidadavolu, P. S., Vasquez, J. C., Yancheva, M., Vann, A., Vogler, N., 2017b. On the Impact of Non-modal Phonation On Phonological Features. In: Proc. of ICASSP. IEEE.

Cernak, M., Potard, B., Garner, P. N., Apr. 2015. Phonological vocoding using artificial neural networks. In: Proc. of ICASSP. IEEE, pp. 4844–4848.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York, NY.

Darley, F. L., Aronson, A. E., Brown, J. R., 1969. Differential diagnostic patterns of dysarthria. Journal of Speech, Language, and Hearing Research 12 (2), 246–269.

Drugman, T., Kane, J., Gobl, C., Sep. 2014. Data-driven detection and analysis of the patterns of creaky voice. Computer Speech & Language 28 (5), 1233–1253.

Enderby, P. M., 1983. Frenchay dysarthria assessment. College Hill Press.

Enderby, P. M., Palmer, R., 2008. FDA-2: Frenchay Dysarthria Assessment: Examiner's Manual. Pearson.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. NASA STI/Recon technical report 93.

Hansen, J. H., Hasan, T., 2015. Speaker recognition by machines and humans: A tutorial review. IEEE Signal Processing Magazine 32 (6), 74–99.

Hanson, H. M., Jan. 1997. Glottal characteristics of female speakers: acoustic correlates. The Journal of the Acoustical Society of America 101 (1), 466–481.

Hinton, G. E., Osindero, S., Teh, Y. W., Jul. 2006. A Fast Learning Algorithm for Deep Belief Nets. Neural Comput. 18 (7), 1527–1554.

Hirano, M., 1981. Clinical examination of voice. Vol. 5 of Disorders of human communication. Springer.

J Holmes, R., M Oates, J., J Phyland, D., J Hughes, A., 2000. Voice characteristics in the progression of Parkinson's disease. International Journal of Language & Communication Disorders 35 (3), 407–418.

Kane, J., Sep. 2012. Tools for analysing the voice. Ph.D. thesis, Trinity College Dublin, Dublin.

Klatt, D. H., Klatt, L. C., Feb. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America 87 (2), 820–857.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., Berke, G. S., 1993. Perceptual evaluation of voice qualityreview, tutorial, and a framework for future research. Journal of Speech, Language, and Hearing Research 36 (1), 21–40.

Ladefoged, P., Johnson, K., Jan. 2014. A Course in Phonetics, 7th Edition. Cengage Learning.

Laver, J., Mar. 1980. The Phonetic Description of Voice Quality. Cambridge Studies in Linguistics. Cambridge University Press.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., Ramig, L. O., et al., 2009. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering 56 (4), 1015–1022.

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., Moroz, I. M., 2007. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMedical Engineering OnLine 6 (1), 23.

Malyska, N., 2008. Analysis of nonmodal glottal event patterns with application to automatic speaker recognition. Ph.D. thesis, Harvard University – MIT Division of Health Sciences and Technology, USA.

Malyska, N., Quatieri, T. F., Dunn, R. B., 2011. Sinewave Representations of Nonmodality. In: Proc. of Interspeech. pp. 69–72.

Maryn, Y., De Bodt, M., Roy, N., 2010. The acoustic voice quality index: toward improved treatment outcomes assessment in voice disorders. Journal of Communication Disorders 43 (3), 161–174.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., Corthals, P., 2009. Acoustic measurement of overall voice quality: A meta-analysis. The Journal of the Acoustical Society of America 126 (5), 2619–2634.

Oates, J., 2009. Auditory-perceptual evaluation of disordered voice quality. Folia Phoniatrica et Logopaedica 61 (1), 49–56.

Orozco-Arroyave, J. R., 2016. Analysis of Speech of People with Parkinson's Disease. Vol. 41. Logos Verlag Berlin GmbH.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Bonilla, J. F. V., Gonzalez-Rátiva, M. C., Nöth, E., 2014. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In: LREC. pp. 342–347.

Orozco-Arroyave, J. R., Vásquez-Correa, J. C., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., Nöth, E., 2016. Towards an automatic monitoring of the neurological state of the Parkinson's patients from speech. In: 41st International Conference on Acoustic, Speech, and Signal Processing (ICASSP).

Paul, D. B., Baker, J. M., 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the workshop on Speech and Natural Language. HLT '91. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 357–362.

Podesva, R. J., 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. Journal of Sociolinguistics 11 (4), 478–504.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., Dec. 2011. The Kaldi Speech Recognition Toolkit. In: Proc. of ASRU. IEEE SPS, iEEE Catalog No.: CFP11SRW-USB.

Rasipuram, R., Magimai.-Doss, M., May 2011. Integrating articulatory features using Kullback-Leibler divergence based acoustic model for phoneme recognition. In: Proc. of ICASSP. IEEE, pp. 5192–5195.

Rusz, J., Cmejla, R., Ruzickova, H., Ruzicka, E., 2011. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinsons disease. The journal of the Acoustical Society of America 129 (1), 350–367.

San Segundo, E., Tsanas, A., Gómez-Vilda, P., 2017. Euclidean distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. Forensic Science International 270, 25–38.

Schuller, B., Batliner, A., Nov. 2013. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley.

Shinoda, K., Watanabe, T., 1997. Acoustic modeling based on the MDL principle for speech recognition. In: Proc. of Eurospeech. pp. I –99–102.

Stouten, F., Martens, J.-P., May 2006. On The Use of Phonological Features for Pronunciation Scoring. In: Proc. of ICASSP. Vol. 1. IEEE, p. I.

Titze, I. R., 1995. Workshop on acoustic voice analysis: Summary statement. National Center for Voice and Speech.

Vasquez-Correa, J. C., Orozco-Arroyave, J. R., Arora, R., Nöth, E., Dehak, N., Christensen, H., Rudzicz, F., Bocklet, T., Cernak, M., Chinaei, H., Hannink, J., Nidadavolu, P. S., Yancheva, M., Vann, A., Vogler, N., 2017. Multi-view Representation Learning Via GCCA for Multimodal Analysis of Parkinson's Disease. In: Proc. of ICASSP.

Wuyts, F. L., De Bodt, M. S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K., Raes, J., Van de Heyning, P. H., 2000. The dysphonia severity indexan objective measure of vocal quality based on a multiparameter approach. Journal of Speech, Language, and Hearing Research 43 (3), 796–809.

Yu, D., Siniscalchi, S., Deng, L., Lee, C.-H., March 2012. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In: Proc. of ICASSP. IEEE SPS.

Zeißler, V., Adelhardt, J., Batliner, A., Frank, C., Nöth, E., Shi, R. P., Niemann, H., 2006. The prosody module. In: SmartKom: foundations of multimodal dialogue systems. Springer, pp. 139–152.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K., 2007. The HMM-based Speech Synthesis System Version 2.0. In: Proc. of ISCA SSW6. pp. 131–136.