Analysis and classification of phonation modes in singing

STOLLER, D; Dixon, S; 17th International Society for Music Information Retrieval Conference (ISMIR 2016)

ANALYSIS AND CLASSIFICATION OF PHONATION MODES IN SINGING

Daniel Stoller
Queen Mary University of London
d.stoller@qmul.ac.uk

Simon Dixon
Queen Mary University of London
s.e.dixon@qmul.ac.uk

ABSTRACT

Phonation mode is an expressive aspect of the singing voice and can be described using the four categories *neutral*, *breathy*, *pressed* and *flow*. Previous attempts at automatically classifying the phonation mode on a dataset containing vowels sung by a female professional have been lacking in accuracy or have not sufficiently investigated the characteristic features of the different phonation modes which enable successful classification. In this paper, we extract a large range of features from this dataset, including specialised descriptors of pressedness and breathiness, to analyse their explanatory power and robustness against changes of pitch and vowel. We train and optimise a feed-forward neural network (NN) with one hidden layer on all features using cross validation to achieve a mean F-measure above 0.85 and an improved performance compared to previous work. Applying feature selection based on mutual information and retaining the nine highest ranked features as input to a NN results in a mean F-measure of 0.78, demonstrating the suitability of these features to discriminate between phonation modes. Training and pruning a decision tree yields a simple rule set based only on *cepstral peak prominence* (CPP), *temporal flatness* and *average energy* that correctly categorises 78% of the recordings.

1. INTRODUCTION

Humans have the extraordinary capability of producing a wide variety of sounds by manipulating the complex interaction between the vocal folds and the vocal tract. This flexibility also manifests in the singing voice, giving rise to a large number of different types of expression such as vibrato or glissando. In this paper, we focus on *phonation mode* as one of these expressive elements. Sundberg [22] defines four phonation modes within a two-dimensional space spanned by *subglottal pressure* and *glottal airflow*: *Neutral* and *breathy* phonations involve less subglottal pressure than *pressed* and *flow* phonations, while neutral and pressed phonations have lower glottal airflow than breathy and flow phonations.

The phonation mode is an important part of singing and can be seen as an expressive dimension along with pitch and loudness [23]. This additional degree of control allows more room for interpretation and expression - a breathy voice for example can be used to portray sweetness and sexuality, while pressed voices can seem forceful and tense [20]. In addition to individual differences between singers, the phonation mode tends to vary depending on the musical style, as shown in a study with four different genres by Borch and Sundberg [5]. Automatically detecting the phonation mode could help diagnose certain vocal disorders such as the hypofunction and hyperfunction of the glottis [10]. Because many singing students in particular exhibit varying degrees of these malfunctions throughout the course of their studies, teachers could be assisted to correct this behaviour during lessons. Apart from music, phonation modes also play an important role in speech. For the task of speaker-independent emotion recognition, phonation mode is one of the features of voice quality that can be useful for reliably detecting emotion in speech [16].

2. RELATED WORK

Several studies have investigated phonation modes from a physiological and a signal processing perspective.

By using direct body measurements, Grillo and Verdolini [11] showed that laryngeal resistance as the ratio of subglottal pressure and average glottal airflow can reliably account for the difference between pressed, neutral and breathy phonation, although not between neutral and resonant voice. Subglottal pressure was also found to correlate with the amount of phonatory pressedness in a similar study, along with the closing quotient of the glottis and the difference in amplitudes of the first two harmonics in the voice source spectrum [17]. Without direct body measurements however, it is difficult to estimate subglottal pressure based only on auditory information.

As a result, signal-based feature descriptors have been developed to estimate the degree of pressedness. Most notably, the *normalised amplitude quotient* (NAQ) describes the glottal closing phase and was shown to be more robust than the closing quotient when separating breathy, neutral and pressed spoken vowels [1, 3]. This capability apparently transfers to the singing voice: Given vocal recordings featuring the four different phonation modes rated by a panel of experts, the NAQ accounted for 73% of the variation in the ratings of perceived pressedness [24]. Other descriptors have been proposed for discriminating breathy from tense voices, such as the *peak slope* [14] and the *maxima dispersion quotient* (MDQ) [15]. The *cepstral*

peak prominence (CPP) [12] feature was shown to correlate strongly with ratings of perceived breathiness. In the context of singing however, the suitability of these features to capture the characteristics of all four phonation modes remains largely unknown and is investigated in this paper.

For the automatic detection of phonation modes, a dataset containing vowels sung by a female professional was created [21]. On this dataset, an automatic classification method [20] based on modelling the human vocal tract and estimating the glottal source waveform was developed. The physiological nature of the model can give insight into how humans produce phonation modes by interpreting optimised model parameters. However, only moderate accuracies between 60% and 75% were achieved, despite training the model on each vowel individually, resulting in a less general classification problem where vowel-dependent effects do not have to be taken into account. Another classification attempt on the same dataset using features derived from *linear predictive coding* (LPC) such as formant frequencies achieved a mean F-measure of 0.84 [13] with a *logistic model tree* as classifier. However, the accuracy may be high partly due to not excluding the higher pitches in the dataset, which the singer was only able to produce in breathy and neutral phonation. As a result, only two instead of four classes have to be distinguished in the higher pitch range, incentivising the classifier to extract pitch-related information to detect this situation. Although the authors identify CPP and the difference between the first two harmonics of the voice source as useful features, they do not systematically analyse how their features allow for successful classification to derive an explanation for phonation modes as an acoustic phenomenon.

This paper focuses on finding the features that best explain the differences between phonation modes in the context of singing. We investigate whether individual features, especially descriptors such as NAQ and MDQ, can directly distinguish some of the phonation modes. Different sets of features are constructed and used for the automatic classification of phonation modes to compare their explanatory power. In its optimal configuration, our classifier significantly outperforms existing approaches.

3. DATASET

We use the dataset provided by [21], which contains single sustained vowels sung by a female professional recorded at a sampling frequency of 44.1 KHz. Every phonation mode is reproduced with each of the nine vowels A, AE, I, O, U, UE, Y, OE and E and with pitches ranging from A3 to G5. However, pitches above B4 do not feature the phonation modes flow and pressed. To create a balanced dataset, where all four classes are approximately equally represented for each combination of pitch and vowel, we only use pitches between A3 and B4 and also exclude alternative recordings of the same phonation mode. If not stated otherwise, the balanced dataset called *DS-Bal* is used in this study. The full dataset *DS-Full* is only used to enable a comparison with classification results from previous work [13].

No.	Feature	No.	Feature
F1	MFCC40B	F15	Harmonic 1-6 amp.
F2	MFCC80B	F16	HNR 500
F3	MFCC80B0	F17	HNR 1500
F4	MFCC80BT	F18	HNR 2500
F5	Temp. Flatness	F19	HNR 3500
F6	Spec. Flatness	F20	HNR 4500
F7	ZCR	F21	Formant 1-4 amp.
F8	Spec. Flux Mean	F22	Formant 1-4 freq.
F9	Spec. Flux Dev.	F23	Formant 1-4 bandw.
F10	Spec. Centroid	F24	CPP
F11	HFE1	F25	NAQ
F12	HFE2	F26	MDQ
F13	F0 Mean	F27	Peak Slope
F14	F0 Dev.	F28	Glottal Peak Slope

Table 1. List of features used in this paper. A detailed explanation can be found in section 4.

4. FEATURES

A large number of features listed in table 1 is extracted to facilitate an extensive comparison and evaluation. Apart from the first three features, *trimmed* audio samples containing only the centre 600 ms of every recording are used for extraction to remove potential silences and note transients, keeping the stable part of the phonation and ensuring the reliability of LPC-derived features. For time-dependent features, frames of 50 ms with a Hanning window and 50% overlap are used for extraction before the mean of all frames is calculated, unless otherwise noted. In addition to common spectral features, we include features introduced in section 2 specifically designed to estimate phonatory pressedness or breathiness, because they should be particularly useful in this task.

Mel-frequency cepstral coefficients (MFCCs, F1-F4) are timbre descriptors used widely in MIR and speech research. In this paper, the n-th coefficient of an MFCC vector will be denoted as MFCC(n). The first feature MFCC40B (F1) is a 40-dimensional MFCC vector using the standard number of 40 Mel bands for the summarisation of the spectrum and including the 0-th coefficient representing energy. Presumably, the lower coefficients capture information more relevant to phonation modes, as they encode timbral properties that are independent of pitch. Therefore, we additionally include MFCC80B (F2), which is the first 40 coefficients of the MFCCs computed with 80 instead of 40 Mel bands, giving increased resolution in the lower coefficients. To determine the importance of energy as a feature for successful classification, MFCC80B0 (F3) is introduced as a variant of MFCC80B that is also 40-dimensional, but does not include the 0-th coefficient. As an additional variant of MFCC80B (F2), we extract MFCC80BT (F4), not from the full but from the trimmed recordings of the sung vowels, to investigate the importance of timbral information at the vowel onset and release.

Although MFCCs represent the audio signal very efficiently, as they encode most of the energy in the spec-

trum in the lower coefficients using the discrete cosine transform, they are hard to interpret as they mostly lack an intuitive description. We add a range of spectral features (F5-F12) to allow for a more comprehensible explanation of the phonation mode as an acoustic phenomenon. More specifically, temporal flatness (F5) and spectral flatness (F6) compute the ratio between the geometric and the arithmetic mean of the audio signal in the time and in the frequency domain, respectively, and describe whether the signal is smooth or spiky. The spectral flux is summarised by its mean (F8) and standard deviation (F9). As an estimation of high-frequency energy (HFE), HFE1 (F11) determines the frequency above which only 15% of the total energy resides and HFE2 (F12) calculates the amount of energy present above a frequency of 1500 Hz. We apply the pitch tracking algorithm from [9] and compute the mean (F13) and standard deviation (F14) of the resulting series of pitches to determine the amplitudes of the first six harmonics (F15). As a potential discriminator for the breathy voice, the harmonic-to-noise ratio (HNR) designed for speech signals [8] is extracted for the frequencies below 500 (F16), 1500 (F17), 2500 (F18), 3500 (F19) and 4500 (F20) Hz. Using LPC with a filter of order $\frac{f}{1000} + 2$ and f as the sampling frequency in Hz, we retain the amplitudes (F21), frequencies (F22) and bandwidths (F23) of the first four formants. We further include CPP (F24), NAQ (F25) and MDQ (F26) introduced in section 2. Finally, the peak slope is computed by determining the slope of a regression line that is fitted to the peaks in the spectrum of the audio signal (F27) and the glottal waveform (F28) obtained by the *iterative adaptive inverse filtering algorithm* [2].

5. FEATURE ANALYSIS

5.1 MFCC Visualisation

In contrast to most of the other features listed in table 1, MFCCs (F1-F4) can be difficult to interpret. To make sense of this high-dimensional feature space and how it potentially differentiates phonation modes, we normalise the coefficients in MFCC40B to have zero mean and unit standard deviation across the dataset. The resulting coefficients are visualised in Figure 1, where the recordings on the horizontal axis are grouped by phonation mode and sorted in ascending order of pitch within each group. Within each phonation mode, multiple diagonal lines extending across the 10-th and higher coefficients imply a dependency of these feature coefficients on pitch, which a classifier would have to account for to reach high accuracy. The first 10 coefficients on the other hand do not exhibit this behaviour and also partly differ between phonation modes, especially when comparing breathy to non-breathy phonation. In particular, MFCC40B(0) as a measure of average energy increases in value from breathy, neutral, pressed to flow phonation. Although this visualisation does not reveal dependencies on vowel, it demonstrates the importance of the lower MFCCs and motivates the usage of MFCC80B, as more Mel bands increase the resolution in this range.

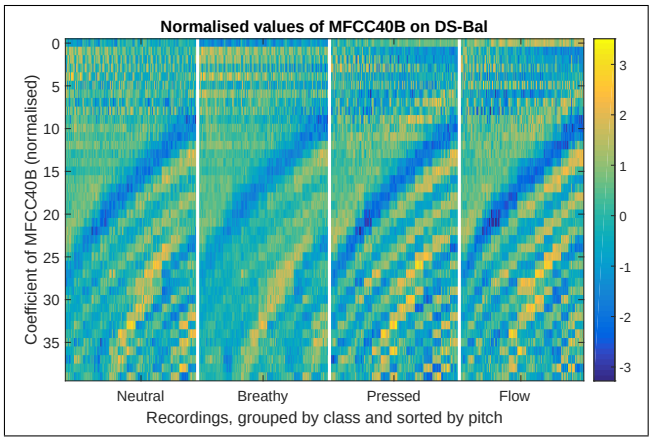

Figure 1. Visualisation of normalised MFCC40B values. Vertical gaps separate the different phonation modes. For each phonation mode, corresponding recordings are sorted in ascending order of pitch.

5.2 Class separation

In this section, we will investigate whether individual feature coefficients can directly separate some of the phonation modes by using an *analysis of variance* (ANOVA). Assuming that of all MFCC variants, MFCC80B is best suited for phonation mode detection, we subject MFCC80B and all remaining features (F5-F28) to an ANOVA with feature coefficients as dependent variables and the four phonation modes as independent categorical variables. The resulting *F-Ratio* equates to the ratio of variance between classes to the variance within classes, indicating how clearly phonation modes are separated by a particular feature.

The ten features with the highest resulting F-Ratios in descending order from $F = 441$ to $F = 213$ are CPP, temporal flatness, MFCC80B(0), MFCC80B(1), spectral flatness, HFE1, MDQ, spectral flux mean and deviation, and MFCC80B(3). However, these features are all mutually correlated with absolute correlation coefficients above 0.5 (mean: 0.77), indicating a large degree of redundancy.

In Figure 2, the distribution of feature values depending on phonation mode is shown for CPP, temporal flatness and MFCC80B(0) as they reach the highest F-Ratios, and for MDQ because it correlates least with the three aforementioned features. Figure 2 (a) demonstrates that CPP separates not only breathy from all other phonation modes significantly, as expected due to its design as a measure of breathiness [12], but can also distinguish neutral from pressed and flow phonation and to some degree pressed from flow phonation. Regardless of its simplicity, temporal flatness shown in Figure 2 (b) manages to clearly separate neutral and breathy from pressed and flow phonation. MFCC80B(0) shown in Figure 2 (c) confirms the finding from section 5.1 that each phonation mode features a different loudness on average. Interestingly, MDQ plotted in Figure 2 (d) behaves similar to temporal flatness shown in Figure 2 (b) and does not separate the classes more clearly despite its comparatively complex design intended to directly quantify the degree of pressedness.

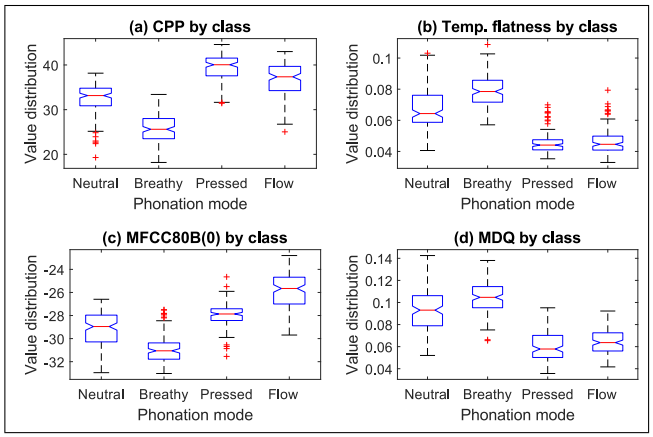

Figure 2. Distributions of (a) CPP, (b) temporal flatness, (c) MFCC80B(0) and (d) MDQ for the four phonation modes.

Regarding the other features, HNR behaves as expected and successfully separates breathy phonation from all other phonation modes, with a cut-off of 2500 Hz (F18) achieving the best F-Ratio of 97.8, but does not differentiate between the remaining phonation modes. Contrary to its purpose of estimating pressedness, NAQ surprisingly exhibits only small differences in mean values and large overlaps of the distributions between different phonation modes ($F = 6.48$), possibly because it was originally proposed for speech [3]. Apart from slightly higher values of peak slope for breathy voices, the feature proves to be uninformative ($F = 22.75$), despite obtaining good results on speech excerpts [14]. Finally, distributions of glottal peak slope for the phonation modes are not significantly different ($F = 0.42$).

In general, separating breathy and neutral phonation from pressed and flow phonation is more readily achieved by individual features than distinguishing pressed from flow phonation. Therefore, we perform the same analysis with only pressed and flow phonation as possible categories of the independent variable to find features that make this particularly difficult distinction. As a result, MFCC80B(0) achieves by far the largest separation ($F = 183.23$), which is hinted at in Figure 2 (c), followed by MFCC80B(1) ($F = 44.51$). Apart from CPP ($F = 38.58$) shown in Figure 2 (a) and MFCC80B(10) ($F = 31.72$), all remaining features exhibit F-Ratios below 15.

5.3 Robustness against pitch and vowel changes

We investigate the robustness of individual features against changes of pitch and vowel by performing an ANOVA with pitch and vowel respectively as independent variables, with one class for each unique pitch or vowel present in the dataset. As well as the formant-based features (F21-F23), the lower MFCC80B coefficients between approximately 4 and 17 are dependent on vowel, a dependency not immediately visible in Figure 1. HFE2 with an F-Ratio of 32.03 is more dependent on vowel than the alternative HFE1 feature ($F = 9.16$), further corroborating the superiority of HFE1 over HFE2 for phonation mode detection. Other particularly vowel-dependent features are the amplitude of the third harmonic ($F = 26.68$) and peak slope ($F = 45.04$). Regarding pitch, dependencies were found in MFCC80B confirming the interpretation of Figure 1, starting with coefficient 18 and increasing in F-Ratio until coefficient 30, where it remains constant for the coefficients 30 to 40. Except for F0 Mean as an estimate of pitch, no other significant dependencies were found, allowing for the construction of a classifier that is mostly robust against pitch changes.

5.4 A simple rule set to explain phonation modes

In this section, possible interactions between features that could explain differences in the phonation modes are analysed to derive a comprehensible rule set that correctly categorises most of the recordings. We construct a decision tree with *Gini's diversity index* [6] as split criterion and prune it so it has only three decision nodes. The result is the following set of rules using only temporal flatness, CPP and the MFCC80B(0) for distinguishing the phonation modes:

- Neutral and breathy phonation have higher temporal flatness (greater than $0.055 = 47$th percentile) than pressed and flow phonation
- Neutral phonation has higher CPP (greater than $29.97 = 30$th percentile) than breathy phonation
- Flow phonation has a higher MFCC80B(0) (greater than $-26.37 = 84$th percentile) than pressed

The above rules assign the correct class to 78% of the recordings in the dataset, thus offering a simple explanation for the main differences between the phonation modes.

6. CLASSIFICATION

6.1 Feature Sets

The eight feature sets listed in table 2 are constructed for training the classifier. The first four feature sets exclusively use the MFCC variants (F1-F4) from section 4. FS5 contains all features except the MFCC variants (F1-F4), while FS6 combines the MFCC variant yielding the best classification accuracy (F2) with all other features (F5-F28).

In the search for a low-dimensional feature representation, we apply *Principal component analysis* (PCA) to the features in FS6. The resulting principal components sorted in descending order of their eigenvalues constitute feature set FS7, for which classification performance will be assessed when including only the first D dimensions. Principal components can be difficult to interpret, because each represents a combination of different features. Therefore, we employ a feature selection method based on mutual information [19] to retrieve a ranking of the dimensions in FS6, enabling the construction of an optimal feature set of dimensionality D with the D highest-ranked feature coefficients. As a result, FS8 contains all feature dimensions from FS6 sorted in descending order of rank.

Name	List of features	Dimensions
FS1	MFCC40B (F1)	40
FS2	MFCC80B (F2)	40
FS3	MFCC80B0 (F3)	40
FS4	MFCC80BT (F4)	40
FS5	Features 5 to 28	38
FS6	FS2 and FS5	78
FS7	FS6, PCA-transformed	78
FS8	FS6, sorted by feature selection	78

Table 2. Feature sets used for classification.

6.2 Method

Feed-forward neural networks (NNs) are used for classification, as they are robust against noise and correlated inputs. We use one hidden layer with a variable number of neurons N, and a *soft-max* output layer. Cross-validation is employed that splits the dataset into 10 evenly distributed subsets, using every combination of two subsets as test and validation set with the remaining 8 subsets as training data, resulting in $10 \cdot 9$ iterations. For training, *stochastic gradient descent* is used to minimise *cross-entropy error* after semi-randomly initialising the network weights with the *Nguyen-Widrow initialisation method* [18].We describe the overall performance with the mean F-measure obtained over all cross-validation iterations.

To find the optimal number of hidden neurons N for every feature set, we determine the mean F-measure achieved for every $N \in \{1, \ldots, 40\}$. To obtain a compact set of features that yields high accuracy, we also optimise the mean F-measure achieved when using only the first D dimensions in the feature sets FS1 to FS4 as well as FS7 and FS8, resulting in a grid search with the number of neurons N and the number of features D as parameters.

6.3 Results

The classification results obtained with varying numbers of hidden neurons N and dimensions D using the feature sets FS1, FS2, FS7 and FS8 are visualised in Figure 3. We excluded the feature sets FS3 and FS4 due to their similar behaviour compared to FS1 and FS2, and FS5 as well as FS6 because only the number of neurons N was varied. With $N < 4$ neurons in the hidden layer, mean F-measures remain at low levels for every feature set regardless of the number of dimensions. Performance with more neurons N improves gradually when using an increasing number of MFCCs, as Figures 3 (a) and (b) demonstrate. The rate of this increase becomes less pronounced for higher MFCCs, implying that the differences in phonation mode are mostly encoded by approximately the first 20 MFCCs. FS2 containing MFCC80B shown in Figure 3 (b) however reaches significantly higher mean F-measures with the same number of coefficients than FS1 comprised of MFCC40B in Figure 3 (a). An increased frequency resolution of the cepstrum representation could be an explanation, as it leads to a more precise description of the relevant low-frequency components in the spectrum. Applying PCA does not lead

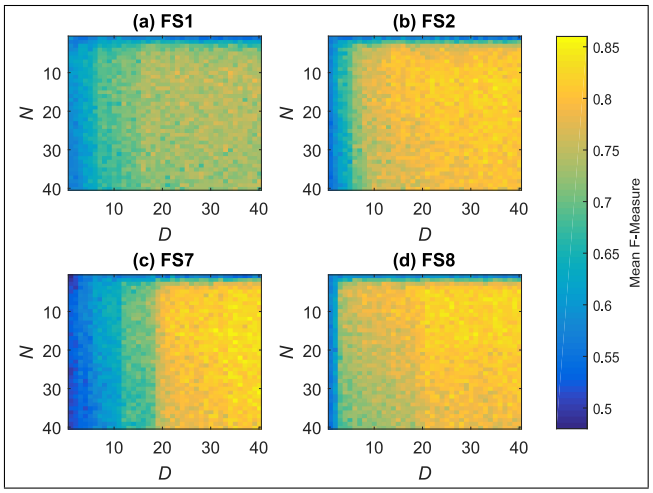

Figure 3. Mean F-measures when using a different number of neurons N and the first D dimensions in the feature sets (a) FS1, (b) FS2, (c) FS7 and (d) FS8.

to a drastically reduced dimensionality of the feature space without a major degradation in performance: Including the first D principal components of feature set FS7 only results in moderate performance for $D < 19$. One reason could be an intrinsically high dimensionality of the feature space, corroborated by the requirement of 32 principal components to explain 95% of the variance. Additionally, the first principal components could encode mostly pitch- and vowel-dependent variances in feature values instead of changes induced by different phonation modes. In contrast, feature selection considers how informative each feature dimension is for classification. As a result, including only the first few dimensions of FS8, which were ranked highest by feature selection, yields high mean F-measures as shown in Figure 3 (d).

Generally, the mean F-measure is subject to considerable variance due to the random selection of subsets performed by cross-validation. Because this impedes the robust selection of the optimal parameters, we interpolate the mean F-measures using *locally weighted regression* [7] with a *span* of 0.1, meaning 10% of the data points along each dimension nearest to an interpolated point determine its position. Intended as a trade-off between classification accuracy and model complexity, we define the optimal combination of parameters N and D as

$$(N_{\text{opt}}, D_{\text{opt}}) = \underset{(N,D)}{\operatorname{argmin}}\{N + D \mid (N, D) \in \mathcal{C}\}, \quad (1)$$

where $\mathcal{C}$ is the set of parameter configurations for which neither adding a dimension nor a hidden neuron increases the smoothed F-measure $s(N, D)$ more than a threshold t:

$$\mathcal{C} = \{(N, D) \mid s(N + 1, D) - s(N, D) < t \quad (2)$$

$$\wedge\, s(N, D + 1) - s(N, D) < t\}. \quad (3)$$

For $t = 0.001$, the mean F-measures for the optimised parameter settings are shown in table 3, including the 95% confidence for the maximum deviation of the mean in both directions, calculated as 1.96 times the standard error of

Feature set	N_{opt}	D_{opt}	Mean F-m.	$1.96 \cdot$ SEM
FS1	10	18	0.7403	0.027
FS2	8	15	0.7965	0.026
FS3	12	17	0.7948	0.027
FS4	9	21	0.7358	0.028
FS5	9	-	0.7681	0.023
FS6	9	-	0.8501	0.024
FS7	9	26	0.8050	0.025
FS8	9	24	0.8302	0.026

Table 3. Classification results for each feature set after optimising the number of neurons N and dimensions D.

the mean (SEM), to determine whether two classification results are significantly different from each other. For every feature set, at least moderate performance is achieved with only a low number of neurons. The usage of 80 Mel bands in MFCC80B (FS2) results in a significantly higher F-measure compared to the standard MFCC40B feature (FS1) with 40 Mel bands. Although FS3 exhibits lower performance with only very few coefficients, removing MFCC80B(0) does not decrease accuracy compared to FS2, perhaps because its correlation with higher MFCCs can be used to gain very similar information. The MFCCs appear to capture relevant timbral information present at the vowel onsets and releases, as the decreased performance for FS4 using the trimmed recordings shows. The interpretable, 38-dimensional feature set FS5 obtains moderate accuracy, but the best mean F-measure of 0.8501 is reached with the 78-dimensional feature set FS6. Application of PCA (FS7) and feature selection based on mutual information (FS8) on feature set FS6 lead to only slightly reduced performance with fewer dimensions. Feature selection is particularly successful, allowing us to construct a NN with only four hidden neurons and the MFCC80B(0), MFCC80B(1), MFCC80B(12), spectral flux mean and deviation, HNR 3500 and temporal flatness as features that still achieves a mean F-measure of 0.78 (SEM = 0.027).

To compare performance, we use the full dataset DS-Full on which the best mean F-measure of 0.84 was achieved by [13], and train a NN in the same manner as described in section 6.2. FS6 is chosen for this experiment, as it exhibits the best performance on the dataset DS-Bal. With $N = 13$ neurons, a mean F-measure of 0.868 with an SEM of 0.015 is obtained, leading to a 95% confidence interval of $[0.846, 0.890]$ for the F-measure and proving a significant improvement over the previously achieved mean F-measure of 0.84.

7. DISCUSSION AND OUTLOOK

Although designed specifically to only determine the amount of breathiness, CPP manages to separate each phonation mode best out of all features ($F = 441$). MDQ reliably distinguishes pressed and flow phonation from neutral and breathy phonation ($F = 310$), but has a correlation of 0.66 with temporal flatness, which achieves a better direct class separation ($F = 394$) with a simpler approach. Peak slope ($F = 22.75$) and glottal peak slope ($F = 0.65$) show weak discriminative power, in contrast to previous work [14]. The same applies to NAQ ($F = 6.48$), which contradicts previous literature demonstrating its suitability to measure the degree of pressedness [1, 3, 24] and warrants further investigation.

Contrary to the assumption in [20] that MFCCs are inapt for phonation mode classification, the lower coefficients alone lead to commensurate performance despite their dependence on vowel. Our classifier is mostly able to account for these effects, but an investigation into how class separation is exactly achieved, for example by using rule extraction from NNs [4], remains for future work. The increase in performance with MFCCs when including the full recording (MFCC80B) instead of an excerpt (MFCC80BT) demonstrates the relevance of timbral information at the vowel onset and release, but a more detailed analysis is needed to find the underlying cause. We show that using 80 Mel bands further increases performance, revealing the importance of optimising this parameter in future work. Every phonation mode features a different loudness, as indicated by MFCC80B(0) as a measure of average energy ($F = 352$). Loudness could vary strongly in more realistic singing conditions and between different singers, therefore making loudness-based phonation mode detection not very generalisable and adaptive to other scenarios.

Overall, the dataset has severe limitations, which reduces the generalisability of classifiers trained on this data: Because it only contains one singer, detection could be using singer-specific effects leading to decreased performance when confronted with other singers. Classification also has to be extended to work on full recordings of vocal performances instead of only isolated vowels. Finally, the recordings in the dataset are monophonic unlike many real-world music pieces, for which performance could be reduced due to the additionally required singing voice separation. Considering this is the only publically available dataset known to the authors that includes annotations of phonation mode, the development of larger, more comprehensive datasets for phonation mode detection seems critical for future progress on this task.

8. CONCLUSION

In this paper, we investigated the discriminative and explanatory power of a large number of features in the context of phonation mode detection. CPP, temporal flatness and MFCC80B(0) representing average energy were found to separate the phonation modes best, as they can correctly explain the phonation mode present in 78% of all recordings. Contrary to previous work, NAQ, peak slope and glottal peak slope did not separate phonation modes well. MFCCs lead to good classification accuracy using NNs as shown in section 6, particularly when using 80 instead of 40 Mel bands. The highest mean F-measure of 0.85 is achieved on the balanced dataset DS-Bal when using all features, demonstrating their explanatory power and the success of our classifier. On the dataset DS-Full, we attain an F-measure of 0.868, thereby significantly outperforming the best classifier from previous work [13].

9. REFERENCES

[1] Matti Airas and Paavo Alku. Comparison of multiple voice source parameters in different phonation types. In *8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1410–1413, 2007.

[2] Paavo Alku. An automatic method to estimate the time-based parameters of the glottal pulseform. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 29–32, 1992.

[3] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[4] Robert Andrews, Joachim Diederich, and Alan B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, 8(6):373–389, 1995.

[5] Daniel Z. Borch and Johan Sundberg. Some phonatory and resonatory characteristics of the rock, pop, soul, and swedish dance band styles of singing. *Journal of Voice*, 25(5):532 – 537, 2011.

[6] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and regression trees*. CRC press, 1984.

[7] William S. Cleveland and Susan J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[8] Guus de Krom. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, 36(2):254–266, 1993.

[9] Thomas Drugman and Abeer Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1973–1976, 2011.

[10] Emil Froeschels. Hygiene of the voice. *Archives of Otolaryngology*, 38(2):122–130, 1943.

[11] Elizabeth U. Grillo and Katherine Verdolini and. Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects. *Journal of Voice*, 22(5):546 – 552, 2008.

[12] James Hillenbrand, Ronald A. Cleveland, and Robert L. Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778, 1994.

[13] Léonidas Ioannidis, Jean-Luc Rouas, and Myriam Desainte-Catherine. Caractérisation et classification automatique des modes phonatoires en voix chantée. In *XXXèmes Journées d'études sur la parole*, 2014.

[14] John Kane and Christer Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 177–180, 2011.

[15] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179, 2013.

[16] Marko Lugger and Bin Yang. The relevance of voice quality features in speaker independent emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–17–IV–20, April 2007.

[17] Moa Millgård, Tobias Fors, and Johan Sundberg. Flow glottogram characteristics and perceived degree of phonatory pressedness. *Journal of Voice. Article in press. DOI: http://dx.doi.org/10.1016/j.jvoice.2015.03.014*, 2015.

[18] Derrick Nguyen and Bernard Widrow. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *IJCNN International Joint Conference on Neural Networks*, pages 21–26 vol.3, June 1990.

[19] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[20] Polina Proutskova, Christophe Rhodes, Tim Crawford, and Geraint Wiggins. Breathy, resonant, pressed automatic detection of phonation mode from audio recordings of singing. *Journal of New Music Research*, 42(2):171–186, 2013.

[21] Polina Proutskova, Christophe Rhodes, Geraint A. Wiggins, and Tim Crawford. Breathy or resonant - A controlled and curated dataset for phonation mode detection in singing. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 589–594, 2012.

[22] Johan Sundberg. *The science of the singing voice*. Illinois University Press, 1987.

[23] Johan Sundberg. What's so special about singers? *Journal of Voice*, 4(2):107 – 119, 1990.

[24] Johan Sundberg, Margareta Thalén, Paavo Alku, and Erkki Vilkman. Estimating perceived phonatory pressedness in singing from flow glottograms. *Journal of Voice*, 18(1):56–62, 2004.