IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 February 2023, accepted 16 March 2023, date of publication 22 March 2023, date of current version 28 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260187

RESEARCH ARTICLE

Classification of Phonation Modes in Classical Singing Using Modulation Power Spectral Features

MANUEL BRANDNER[1], PAUL ARMIN BEREUTER[1], SUDARSANA REDDY KADIRI[2], (Member, IEEE), AND ALOIS SONTACCHI[1]

[1]Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, 8010 Graz, Austria
[2]Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland

Corresponding author: Manuel Brandner (brandner@iem.at)

This work was supported in part by the Academy of Finland under Project 330139.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Advisory Board of the University of Music and Performing Arts, Graz, and performed in line with the Declaration of Helsinki.

ABSTRACT In singing, the perceptual term ''voice quality'' is used to describe expressed emotions and singing styles. In voice physiology research, specific voice qualities are discussed using the term phonation modes and are directly related to the voicing produced by the vocal folds. The control and awareness of phonation modes is vital for professional singers to maintain a healthy voice. Most studies on phonation modes have investigated speech and have used glottal inverse filtering to compute features from an estimated excitation signal. The performance of this method is reported to decrease at high pitches, which limits its usability for the singing voice. To overcome this, this study proposes to use features derived from the modulation power spectrum for phonation mode classification in the singing voice. The exploration of the modulation power spectrum is motivated by the fact that, in singing, temporal modulations (known as vocal vibrato) and spectral modulations hold information of the vocal fold tension. Since there exists no large publicly available dataset of phonation modes in singing, we created a new dataset consisting of six female and four male classical singers, who sang five vowels at different pitches in three phonation modes (breathy, modal, and pressed). Experimental results with a support vector machine classifier reveal that the proposed features show better classification performance compared to state-of-the-art reference features. The performance for the current dataset is at least 10% higher compared to the performance of the reference features (such as glottal source features and MFCCs) in the case of target labels and around 6% higher in the case of perceptually assessed labels.

INDEX TERMS Modulation power spectrum, phonation modes, singing voice analysis, voice qualities.

I. INTRODUCTION

The classification of phonation modes as a computerised aid in classical singing voice training seems vital. Maintaining a healthy voice is an important component of professional singing and is essential for students during the course of their studies. The analysis and classification of phonation modes in classical singing might give a singer valuable insights into their voice production, and in the best case, prevent voice production problems. As voice production problems often occur throughout the course of voice studies, self-monitoring during vocal training and vocal warm-up could be beneficial to prevent more serious problems, which usually entail a longer rest period.

The phonation modes studied in classical singing are not pathological. The transition from one to the other is more tenuous than the pathological phonation modes investigated in clinical studies. Different phonation modes mean different

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

vibration patterns of the vocal folds, and in classical singing four main phonation modes can be distinguished. They are: *breathy*, *flow* (also referred to as resonant in [1]), *modal* and *pressed* phonation. *Breathy* phonation is characterized in [2] by minimal adductive tension, causing the vocal folds to reside in a Y-shaped state which leaves an opening at the top of the vocal folds, even during the closure phase of one glottal cycle. This constant opening lets the turbulent tracheal airflow enter the vocal tract at any time causing a *breathy* voice perception, which is also referred to as aspiration noise [3]. Physiologically, *modal* phonation in singing voice is defined by a full-length vibration of the vocal folds caused by moderate tension and compression, which implicitly leads to a full glottal closure of the vocal folds during vibration [2], [4]. The opposite end of the phonatory dimension described in [2] is given with *pressed* voice, which is defined by strong subglottic pressure and adduction caused by tense muscles surrounding the vocal folds.

The differences between the phonation modes have been studied for speech, using features derived by signal processing methods that attempt to separate vocal fold movement information (excitation signal or glottal source waveform) from the vocal tract contribution (filter). A common approach of calculating glottal waveform characteristics is by using a glottal inverse filtering method (GIF) (based on *source-filter deconvolution*) [5], [6], [7], [8], [9]. An overview of glottal source processing is given in [10]. Although speech and singing share similar basic concepts of voice production, the analysis of singing voices is far more difficult from a signal processing point of view, especially due to higher and rapid variations in pitch. As the pitch increases, the analyzed signal in the frequency domain exhibits an increased sparsity, especially for sustained vowels due to its harmonic structure, which leads to an ill-posed mathematical condition for GIF methods. Automated inverse filtering usually fails at higher pitches and some implementations are reported to be already erroneous at above ca. 300 Hz [11]. Features used for phonation mode classification, which are derived from an estimated glottal waveform using a GIF-method are commonly denoted as voice quality features (VQ-features) [12], [13].

Another common parametric method for separating the vocal tract contribution and calculating glottal source characteristics is cepstral analysis [14], which was initially introduced in [15] for seismic analysis in order to find echo components. The difference between cepstral analysis and glottal inverse filtering lies in the fact that the information of both the source and the filter are found in the same resulting cepstral domain signal, but at different locations along the so-called quefrency axis. The magnitude of the first peak along the quefrency axis has been found to be well suited to determine the breathiness of the voice [16] and, when used as a feature, is called cepstral peak prominence (CPP). In [17], it is concluded that CPP is similar to the first rahmonic and gives meaningful results to detect breathiness. In [18] CPP was reported to separate neutral, breathy, and pressed phonation from each other, but not flow from the other phonation modes. Also, breathy phonation was shown to have high level of turbulent noise and is reported to have a large harmonic to noise ratio (HNR) [19]. Modal phonation in singing results in rich harmonics and pressed phonation is reported to typically show a weaker fundamental and more dominating higher harmonics [20]. Mel-frequency cepstral coefficients (MFCCs) [21] serve as a descriptive representation of the magnitude spectrum and are frequently used for classification of phonation modes in [13] and [18].

Unlike the cepstral analysis, the modulation spectrum comprises temporal information and results from an analysis along the temporal axis and not along the frequency axis. In [22], the significance of low-frequency modulations is discussed along with how to use the modulation spectrum to analyze sound in accordance with the human auditory system. Studies in [23] and [24] showed that temporal modulations influence speech intelligibility. However, the extraction of characteristics from the modulation spectrum is not trivial, due to its high dimensionality. This is why in [25], [26], and [27] it is proposed to apply a Higher Order Singular Value Decomposition (HOSVD) on the modulation spectrum in combination with a feature selection algorithm based on the mutual information. The results for their approach for voice pathology detection achieves a detection rate of around 94% and to classify hoarseness, a global classification rate of 74% is reported. However, they did not use the modulation power spectrum as originally presented in [28] and [29], which combines the advantages of cepstral analysis and the modulation spectrum by extracting both temporal and spectral modulations. Moreover, they limited their investigations to a single vowel (/a/). The most comparative studies investigating phonation modes in classical singing have been [13], [18], and [1], but unfortunately all studies used a small dataset with data of only two singers. According to the authors' knowledge, there are no studies on classification of phonation modes in classical singing using characteristics extracted from the modulation power spectrum and no previous work has investigated phonation modes on a larger dataset consisting of data from more than two classical singers.

II. GOALS OF THE CURRENT STUDY

In classical singing, vocal vibrato is a temporal modulation which lies at 4 to 8 Hz [30], [31] and spectral modulations depend on the spectral composition of a sung vowel, which provide information on the harmonic structure of a sound. The spectral composition can show how breathy or strained the voice sounds, which depends on the singer's physical effort on the vocal folds and the amount of used airflow [32]. In order to investigate both the temporal and spectral modulations of sung vowels, we propose the investigation of novel features extracted from the modulation power spectrum (MPS) [28], [29]. This method combines the benefits of the discussed parametric and non-parametric approaches. We investigate a peak-picking technique similar to CPP, where we additionally include higher harmonics along the temporal and spectral modulation axes, as opposed

to an algebraic approach like the HOSVD utilized in [25], [26], and [27]. As of now, the data made public in [1] and [20] are the only two openly accessible datasets of professional singers, singing with different phonation modes. However, both of them have severe restrictions regarding the number of singers and ratings. Thus, we have created a new dataset including ten singers singing five vowels in three phonation modes (breathy, modal, and pressed) over a large pitch range. We propose two novel feature sets derived from the modulation power spectrum and one feature set derived from an averaged cepstrum over consecutive time frames for the classification of the phonation modes *breathy*, *modal* and *pressed*. The proposed feature sets are compared to three state-of-the-art reference feature sets. The feature sets are compared by means of their classification performance using a support vector machine (SVM) classifier (see section V).

The highlights and novelties of the current study are:

- Investigation of temporal and spectral characteristics extracted from the modulation power spectrum.
- Investigation of automatic classification of phonation modes (breathy, modal, and pressed) on a newly created classical singer dataset.
- Investigation of a feature reduction by using the averaged MPS along the temporal axis.
- Study of the performance of features derived from the averaged cepstrum over consecutive time frames compared to MPS features.
- Comparison of the proposed features with state-of-the-art reference features, which are: voice quality features (VQ-features), cepstral features (MFCCs), and features derived after zero frequency filtering (ZFF features).

The organization of the paper is as follows: Section III describes the data collection including measurements and labelling. The extraction of features and calculation of the modulation power spectrum are described in section IV. The experimental protocol is described in section V, which gives a general overview of the classification framework, the reference features, the classifier, and the evaluation metrics. Results of the classification experiments are presented in section VI. In section VII the results are discussed and section VIII summarizes the study.

III. DATA COLLECTION

For the present work, we created a new dataset of audio recordings of sustained vowels sung at various pitches with three different phonation modes. Furthermore, we conducted a listening assessment in order to obtain the perceived phonation modes of the recorded vowels. Although datasets on phonation modes in singing exist, the current dataset is much larger compared to existing datasets [1], [20]. The already available datasets hold recordings of only two classical singers, whereas the proposed dataset contains recordings of ten classical singers. In contrast to the already available data, where the labels of phonation modes are based solely on the judgements of a single expert, a perceptual assessment was performed for the presented dataset, resulting in 6 ratings per recording. This allows for a statistical analysis of the perceived phonation mode labels. The dataset is named as Voice Qualities in Singing (VQS) and is publicly available at: https://phaidra.kug.ac.at/o:126552.

A. MEASUREMENTS

The measurements were recorded with a microphone (omnidirectional pattern, NTI M2230, Schaan, Liechtenstein) at a distance of 1 m in front of the singer. For the acoustic analysis, dry signals measured in an anechoic environment are ideal. However, in singing, room acoustics support the voice, which is a necessity in a longer recording session. Therefore, we used an augmented acoustic system with zero latency [33], [34], which only gives the singer natural room acoustics via transparent headphones [35] while creating no reverberation on the microphone signals. The augmented acoustic system is fed with the signal of the microphone placed in front of the singer and employs a static, however frequency-dependent directivity to excite the virtual room. The virtual room simulates a shoe-box-like concert hall with a size of roughly $30\,\text{m} \times 24\,\text{m} \times 20\,\text{m}$ and reverberation time of 2.2 s. Typical reverberation times of concert halls are in the range between 1.5 s and 3 s [36], [37].

1) ROOM CONDITIONS

Measurements were carried out in a sound treated measurement room with absorptive material on the walls and floor at the Institute of Electronic Music and Acoustics in Graz. The frequency-dependent room reverberation in the measurement room is less than 75 ms between 400 Hz and 1 kHz, and less than 50 ms above 1 kHz. The volume of the room is approximately 50 m^3 with a floor space of 22.50 m^2.

B. AUDIO RECORDINGS

Four male singers (3 tenors and 1 baritone) were instructed to sing 5 sustained German vowels (/a:/, /e:/, /i:/, /o:/, and /u:/) over the pitch range from H/B2/123 Hz to a^1/A4/440 Hz on a whole-tone scale, except for the baritone, who only sang up to e^1/E4/330 Hz. Furthermore, six female singers (3 sopranos and 3 mezzo-sopranos) sang the vowels from a/A3/220 Hz to a^2/A5/880 Hz. The singers were asked to sing the vowels, starting on the consonant /m/ and sustaining the vowel for 2 seconds. The vowels were repeated three times each with different provoked voice phonation modes (modal, breathy, and pressed). This results in a total number of 2145 audio samples. The dataset was then reduced to a total number of 1140 samples to ensure a listening assessment of reasonable duration (see section III-C). This reduced dataset is consequently used in the classification experiments presented in section V.

All singers were trained classical singers except for the baritone (studied jazz vocals), who said to have the ability to mimic the classical singing technique due to his teaching experience at the music conservatory. The average age was 29.6 years (the youngest was 24 years and the oldest

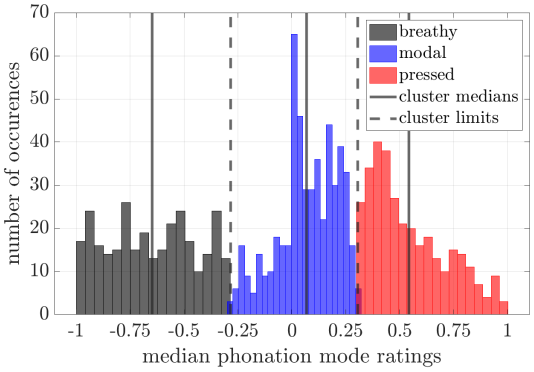


FIGURE 1. Histogram of perceptually assessed median phonation mode ratings along with the cluster limits (which are used to derive the perceptually assessed labels).

was 34 years). The classically trained singers were 4 graduate students (at the end of their current master studies), 5 post-graduate students (with one master's degree or more), and 1 undergraduate student (bachelor's degree). Six of the singers were also teaching at the time. The singers were asked to sing at a comfortable loudness level (mezzo-forte). All the participants were well-trained for the task due to their extensive practice during their classical vocal studies.

a: NOTE ON THE PHONATION MODES

We chose to study three phonation modes: breathy, modal, and pressed, as most singers were unfamiliar with the term ''flow'' phonation. In this sense, the term ''modal'' in our study and in the dataset defines the optimal singing voice phonation, and the other two phonation modes ''breathy'' and ''pressed'' are deviations from this optimal state.

C. TARGET AND PERCEPTUALLY ASSESSED LABELS

The instructions given to the singers during the course of the recordings are used as target labels in this work. Due to the time-consuming nature of a listening assessment, only a portion of the recordings were chosen to be perceptually assessed. This resulted in the selection of 1140 samples. The pitches of these samples are listed in Table 1. The samples were randomly grouped into ten subsets, which each were independently rated by 6 listeners, resulting in six independent ratings per sample. A total of 20 listeners participated in the assessment. As a starting point for the current investigation, a k-mediods clustering algorithm is used to categorize the median of the six independent ratings for each sample. The distributions of the three phonation mode clusters derived using the k-mediods algorithm are visualized in Fig. 1 as histograms. The cluster boundaries are chosen as the upper and lower boundary of the modal cluster, which provides one possible straight-forward approach for the assignment of a fixed label to each recording. Fig. 1 also shows a smaller distance between the cluster medians of the modal and pressed cluster compared to the distance between the modal and breathy cluster. From the resulting 1140 samples, 297 samples were rated as breathy, 516 as modal, and 327 as pressed. The amount of data for target labels and perceptually assessed labels are listed in Table 2 along with information of gender, vowels, and pitch range. The confusion matrix in Table 3 presents the differences between the perceptually assessed and the target labels. When comparing the different labels of phonation mode classification, it can be seen that pressed and modal phonation consistently cause the greatest uncertainty, suggesting that performance differences in classifications are to be expected. However, it is also reasonable to anticipate that there will be some uncertainty in the data, if only target labels (instructions to the singers) are looked into.

IV. FEATURE EXTRACTION

In this section, newly developed features, based on the modulation power spectrum (MPS) including their underlying theory, are discussed along with the features derived from the averaged cepstrum over consecutive time frames. In order to calculate the MPS-based features, a peak-picking procedure is applied, which uses the knowledge of the fundamental frequency. Three sets of features are proposed, one set is based on the two-dimensional MPS-representation, leading to a larger dimension feature set (MPS_{peaks}), the second set builds on a compact, summed version of the MPS resulting in a smaller dimension feature set (MPS_{sum}), and the third set is derived based on cepstral peaks ($Ceps_{peaks}$), as there exists a strong relation between the cepstrum and the spectral modulation dimension of the MPS. A schematic block diagram describing the steps involved in the computation of the three feature sets is shown in Fig. 2.

A. MODULATION POWER SPECTRUM

The origin of the modulation power spectrum (MPS) can be traced back to the field of neuroscience, where it was used to better understand human auditory processing [28]. Temporal modulations which constitute the modulation spectrum along the time axis, have been studied for different tasks such as audio coding, modification, and automatic classification [22]. Subsequently, temporal modulations have also been extensively investigated for pathological voices [25], [26], [27]. The MPS combines the information of temporal modulations and the approach of cepstral analysis, which aims at a separation of vocal tract and voice source information [14]. The MPS is calculated by applying a two-dimensional Fourier transform on the squared and logarithmized amplitude values of a short-time Fourier transform (STFT) ($X(m,k)$), computed with the block-length L and hop-size R. The STFT consists of N positive frequencies and M time-frames, where the current time-block is denoted using the index m and k denotes the discrete frequency indices. Note that instead of using the natural logarithm as mentioned in [28], we use the logarithm with base 10, and represent the spectro-temporal modulation amplitudes $S(k_f, k_t)$ in decibels.

$$S(k_f, k_t) = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} 10 \log_{10}(\|X(m,k)\|^2) e^{-j2\pi U}, \tag{1}$$

TABLE 1. Pitches of the data selected from the full dataset which were perceptually assessed. Frequency differences in Hz between the pitches (delta) are also listed to show the almost linear spacing.

pitch	$c/C3$	$g/G3$	$c^1/C4$	$e^1/E4$	$g^1/G4$	$a^1/A4$	$c^2/C5$	$d^2/D5$	$e^2/E5$	$g^2/G5$	$a^2/A5$
Hz	131	196	262	330	392	440	523	587	659	788	880
delta	-	65	66	68	62	48	83	64	72	129	92

TABLE 2. Overview of the analyzed dataset* with information on gender, vowels, pitch range, and the number of target and perceptually assessed phonation mode labels.

gender	female singers (6)	male singers (4)
vowels	/a:, e:, i:, o:, u:/	/a:, e:, i:, o:, u:/
pitch range	$c^1/C4/261Hz$ to $a^2/A5/880Hz$	$c/C3/131Hz$ to $a^1/A4/440Hz$ **
target	breathy (270), modal (270), pressed (270)	breathy (110), modal (110), pressed (110)
assessed	breathy (221), modal (359), pressed (230)	breathy (76), modal (157), pressed (97)

* Subset of the full dataset at selected pitches.
** The baritone only sung up to e^1/E4/330 Hz.

TABLE 3. Percentage of confusions: ***target labels*** (rows) vs. ***perceptually assessed labels*** (columns). The perceptual listening assessment's results are compared with the reference number of target labels per class (380).

	Breathy [%]	**Modal** [%]	**Pressed** [%]
Breathy	74	20	6
Modal	4	71	25
Pressed	1	45	54

where $U = \left(\frac{mk_t}{M} + \frac{k_f k}{N}\right)$. The indices $k_f = -\left\lfloor \frac{N}{2} \right\rceil, \ldots, \left\lfloor \frac{N}{2} \right\rceil$ and $k_t = -\left\lfloor \frac{M}{2} \right\rceil, \ldots, -\left\lfloor \frac{M}{2} \right\rceil$[1] are the corresponding discrete spectral and temporal modulation frequency bins in the joint modulation frequency domain after the two-dimensional Fourier transform. In our implementations, a Blackman-Harris window with a block-length of 80 ms, a hop-size of 2.5 ms, and a 4096-point fast Fourier transform at a sampling frequency of 16 kHz are used. The block length is longer than commonly used in speech signal processing (25 to 50 ms) to accommodate the nature of the singing voice. The most important aspect is the choice of a block length and a hop size that allow to study the vibrato characteristics of the classical singing voice, which has a vibrato frequency around 4 to 8 Hz [30].

B. EXTRACTION OF MODULATION POWER SPECTRAL FEATURES

The MPS represents a high dimensional feature space and exhibits pitch-dependent regions of high and low spectro-temporal energy, especially in the case of sustained vowels. Therefore, we propose a pitch-normalized peak-picking strategy to extract only the high energy components of the MPS. Fig. 3 shows illustrations of modulation power spectra for three phonation modes (breathy, modal, and pressed). The search regions depicted in Fig. 3 on the spectral modulation axis (y-axis) with a width of $\pm\frac{1}{3}\tau_0$ are centered around $n_f \cdot \tau_0$, with $n_f = \{1, 2, \ldots, N_f = 8\}$ being positive multiples of the fundamental period $\tau_0 = \frac{1}{f_0}$. The fundamental periods are determined using the reference pitches listed in Table 1. The search regions on the temporal modulation axis are fixed around $N_t = 5$ multiples, centered at $n_t = \{-2, -1, 0, 1, 2\}$, times a pre-selected vibrato frequency of $f_{vib} = 6$ Hz.[2] The search region width along the x-axis is chosen to be $\pm\frac{1}{3}f_{vib}$. The boundaries of the spectral and temporal search regions are formulated in (2) and (3).

$$\tau_i = n_{f,i} \cdot \tau_0 \pm \frac{1}{3}\tau_0 \tag{2}$$

$$f_{t_{mod},i} = n_{t,i} \cdot f_{vib} \pm \frac{1}{3}f_{vib} \tag{3}$$

The modulation frequencies $f_{t_{mod}}$ and τ denote the temporal and spectral frequencies used in the modulation power spectrum. They can be calculated using the linear relationship between the discrete modulation frequency bins k_t and k_f and the spectral modulation frequency resolution Δ_τ and temporal modulation frequency resolution $\Delta_{f_{t_{mod}}}$ (see (4) and (5)).

$$\tau = \Delta_\tau \cdot k_f \tag{4}$$

$$f_{t_{mod}} = \Delta_{f_{t_{mod}}} \cdot k_t \tag{5}$$

These modulation frequencies are used below to describe the process of calculating the newly proposed features from the modulation power spectrum.

$\mathbf{MPS_{peaks}}$: The full set of peak amplitudes derived as in (6) is denoted as MPS_{peaks}.

$$MPS_{peaks} = \max\{S(\tau_i, f_{t_{mod},i})\} \tag{6}$$

The dimension of this feature set is: $N_t \cdot N_f = 5 \cdot 8 = 40$ peak amplitudes.

$\mathbf{MPS_{sum}}$: In order to further reduce the dimensionality of the full set of peak values MPS_{peaks}, the MPS is summed along the temporal modulation axis, see (7).

$$S_\Sigma(k_f) = \sum_{k_t} S(k_t, k_f) \tag{7}$$

[1] $\lfloor \cdot \rceil$ denotes a rounding operation.

[2] f_{vib} is chosen as the mean value of the vibrato range 4 to 8 Hz reported in [30].

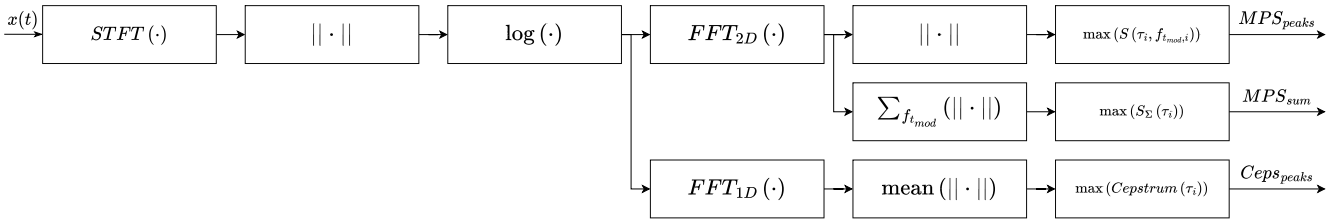
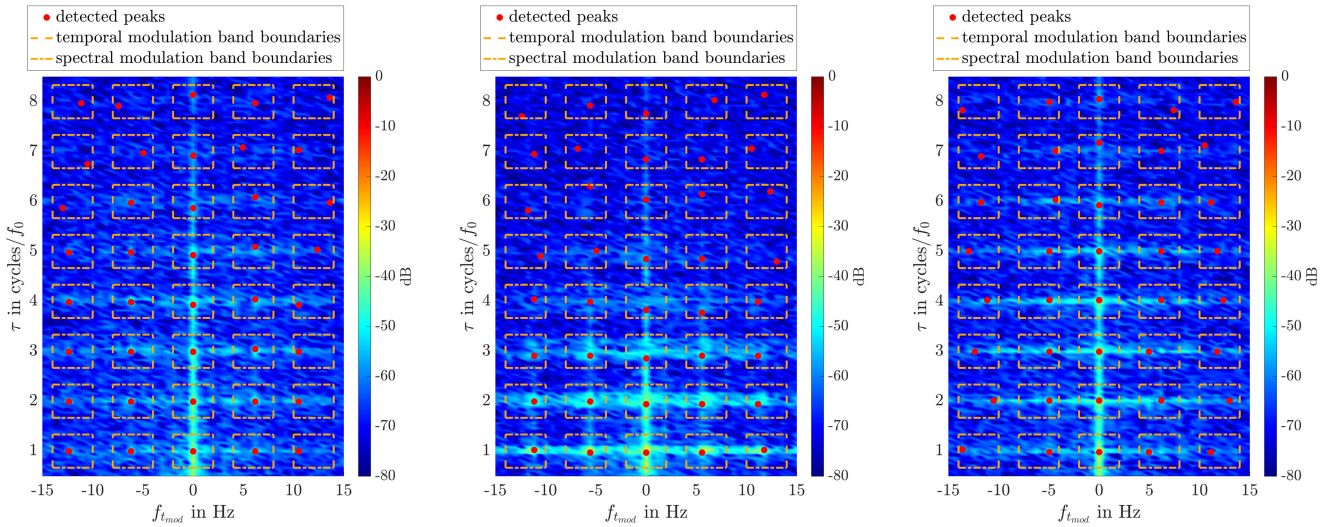


FIGURE 2. Schematic block diagram for the extraction of features from the modulation power spectrum (MPS_{peaks} and MPS_{sum}) and the averaged cepstrum ($Ceps_{peaks}$).



(a) breathy. (b) modal. (c) pressed.

FIGURE 3. Illustrations of modulation power spectra for the three phonation modes. Shown are the extracted peak amplitudes within a search grid referenced to the fundamental frequency ($\tau_0 = \frac{1}{f_0}$) and the average temporal modulation frequency for vibrato $f_{vib} = 6$ Hz along both the temporal and spectral axes. The search grid regions along the spectral axis lie within $n_f \cdot \tau_0 \pm \frac{1}{3}\tau_0$ with $n_f \in \{1, 2, \ldots, 8\}$, and along the temporal axis within $n_t \cdot f_{vib} \pm \frac{1}{3} f_{vib}$ with $n_t \in \{-2, -1, 0, 1, 2\}$.

Again, 8 peaks along the spectral modulation axis are picked (see (8)), which is the dimension of the MPS_{sum} feature set.

$$MPS_{sum} = \max\{S_\Sigma(\tau_i)\} \tag{8}$$

Fig. 3 shows the modulation power spectra for the breathy, modal, and pressed phonation modes. The MPS for breathy phonation shown in Fig. 3(a), is characterized by low energy along the temporal modulation axis, and evenly distributed energy along the spectral modulation axis around $f_{t_{mod}} = 0$ Hz. Fig. 3(b) shows the illustration of the MPS for modal phonation, exhibiting strong temporal modulation components, but also a substantial energy decrease along the spectral modulation axis. The MPS for pressed phonation shown in Fig. 3(c) shows more energy at higher spectral modulation frequencies. Fig. 4 shows the illustrations of the summed modulation power spectra for the three phonation modes. From the illustrations of the modulation power spectra (shown in Fig. 3 and Fig. 4), it is clearly evident that temporal modulation components and spectral harmonic structures vary among the three phonation modes.

C. EXTRACTION OF FEATURES FROM THE AVERAGED CEPSTRUM

There exists a strong relationship between the spectral modulation dimension of the MPS and the cepstrum. Therefore, peak values extracted from the averaged cepstrum over consecutive time frames are also considered as possible features. The calculation of the cepstrum is presented in Fig. 2, whereas the features extracted from it are denoted as $Ceps_{peaks}$. We extracted the peak amplitudes within the averaged cepstrum using the same search regions as for the spectral modulation axis in the MPS (see Sec. IV-B).

V. EXPERIMENTAL PROTOCOL

In order to investigate the phonation mode classification performance of the newly proposed modulation power spectral features we set up a classification problem. The newly developed features are compared with state-of-the-art reference features that are commonly used and have been employed in previous comparative work [1], [13], [18]. In the current study, we use a support vector machine (SVM) including a hyperparameter optimization, and a leave-one-singer-out

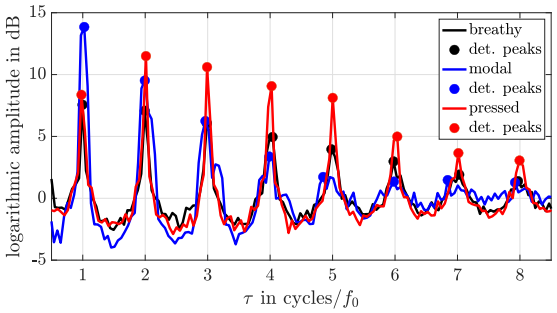


FIGURE 4. Illustrations of the summed modulation power spectra for the three phonation modes. The circular markers indicate the peak extraction for the corresponding MPS_{sum} feature. The summed modulation power spectra are detrended by subtracting a fitted first-order polynomial for better visualization.

(LOSO) cross-validation technique. This may reduce overall performance accuracy, but should increase the generalizability of the current results. The basic processing and classification framework is shown as a block diagram in Fig. 5. The feature extraction block is preceded by the pre-processing steps, which includes segmenting the audio samples to the sustained part of the vowels and excluding the consonant /m/ at the beginning. The feature extraction block summarizes the computation of the reference features, as well as the steps for calculating the newly developed features based on the MPS and the averaged cepstrum (averaged over consecutive time frames, see section IV). The last block indicates the classification process with the SVM, which predicts one of the three phonation modes for each audio sample after the classifier has been trained according to the LOSO cross-validation and hyperparameter optimization technique. The following sections present the reference features, the details of the classifier and the evaluation metrics along with the classification framework.

A. REFERENCE FEATURES

The proposed features extracted from the modulation power spectrum are compared to three state-of-the-art reference feature sets (see Fig. 5), which are briefly described in the following paragraphs.

1) VOICE QUALITY FEATURES (VQ)

The VQ feature set consists of six features, derived from a glottal waveform estimate, which is calculated using a GIF-method [12], [13]. The six features are: (1) normalized amplitude quotient (NAQ) [38], (2) quasi-open quotient (QOQ) [10], [39], (3) amplitude difference between fundamental and first harmonic (H1-H2) [39], (4) parabolic spectral parameter (PSP) [40], (5) harmonic richness factor (HRF) [39] and (6) maximum dispersion quotient (MDQ) [12]. The literature shows that voice quality features work well for speech, but their applicability to singing is known to be limited at high pitches, due to erroneous glottal inverse filtering [11]. Nevertheless, we use these features as reference features in the current study.

2) ZERO FREQUENCY FILTERING (ZFF)

The ZFF method provides an approximate voice source waveform without explicitly using the source-filter model of speech production. The ZFF feature set consists of four features, which are: the strength of excitation (SoE), the energy of excitation (EoE), the loudness measure and the ZFF signal energy. These features were shown to be useful for discriminating phonation types in speech and singing [13], [41], [42]. SoE was shown to be proportional to the rate of glottal closure, the EoE feature was shown to capture the vocal effort, and the loudness measure was shown to capture the abruptness of the glottal closure [43], [44]. The energy of the ZFF signal at glottal closure is also used as a feature which was shown to capture low frequency energy [13]. Zero frequency filtering features have been designed to overcome the problem of the classical voice quality features and have been extensively investigated in [13], but it has been shown that the performance for singing voices could still not be improved significantly.

3) MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs)

MFCCs are popular features used in many tasks, such as automatic speech recognition [45], [46], music information retrieval [47], [48], including phonation modes classification in speech and singing [18]. The MFCCs are derived using the same parameters as for the MPS (Blackman-Harris window, 80 ms window length and 2.5 ms hop-size). From the mel-cepstrum, the first 36 cepstral coefficients are derived. The 0^{th} coefficient is not considered, which results in a 35-dimensional feature vector. MFCCs have been shown to be versatile descriptors for speech recognition tasks and phonation mode classification in several previous works [13], [18]. In comparison to other features, MFCCs are harder to interpret and their descriptive quality usually depends largely on the number of used coefficients.

B. FEATURE SET COMBINATIONS

Additionally, the reference and proposed features described above were combined in order to study the complementary information among the features. In total, 9 combinations of feature sets (FSCs) were created in addition to the single feature sets and their corresponding performances are discussed in subsection VI-C.

C. CLASSIFIER

We use a SVM with a radial basis function kernel as a classifier [49], because it is reported to perform well even on a smaller number of training data. We perform a hyperparameter tuning with GridSearchCV [49] within a LOSO cross-validation strategy to avoid over-fitting and increase the generalizability of the model. For our phonation modes classification task, we use two types of labels: (i) the instructed phonation modes given to the singers during recordings (target labels) and (ii) the perceptually asssessed phonation mode labels from the listening assessment (perceptually

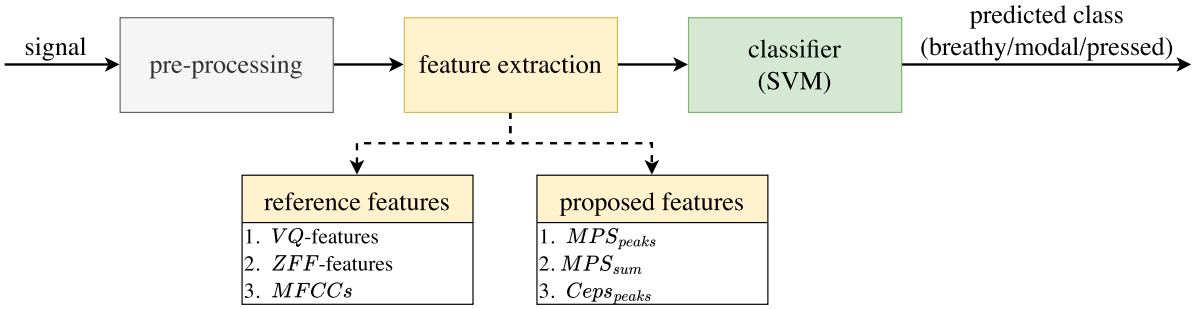


FIGURE 5. Block diagram of the basic processing and classification framework including the pre-processing stage, the feature extraction for the reference and proposed features, and the classification using a support vector machine to predict one of the three phonation modes for an audio sample. The depicted framework indicates the processing after the SVM classifier has been trained according to the LOSO cross-validation and hyperparameter optimization technique.

assessed labels). Experiments are conducted by considering both genders (including both male and female singers, totally 10 singers) and only female singers (6 singers).

D. EVALUATION METRICS

Performance measures are the mean and standard deviation of the test accuracy over the runs of the LOSO cross-validation (for the whole dataset and the female-only dataset). We omit the male-only dataset due to its small sample size. We have computed the standard deviation of the accuracy for each feature set to see the reliability of the features across varying singers. As an additional metric, we provide confusion matrices for the test sets averaged over all runs of the LOSO cross-validation to examine the confusions among phonation modes.

VI. RESULTS

In this section, the results of the classification problem described in section V are presented for the target and perceptually assessed labels, in terms of accuracy and confusion matrices (see sections VI-A and VI-B). In each of the subsections, we present the results (for both the reference and proposed features) for the whole dataset, i.e., combination of male and female singers (see Table 2 for an overview of the data and labels) and the female-only dataset. Finally, the results of the feature set combinations are listed in section VI-C.

A. CLASSIFICATION RESULTS FOR TARGET LABELS

This section presents the classification results (in terms of mean and standard deviation of accuracies) obtained for the target labels. The classification accuracies for the whole data are given in Table 4. From the table, it is observed that the proposed feature sets (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) perform better than the reference features. All the proposed features show a mean accuracy which is around 10% higher than the mean accuracy of the reference features. However, the standard deviations of the accuracies are all larger for the proposed features. The most striking aspect in Table 4 is the similar performance of the 40-dimensional MPS_{peaks} feature set and the 8-dimensional MPS_{sum} and $Ceps_{peaks}$ feature sets. In order to gain more information on the misclassifications, confusion matrices are given in Table 5 and Table 6 for one proposed feature set (MPS_{sum}) and one reference feature set (VQ-features). The confusion matrices clearly show that there exists greater confusion between pressed and modal, and between breathy and modal phonation modes in the VQ feature set, compared to the proposed MPS_{sum} feature set.

The results for the female-only data are given in Table 7. It is expected that the female-only data, still consisting of 810 samples, will be more homogeneous because all singers

TABLE 4. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the classical singers (female+male) using the *target labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference features:		
MFCCs (35)	57	± 9
VQ (6)	46	± 6
ZFF (4)	51	± 6
Proposed features:		
MPS_{peaks} (40)	**68**	± **14**
MPS_{sum} (8)	67	± 12
$Ceps_{peaks}$ (8)	66	± 11

TABLE 5. Confusion matrix in % for the MPS_{sum} feature set of the whole dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	77	12	11
Modal	8	76	16
Pressed	12	32	56

TABLE 6. Confusion matrix in % for the VQ feature set of the whole dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	60	29	11
Modal	27	47	26
Pressed	23	45	32

TABLE 7. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the female classical singers using the *target labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference Features:		
MFCCs (35)	61	$\pm$ 11
VQ (6)	48	$\pm$ 4
ZFF (4)	56	$\pm$ 3
Proposed Features:		
MPS_{peaks} (40)	74	$\pm$ 9
MPS_{sum} (8)	**75**	$\pm$ **7**
$Ceps_{peaks}$ (8)	68	$\pm$ 10

TABLE 8. Confusion matrix in % for the MPS_{sum} feature set of the female-only dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	88	5	7
Modal	6	74	20
Pressed	9	27	64

TABLE 9. Confusion matrix in % for the ZFF feature set of the female-only dataset using the *target labels*.

	Breathy [%]	Modal [%]	Pressed [%]
Breathy	64	24	12
Modal	30	44	26
Pressed	12	28	60

sang the same pitches. Similar to the results of the whole dataset, it can also be observed that the proposed feature sets (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) perform better than the reference features. Moreover, it can be seen that the results show lower standard deviations compared to the results for the whole dataset, except for the MFCCs. The classification accuracies are also increased for all the feature sets by 4-8% for the female-only data, where the MPS_{sum} feature set showed the highest accuracy among all features, and the lowest standard deviation among the proposed feature sets. The confusion matrix shown in Table 8 demonstrates the increased performance for the MPS_{sum} feature set, concerning less confusion between breathy and pressed phonation modes compared to the results for the whole dataset, but no performance increase for the classification of modal phonation. On the other hand, the confusion matrix for the ZFF feature set given in Table 9 indicates that there exists greater confusion with modal for breathy and pressed phonation modes compared to the MPS_{sum} feature set in the female-only data.

B. CLASSIFICATION RESULTS FOR PERCEPTUALLY ASSESSED LABELS

This section gives the classification results obtained for the perceptually assessed labels. The classification accuracies for the whole data are given in Table 10. The table, shows that the proposed feature sets (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) again perform better than the reference features. In general, the performance differences between the feature sets decrease when perceptually assessed labels are used.

The results for the female-only data evaluated for the perceptually assessed labels are given in Table 11. Again,

TABLE 10. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the classical singers (female+male) using the *perceptually assessed labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference Features:		
MFCCs (35)	59	$\mp$ 7
VQ (6)	52	$\mp$ 9
ZFF (4)	54	$\mp$ 9
Proposed Features:		
MPS_{peaks} (40)	64	$\mp$ 6
MPS_{sum} (8)	**65**	$\mp$ **8**
$Ceps_{peaks}$ (8)	61	$\mp$ 6

TABLE 11. Phonation mode classification accuracies (mean and standard deviation (Std.)) for the female classical singers using the *perceptually assessed labels*.

Features (x)	Mean accuracy [%]	Std. [%]
Reference Features:		
MFCCs (35)	65	$\mp$ 3
VQ (6)	53	$\mp$ 5
ZFF (4)	55	$\mp$ 4
Proposed Features:		
MPS_{peaks} (40)	64	$\mp$ 6
MPS_{sum} (8)	**69**	$\mp$ **5**
$Ceps_{peaks}$ (8)	64	$\mp$ 5

the highest mean accuracies are obtained for the MPS_{sum}, followed by the MPS_{peaks}, $Ceps_{peaks}$ and MFCC feature set.

Overall, the results in the classification experiments (both for target and perceptually assessed labels) show that the features extracted from the modulation power spectrum (MPS_{peaks}, and MPS_{sum}), and the cepstral features ($Ceps_{peaks}$) perform better than the reference features. Most striking is the performance of the MPS_{sum} feature set, which only consists of 8 features. This suggests that including modulation characteristics in phonation mode analysis is beneficial, especially in the analysis of classical singing.

C. CLASSIFICATION RESULTS FOR COMBINATIONS OF FEATURE SETS

This section reports the results for the combination of the proposed and the reference features on the whole dataset for the target and perceptually assessed labels. The feature sets are combined to investigate the complementary information among the feature sets. In total, 9 feature set combinations (FSC) for each label group (target and perceptually assessed labels) were created as listed in Table 12. FSC1 and FSC2 include combinations of the reference feature sets. FSC3 to FSC5 combine the best reference feature set combination with the proposed feature sets. FSC6 to FSC8 combines the proposed feature sets to investigate their corresponding complementary information. FSC9 combines the best performing feature sets of the reference feature set combinations and the proposed feature set combinations. The feature set combination FSC8 produces the highest mean accuracies for the target labels and FSC9 for the assessed labels. Interestingly, the combination of the newly proposed features FSC8 (MPS_{peaks} and $Ceps_{peaks}$) and FSC7 (MPS_{sum} and $Ceps_{peaks}$) perform

TABLE 12. Feature set combinations FSC1 to FSC9 and their corresponding accuracies (mean and standard deviation in percent) for the target and perceptually assessed labels.

	target (f.+m.)			percep.ass.(f.+m.)		
ID	**combined features**	**Mean acc.** [%]	**Std.** [%]	**combined features**	**Mean acc.** [%]	**Std.** [%]
FSC1	VQ+ZFF	51	7	VQ+ZFF	58	6
FSC2	ZFF+MFCCs	59	11	VQ+ZFF+MFCCs	65	8
FSC3	ZFF+MFCCs+MPS_{peaks}	69	14	VQ+ZFF+MFCCs+MPS_{peaks}	67	7
FSC4	ZFF+MFCCs+MPS_{sum}	67	14	VQ+ZFF+MFCCs+MPS_{sum}	68	7
FSC5	ZFF+MFCCs+$Ceps_{peaks}$	66	14	VQ+ZFF+MFCCs+$Ceps_{peaks}$	67	11
FSC6	MPS_{peaks}+MPS_{sum}	68	15	MPS_{peaks}+MPS_{sum}	67	7
FSC7	MPS_{sum}+$Ceps_{peaks}$	68	13	MPS_{sum}+$Ceps_{peaks}$	68	7
FSC8	MPS_{peaks}+$Ceps_{peaks}$	**70**	**14**	MPS_{peaks}+$Ceps_{peaks}$	65	5
FSC9	ZFF+MFFCs+MPS_{peaks}+$Ceps_{peaks}$	70	15	VQ+ZFF+MFCCs+MPS_{sum}+$Ceps_{peaks}$	**70**	**7**

TABLE 13. Confusion matrices for all the reference features (MFCC, VQ, and ZFF) and proposed features (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) of the whole dataset (combination of female and male data) and the female-only dataset using the target labels and perceptually assessed labels. Here B, M and P refer to breathy, modal, and pressed phonation modes, respectively.

		target (f.+m.)			target (f.)			percep.ass. (f.+m.)			percep.ass. (f.)		
		B	**M**	**P**	**B**	**M**	**P**	**B**	**M**	**P**	**B**	**M**	**P**
		Reference Features:											
MFCCs	B	**71**	18	11	**74**	17	9	**72**	23	5	**79**	19	2
	M	22	**56**	22	18	**63**	19	16	**64**	20	12	**73**	15
	P	20	33	**47**	12	42	**46**	11	46	**43**	7	54	**38**
VQ	B	**60**	29	11	**54**	25	21	**42**	37	21	**49**	27	24
	M	27	**47**	26	25	**42**	33	13	**60**	27	17	**52**	31
	P	23	45	**32**	17	35	**48**	9	41	**50**	11	31	**58**
ZFF	B	**63**	26	11	**64**	24	12	**41**	48	11	**47**	41	12
	M	28	**52**	20	30	**44**	26	19	**65**	16	21	**60**	19
	P	24	34	**42**	12	28	**60**	13	37	**50**	4	41	**55**
		Proposed Features:											
MPS_{peaks}	B	**79**	9	12	**90**	5	5	**80**	15	5	**85**	10	5
	M	6	**77**	17	4	**78**	18	9	**75**	16	8	**78**	14
	P	12	34	**54**	7	40	**53**	11	55	**34**	7	54	**39**
MPS_{sum}	B	**77**	12	11	**88**	5	7	**72**	23	5	**82**	13	5
	M	8	**76**	16	6	**74**	20	8	**76**	16	10	**74**	16
	P	12	32	**56**	9	27	**64**	7	46	**47**	6	47	**47**
$Ceps_{peaks}$	B	**76**	8	16	**82**	8	10	**79**	17	4	**87**	9	4
	M	8	**73**	19	10	**72**	18	8	**80**	12	8	**71**	21
	P	17	31	**52**	11	40	**49**	9	74	**17**	6	65	**29**

better or nearly similarly to FSC9, for both target and perceptually assessed labels, even though FSC9 holds some or all reference features. Note that the feature set combinations for each corresponding set of labels (target and perceptually assessed) include the best performing feature sets of the reference features and the proposed features. Thus, the feature set combinations vary for the different set of labels.

VII. DISCUSSION

From the results in Tables 4, 7, 10, and 11, it is clearly evident that the performance for the MPS_{sum} (8-dimensional) features is similar or better than the MPS_{peaks} (40-dimensional) features.

This is explainable by the summation included in calculating the MPS_{sum} features. The summed modulation power spectrum is calculated by summing all the energy along the temporal axis for each spectral modulation frequency τ, whereas the MPS_{peaks} features only hold information on the peak amplitudes within the search regions. The extracted modulation information contained in the MPS_{peaks} and MPS_{sum} features seems to increase the capability of distinguishing modal from the other phonation modes. The peaks extracted from the cepstrum, averaged over consecutive time frames, contain similar information as the peaks extracted from the summed MPS, but without the temporal modulation information (i.e., vocal vibrato), which decreases the performance compared to the MPS_{sum} and MPS_{peaks} features.

The confusion matrices for all the features (reference and proposed) with the target and perceptually assessed labels in the whole dataset and in the female-only dataset scenario are given in Table 13. These confusion matrices show that the proposed MPS feature sets and the MFCCs exhibit the best performance in classifying breathy phonation, while the MPS_{peaks} feature set performs best over the whole dataset for both the target and the perceptually assessed labels.

The pronounced difference between breathy and other phonation modes can be visualized in the averaged modulation power spectra (averaged over all singers for each corresponding phonation mode), depicted in Fig. 6.[3] Breathy phonation shows the lowest values for the temporal

[3]We subtracted a fitted first order polynomial to detrend the MPS data for better visualization.

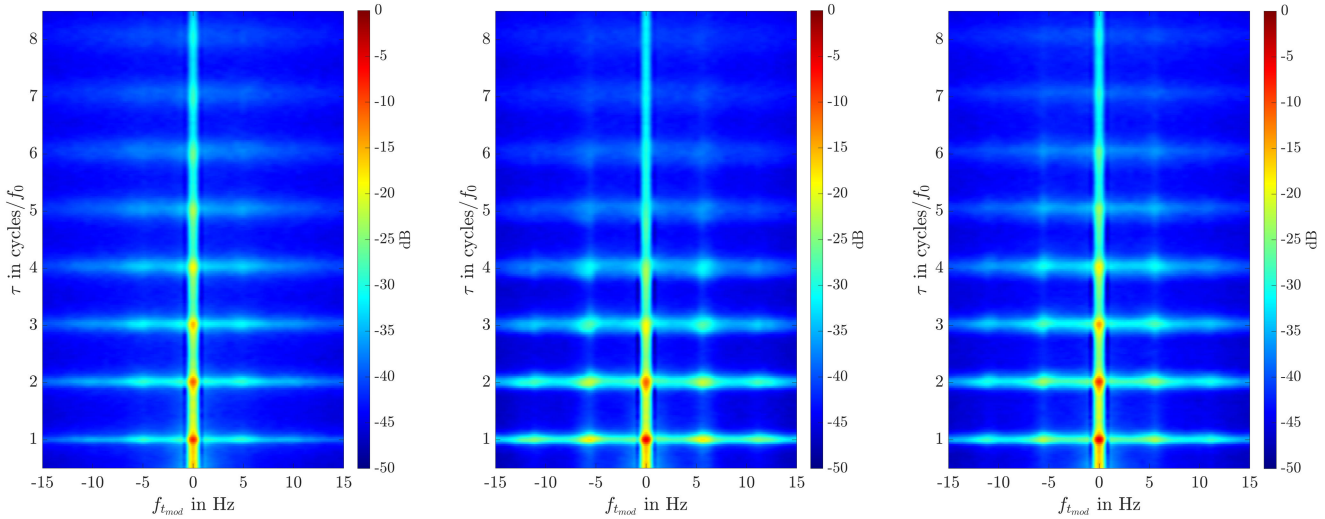


FIGURE 6. Averaged modulation power spectra of all singers (whole dataset) for breathy, modal, and pressed phonation. The modulation power spectrum for each sample is detrended by subtracting a fitted a first-order polynomial before averaging in order to highlight the differences between the phonation modes.

modulation components. Regarding the classification of pressed phonation, the confusion matrices of Table 13 overwhelmingly show a reduced performance for all investigated feature sets. The less distinct difference between modal and pressed phonation is also visible in the averaged MPS illustrated in Fig. 6. Nonetheless, slight differences between modal and pressed phonation modes are still detectable, especially at higher spectral modulation frequencies starting at $\tau = 3$ cycles/f_0, but they are less pronounced than in the visualizations presented in Fig. 3.

Additionally, Fig. 7[3] shows the summed MPS averaged over all singers for each corresponding phonation mode. Again, the discussed reduced differences between modal and pressed are visible. A decreased amplitude at $\tau = 1$ cycles/f_0, as seen in Fig. 4 for pressed phonation, diminishes in the averaged data.

The classification experiments with the perceptually assessed labels show a slightly lower performance compared to target labels for almost all feature sets. However, an approach other than using the median rating in combination with k-mediods clustering could lead to better performance, offering potential for future research. Overall, the results of the classification experiments show a higher mean accuracy for the novel MPS features when using the target labels. This suggests that the target labels, in combination with the new features, provide a better discrimination of phonation modes for the present dataset.

The variance of the classification accuracy is strongly influenced by the underlying data and the labels. The lowest standard deviations for all feature sets are present when the data is reduced to the female-only data and by using the labels from the perceptual listening assessment (see Table 11). This reduced variance is generally noticeable for the perceptually assessed labels, most likely due to the involved strategy, which entails categorizing these labels with the median value of six ratings. The variance seen for the target labels may

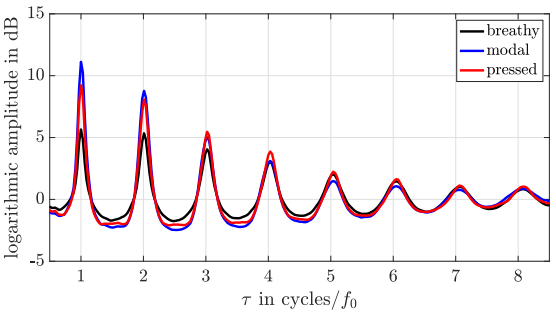


FIGURE 7. Averaged summed modulation power spectra of all singers for each phonation mode (breathy, normal, and pressed). The summed modulation power spectrum is detrended for each sample by subtracting a fitted first order polynomial for better visualisation.

be explained by the leave-one-singer-out strategy. A greater variance is to be anticipated if one or more singers perform the instructed phonation modes (target labels) with more consistency than the others. This is also evident in the experiments of the combined feature sets (for both target and perceptually assessed labels). Furthermore, the results demonstrate that there exists a weaker complementary information between the various feature sets.

In addition, a comparison between target labels and labels from the perceptual listening assessment (see Table 3) shows similar confusions (i.e., between modal and pressed phonations). This implies that either the singers were unable to fully reproduce the target phonation modes in the recordings or it was too challenging for listeners to distinguish the phonation modes. Most likely, both of these factors are at play. This limits, to some extent, the discussion on the performance of the current feature sets for classifying pressed phonation, and the comparison of performance between target or perceptually assessed labels. However, to the authors' knowledge, the data investigated in this work is currently the largest publicly available annotated dataset for phonation modes in singing and further investigations on the dataset should follow.

VIII. CONCLUSION

In this article, we have proposed three new feature sets based on the modulation power spectrum and the averaged cepstrum for the classification of phonation modes (breathy, modal, pressed) in classical singing. We have also presented a newly collected phonation modes dataset, which consists of ten classical singers singing several vowels at several pitches, which is publicly available at: https://phaidra.kug.ac.at/o:126552. Experiments were carried out on the whole (combination of female and male data) and the female-only data using target labels (instructed phonation modes during the recordings) as well as perceptually assessed labels (derived from a perceptual listening assessment). We have compared the proposed three features (MPS_{peaks}, MPS_{sum}, and $Ceps_{peaks}$) with state-of-the-art reference features (VQ, ZFF, and MFCCs) in a phonation mode classification task. The results of the classification experiments reveal that the proposed features based on the MPS have a slightly better ability to accurately assess the phonation modes. In terms of performance and most striking in number of features, MPS_{sum} is the best performing feature set compared to other feature sets, which by itself produces a similar classification accuracy as the best performing feature set combination. Additionally, it has been found that the target labels result in a better performance than using the labels derived from the perceptually assessed ratings. It was found that the influence of the underlying data and the corresponding labelling play an important role in the classification. This should be taken into account in future approaches. Further research is still needed on deriving phonation mode specific features without using the fundamental frequency, as well as on thoroughly examining the listening assessment data.

IX. ETHICS AND CONSENT

This work involved human subjects or animals in its research. Ethical and experimental procedures and protocols were designed to comply with the proposal reviewed by the ethics advisory board of the University of Music and Performing Arts, Graz, and performed in line with the Helsinki declaration. All participants were informed that their participation was voluntary and could be withdrawn any time. Participants received an expense allowance for their voluntary participation.

REFERENCES

[1] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed—Automatic detection of phonation mode from audio recordings of singing," *J. New Music Res.*, vol. 42, no. 2, pp. 171–186, Jun. 2013, doi: 10.1080/09298215.2013.821496.

[2] J. Sundberg, *The Science of the Singing Voice*. DeKalb, IL, USA: Northern Illinois Univ. Press, 1987.

[3] H.-L. Lu and J. O. Smith, "Glottal source modeling for singing voice synthesis," in *Proc. ICMC*, 2000, pp. 1–8. [Online]. Available: http://hdl.handle.net/2027/spo.bbp2372.2000.186

[4] C. Gobl, "A preliminary study of acoustic voice quality correlates," *Quart. Prog. Status Rep.*, vol. 4, pp. 9–22, Jan. 1989.

[5] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Amer.*, vol. 31, no. 6, pp. 667–677, Jun. 1959, doi: 10.1121/1.1907771.

[6] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1625–1629, 1974, doi: 10.1121/1.1903487.

[7] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, pp. 109–118, Jan. 1992, doi: 10.1016/0167-6393(92)90005-R.

[8] P. Alku and E. Vilkman, "Preliminary experiences in using automatic inverse filtering of acoustical signals for the voice source analysis," *Scandin. J. Logopedics Phoniatrics*, vol. 17, no. 2, pp. 128–135, Jan. 1992, doi: 10.3109/14015439209098723.

[9] P. Alku, "Glottal inverse filtering analysis of human voice production—A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, Oct. 2011, doi: 10.1007/S12046-011-0041-5.

[10] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1117–1138, Sep. 2014, doi: 10.1016/j.csl.2014.03.003.

[11] I. Arroabarren and A. Carlosena, "Inverse filtering in singing voice: A critical analysis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1422–1431, Jul. 2006, doi: 10.1109/TSA.2005.858013.

[12] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 6, pp. 1170–1179, Jun. 2013, doi: 10.1109/TASL.2013.2245653.

[13] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Commun.*, vol. 118, pp. 33–47, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639319303358

[14] L. Rabiner and R. Schafer, *The Cepstrum and Homomorphic Speech Processing*. London, U.K.: Pearson, 2011.

[15] B. Bogert, M. Healy, and J. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and Saphe cracking," in *Proc. Symp. Time Ser. Anal.*, vol. 15, 1963, pp. 209–243.

[16] J. M. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech Hearing Res.*, vol. 37, no. 4, pp. 769–778, 1994, doi: 10.1044/JSHR.3704.769.

[17] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomed. Signal Process. Control*, vol. 14, pp. 42–54, Nov. 2014, doi: 10.1016/j.bspc.2014.07.001.

[18] D. Stoller and S. Dixon, "Analysis and classification of phonation modes in singing," in *Proc. 17th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, New York, NY, USA, 2016, pp. 80–86.

[19] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[20] J.-L. Rouas and L. Ioannidis, "Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings," in *Proc. Interspeech*, Sep. 2016, pp. 150–154, doi: 10.21437/Interspeech.2016-1135.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[22] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 7, pp. 1–8, Dec. 2003. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1155/S1110865703305013

[23] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[24] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 1647–1650.

[25] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 2514–2517, doi: 10.1109/IEMBS.2009.5334850.

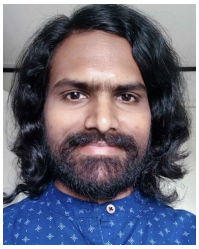
[26] M. Markaki and Y. Stylianou, "Modulation spectral features for objective voice quality assessment," in *Proc. 4th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, Mar. 2010, pp. 1–4, doi: 10.1109/ISCCSP.2010.5463313.

[27] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 1938–1948, Sep. 2011, doi: 10.1109/TASL.2010.2104141.

[28] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Amer.*, vol. 114, no. 6, pp. 3394–3411, 2003, doi: 10.1121/1.1624067.
[29] A. Hsu, S. M. N. Woolley, T. Fremouw, and F. E. Theunissen, "Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons," *J. Neurosci.*, vol. 24, pp. 9201–9211, Jan. 2004, doi: 10.1523/JNEUROSCI.2449-04.2004.
[30] R. Husson and C. E. Seashoee, "Psychology of the vibrato in voice and instrument," *Revue Musicologie*, vol. 19, no. 66, p. 115, May 1938, doi: 10.2307/925282.
[31] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neurosci. Biobehav. Rev.*, vol. 81, pp. 181–187, Oct. 2017, doi: 10.1016/j.neubiorev.2017.02.011.
[32] Z. Zhang, "Mechanics of human voice production and control," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. 2614–2635, Oct. 2016.
[33] M. Frank, D. Rudrich, and M. Brandner, "Augmented practice-room—Augmented acoustics in music education," in *Proc. 46th Conf. Fortschritte Akustik, Deutsche Gesellschaft Akustik*, vol. 46, Mar. 2020, pp. 151–154. [Online]. Available: https://www.dega-akustik.de/publikationen/online-proceedings/
[34] N. Klanjscek, L. David, and M. Frank, "Evaluation of an E-learning tool for augmented acoustics in music education," *Music Sci.*, vol. 4, Aug. 2021, Art. no. 20592043211037511, doi: 10.1177/20592043211037511.
[35] N. Meyer-Kahlen, D. Rudrich, M. Brandner, S. Wirler, S. Windtner, and M. Frank, "DIY modifications for acoustically transparent headphones," *J. Audio Eng. Soc.*, vol. 148, pp. 1–5, May 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20841
[36] L. L. Beranek, "Concert Hall acoustics—1992," *J. Acoust. Soc. Amer.*, vol. 92, no. 1, pp. 1–39, 1992.
[37] M. Skålevik, "Reverberation time—The mother of all room acoustic parameters," in *Proc. 20th Int. Congr. Acoustic, Integr. Comput.-Aided Eng.*, vol. 10, 2010, pp. 2508–2513.
[38] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, Feb. 2002, doi: 10.1121/1.1490365.
[39] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. Interspeech*, Aug. 2007, pp. 1410–1413.
[40] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter—A new method for quantification of the glottal flow," *Speech Commun.*, vol. 22, no. 1, pp. 67–79, 1997, doi: 10.1016/S0167-6393(97)00020-4.
[41] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ZTWCCs)," in *Proc. Interspeech*, Sep. 2018, pp. 232–236, doi: 10.21437/Interspeech.2018-2498.
[42] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *Proc. Interspeech*, Sep. 2018, pp. 441–445, doi: 10.21437/Interspeech.2018-2502.
[43] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. Interspeech*, Aug. 2013, pp. 1916–1920.
[44] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. Interspeech*, Sep. 2015, pp. 1324–1328.
[45] S. J. Young, D. K. J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[46] H. K. Kathania, S. Shahnawazuddin, W. Ahmad, and N. Adiga, "Role of linear, Mel and inverse-Mel filterbanks in automatic recognition of speech from high-pitched speakers," *Circuits, Syst., Signal Process.*, vol. 38, no. 10, pp. 4667–4682, Oct. 2019, doi: 10.1007/s00034-019-01072-7.
[47] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. ISMIR*, 2000. Accessed: Mar. 23, 2023. [Online]. Available: https://ismir2000.ismir.net/indexnoframes.html and https://ismir2000.ismir.net/papers/logan_paper.pdf
[48] D. Grzywczak and G. Gwardys, "Audio features in music information retrieval," in *Active Media Technology*, D. Ślęzak, G. Schaefer, S. T. Vuong, and Y.-S. Kim, Eds. Cham, Switzerland: Springer, 2014, pp. 187–199.
[49] A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

MANUEL BRANDNER received the M.S. degree in electrical engineering and audio engineering from the Graz University of Technology, in 2014. He is currently pursuing the Ph.D. degree with the Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria. He was a Scientific Project Assistant with the University of Music and Performing Arts Graz, in 2014, where he worked in the research fields of active noise control, acoustic measurement design, sound zoning, and machine learning. He was an university assistant, from 2018 to 2020, and has been a lecturer, since 2020. His research interests include singing voice analysis, musical acoustics, microphone array processing, and directivity analysis.

PAUL ARMIN BEREUTER received the dual master's degree in electrical engineering and audio engineering from the University of Music and Performing Arts Graz, and the Graz University of Technology, in 2022, and currently pursues his Ph.D. degree with the Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, where he is also employed as a Research Assistant. During his master's degree, he was a Student Assistant with the Signal Processing and Speech Communication Laboratory, Graz. His research interests include audio signal processing, music information retrieval, machine learning, and acoustics.



SUDARSANA REDDY KADIRI (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, and the Ph.D. degree from the Department of Electronics and Communication Engineering (ECE), International Institute of Information Technology Hyderabad (IIIT-H), Hyderabad, in 2018. He was a Teaching Assistant for several courses at IIIT-H, from 2012 to 2018. Since 2019, he has been involved in teaching and mentoring activities at Aalto University, Espoo, Finland. He was a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, from 2019 to 2021, where he is currently a Research Fellow. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience. He has published more than 60 research papers in peer-reviewed journals and conferences in his research areas. He is a reviewer of several reputed journals and conferences. He was awarded the Tata Consultancy Services (TCS) Fellowship for his Ph.D. degree.

ALOIS SONTACCHI received the Diploma degree in electrical engineering and audio engineering and the Ph.D. degree in technical science from the Graz University of Technology, in 1999 and 2003, respectively. In 1999, he joined the Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz (KUG). He directed the IEM, from 2010 to 2016. Since 2016, he has been a Professor of acoustics and audio engineering with KUG. His research interests include perceptual evaluation, music information retrieval, and spatial audio signal processing. He is a member of the Audio Engineering Society and the German Acoustical Society and the Scientific Board Member of DAFx and the Ambisonics Symposium.