ANALYSIS AND CLASSIFICATION OF PHONATION MODES IN SINGING

Daniel Stoller and Simon Dixon d.stoller@qmul.ac.uk

Queen Mary University of London

centre for digital music

Contributions

We analyse the suitability of a large range of features to distinguish between different **phonation modes** as an expressive aspect of the singing voice.

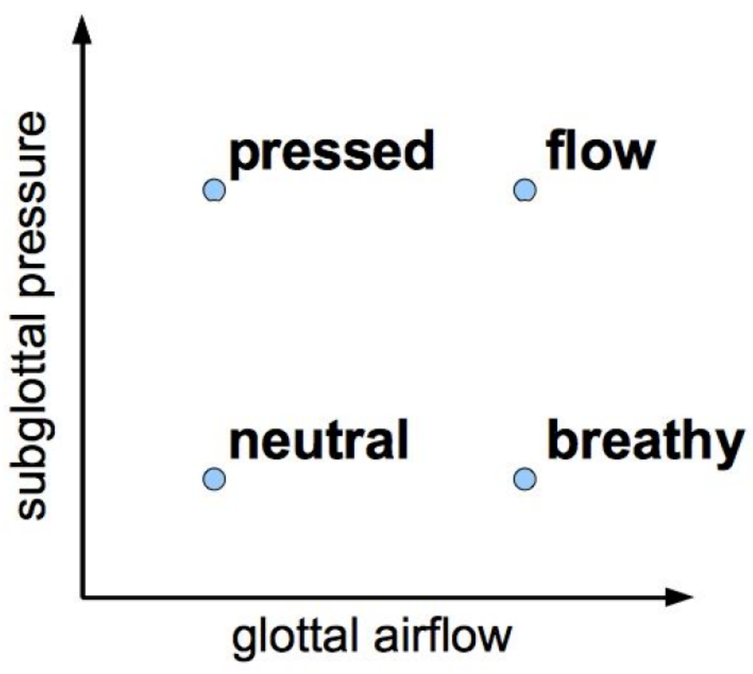
On a dataset with annotated recordings of sustained vowels, we derive a simple rule set based only on **cepstral peak prominence (CPP)**, **temporal flatness** and the **average energy** that separates the phonation modes in 78% of cases.

Among the investigated descriptors of breathiness and pressedness, the **normalised amplitude quotient (NAQ)** and the **(glottal) peak slope** are surprisingly uninformative, in contrast to **CPP** and the **maxima dispersion quotient (MDQ)**.

Accurate classification is achieved using neural networks with only a few hidden neurons and feature coefficients, and we outperform the state of the art using all features.

Phonation modes

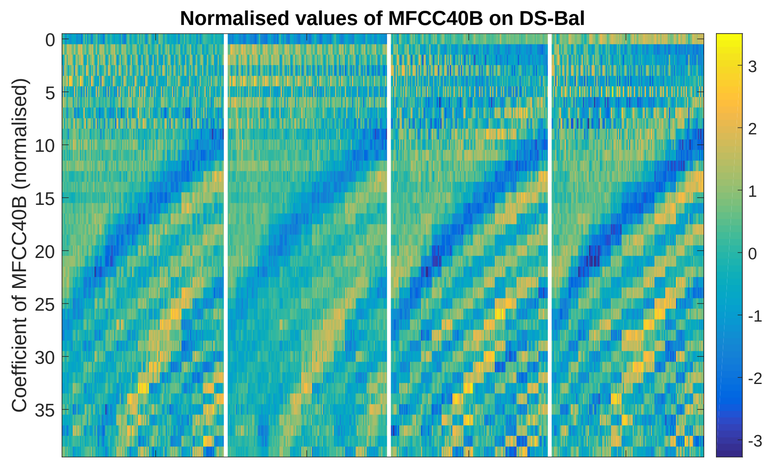
Sundberg [1] defines **four phonation modes** which span the quadrants of the pressure-airflow plane:



Dataset

We use the dataset provided by Proutskova et al. [2], which contains single recordings of isolated sustained vowels sung by a female professional. 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.

MFCC visualisation



The pitch dependency in the higher and the class differences in the lower coefficients motivate the use of a higher frequency resolution (80 Mel bands).

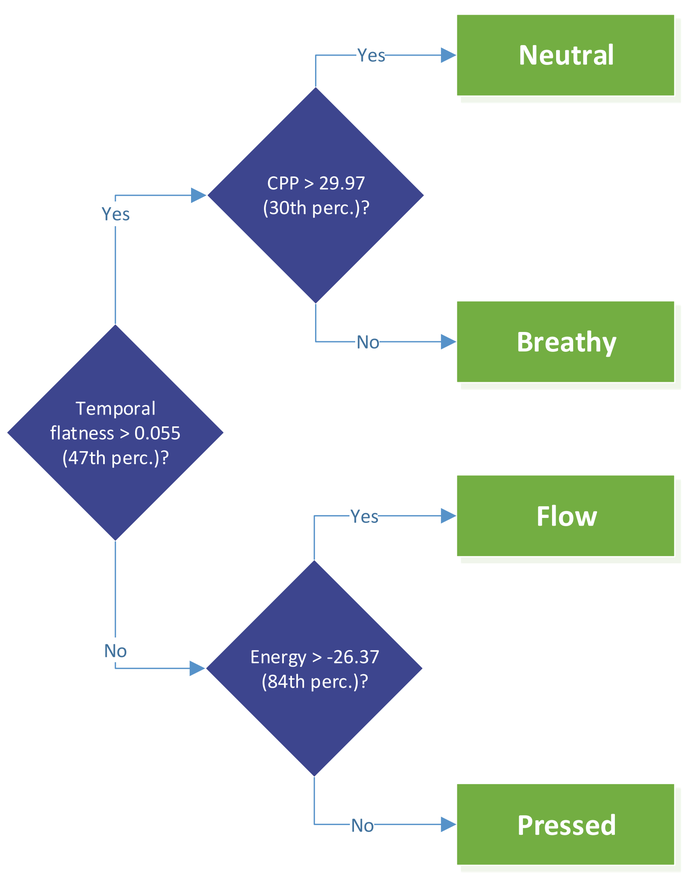
Feature list

Apart from the first three features, *trimmed* audio samples containing only the centre 600 ms of every recording are used for extraction to keep only the stable part of phonation. Feature values are averaged for each sample.

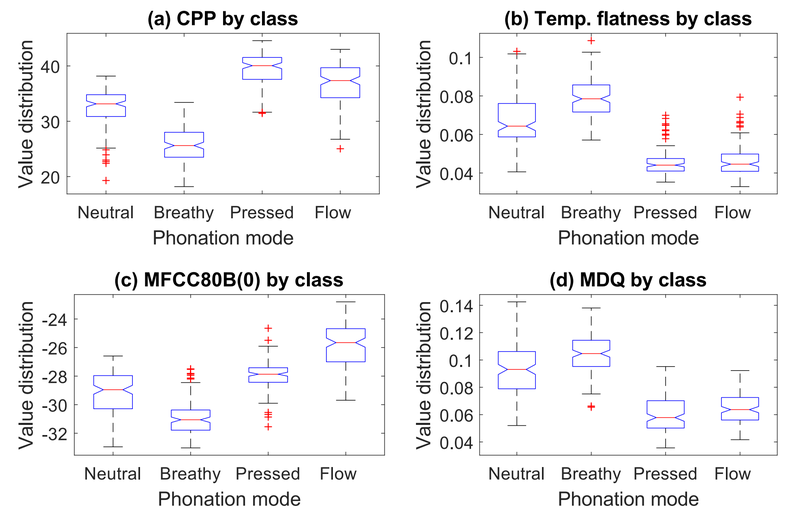
No.	Feature	No.	Feature
F1	MFCC40B	F15	Harmonic 1-6 amp.
F2	MFCC80B	F16	HNR 500
F3	MFCC80B0	F17	HNR 1500
F4	MFCC80BT	F18	HNR 2500
F5	Temp. Flatness	F19	HNR 3500
F6	Spec. Flatness	F20	HNR 4500
F7	ZCR	F21	Formant 1-4 amp.
F8	Spec. Flux Mean	F22	Formant 1-4 freq.
F9	Spec. Flux Dev.	F23	Formant 1-4 bandw.
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

A simple explanation of phonation

Constructing and pruning a decision tree on the dataset leads to a simple ruleset correctly categorising 78% of the audio samples:



Class separation

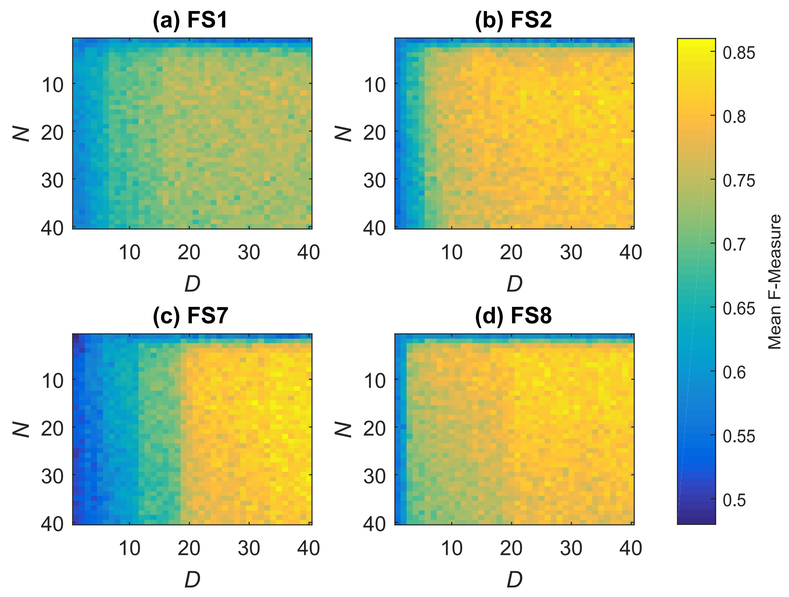


Classification

Feature sets used:

Name	List of features	Dimensions
FS1	MFCC40B (F1)	40
FS2	MFCC80B (F2)	40
FS3	MFCC80B0 (F3)	40
FS4	MFCC80BT (F4)	40
FS5	Features F5 to F28	38
FS6	FS2 and FS5	78
FS7	FS6, PCA-transformed	78
FS8	FS6, sorted by feature selection	78

Mean F-Measure achieved with single-layer neural network depending on number of hidden neurons N with first F features as input:



Minimal number of hidden neurons N_{opt} and features D_{opt} for which performance does not increase significantly when adding more neurons/features:

Feature	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

References

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

[2] 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.