# Prediction of three articulatory categories in vocal sound imitations using models for auditory receptive fields

Anders Friberg,[1,a)] Tony Lindeberg,[2] Martin Hellwagner,[1] Pétur Helgason,[1] Gláucia Laís Salomão,[1] Anders Elowsson,[1] Guillaume Lemaitre,[3] and Sten Ternström[1]

[1]*Speech, Music and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Lindstedtsvägen 24, 10044 Stockholm, Sweden*

[2]*Computational Brain Science Lab, Computational Science and Technology, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Lindstedtsvägen 5, 10044 Stockholm, Sweden*

[3]*Institute for Research and Coordination in Acoustics and Music, 1 Place Igor Stravinsky, Paris 75004, France*

(Received 2 February 2018; revised 31 July 2018; accepted 16 August 2018; published online 19 September 2018)

Vocal sound imitations provide a new challenge for understanding the coupling between articulatory mechanisms and the resulting audio. In this study, the classification of three articulatory categories, *phonation*, *supraglottal myoelastic vibrations*, and *turbulence*, have been modeled from audio recordings. Two data sets were assembled, consisting of different vocal imitations by four professional imitators and four non-professional speakers in two different experiments. The audio data were manually annotated by two experienced phoneticians using a detailed articulatory description scheme. A separate set of audio features was developed specifically for each category using both time-domain and spectral methods. For all time-frequency transformations, and for some secondary processing, the recently developed Auditory Receptive Fields Toolbox was used. Three different machine learning methods were applied for predicting the final articulatory categories. The result with the best generalization was found using an ensemble of multilayer perceptrons. The cross-validated classification accuracy was 96.8% for phonation, 90.8% for supraglottal myoelastic vibrations, and 89.0% for turbulence using all the 84 developed features. A final feature reduction to 22 features yielded similar results. © *2018 Acoustical Society of America*.
https://doi.org/10.1121/1.5052438

[JFL] Pages: 1467–1483

## I. INTRODUCTION

By imitating sounds with the voice, humans can portray the surrounding world and convey meaningful information; e.g., the noise of a broken engine in the neighbor's lawn mower. Although these sounds have articulatory similarities to speech and singing, they also have many unique characteristics. As these sounds have not been studied extensively in the past, they represent a new interesting research field, combining phonetics, vocal production and audio recognition.

This study was part of the EU project SkAT-VG, the goal of which was to provide tools for "vocal sketching" of sounds, for the purpose of facilitating sound design. Conventionally, product designers sketch with pencil on paper. Sound designers need a similarly effective tool, and a given candidate is one's own voice, hence the term "vocal sketching." In previous experiments within the project (Lemaitre *et al.*, 2016a, 2016b, 2017), a large number of vocal imitations were collected. For example, Lemaitre *et al.* (2016b) showed that listeners could effectively recognize which sounds the vocal imitations were referring to, which suggests that vocal imitations of everyday sounds convey the sound features that are necessary for sound identification.

These vocal imitations consist of imitations of either basic mechanical interactions (e.g., hitting a table, pouring water) or manufactured products (e.g., vehicles, domestic appliances, video game sounds). The recordings were further annotated by experienced phoneticians according to the detailed articulatory function. In order to incorporate vocal sketching into computer tools for sound design, the SkAT-VG researchers deemed it necessary to understand the imitations from an articulatory point of view, and to develop computer tools that analyze them in such terms. Therefore, in this study, we wanted to explore if some of the manual annotations of articulations could be predicted from the audio recordings.

A key aspect for quantifying and understanding voice, speech and music has been the development of specific audio features that provide relevant information from a recorded audio signal. Therefore, a large number of features have previously been developed (see overview in Alías *et al.*, 2016), ranging from low-level physical descriptions such as zero-crossing rate (e.g., Burred and Lerch, 2004) to perceptually modeled features such as musical speed (e.g., Elowsson *et al.*, 2013). In addition, voice specific features have been developed for clinical applications, with the purpose of characterizing voice qualities commonly perceived in dysphonic voices, such as hoarseness, breathiness, or roughness. These features include jitter, shimmer, cepstral peak prominence

a)Electronic mail: afriberg@kth.se

(CPP), and noise-to-harmonic ratio (NHR). To a varying extent, they can predict perceptually estimated voice quality measures. For example, Gorham-Rowan and Laures-Gore (2004) found moderate correlations between the perception of hoarseness and breathiness versus several acoustic measures including NHR, amplitude perturbation quotient, and fundamental frequency standard deviation. Carding *et al.* (2004), found that jitter, shimmer, and NHR were shown to have a low or moderate reliability and effect size relating to voice quality when comparing dysphonic patients before and after treatment. The CPP measure was found to predict perceived breathiness to a rather large extent (Hillenbrand *et al.*, 1994) and correlated well with dysphonia (Heman-Ackah *et al.*, 2002). In a meta-study, summarizing 25 studies which compared perceived voice qualities with acoustical features, Maryn *et al.* (2009) found that only six out of 39 acoustical features were considered "superior," meaning that they had an average correlation to voice quality higher than $r = 0.6$. The CPP measure had the overall best average correlation with $r = 0.88$.

Given that the present dataset contains sounds produced by unusual combinations of vocal production mechanisms, not found in singing or speaking, and that previous voice features had limited success in describing voice quality (with the exception of, e.g., CPP), it was natural to develop new features that were tailor-made for each articulation category. For example, Lemaitre *et al.* (2016b) studied how human imitators imitated basic acoustic parameters (pitch, attack time, spectral centroid, etc.). For some of these parameters, general-purpose audio features (Peeters *et al.*, 2011) were sufficient to describe some of the characteristic features of the referent sounds. For some other cases, specific features had to be defined to analyze the unique articulatory mechanisms used by the imitators. In addition, we had the possibility to use the recently developed Auditory Receptive Fields (ARF) Toolbox, which provides a new starting point for audio analysis in general (Lindeberg and Friberg, 2015a,b).

The purpose of the present study was to predict three articulatory categories from recorded audio of vocal imitations. The approach was not to find a specific descriptor for each articulation category, but rather to define a range of features for each category that potentially could be used in a subsequent machine learning step in order to predict the final articulation. The focus was on the development of new features for voice analysis using the ARF Toolbox. We start by describing the data set, give an overview of the auditory receptive fields methods, the specific methods used to define audio features, and finally present the results.

## II. DATA SET

### A. French recordings

Four French imitators (two male, 21 and 39 years old; two female, 21 and 41 years old) with normal hearing were selected for the experiment. None of them had received formal training in music, audio, dance, or theater, nor any practice of vocal imitation or Foley artistry.

There were 52 referent sounds selected from three families. The first family (20 sounds) consisted of *basic mechanical interactions*: a hit on a board, the friction of a wheel on the ground, aerodynamic turbulences, water dripping, etc. The selection balanced an equal number of sounds produced by solid objects, liquids, and gases (Lemaitre *et al.*, 2010). The second family of sounds (20 sounds) focused on the sounds of *manufactured products*: vehicles (cars, buses, motorcycles), domestic appliances (refrigerator, etc.), and alarms. The third family (12 *abstract* sounds) included artificial sounds recorded from human computer interfaces (mobile phones, video games, computer operating systems) or synthesized.

The imitators used a custom-made Max/MSP v.6.1 (Ircam/Cycling74) user interface and were seated in a double-walled IAC sound isolated booth. The setup included a microphone (DPA Microphones, model d:fine omni), and an audio interface (RME model Fireface 800). The imitators was recorded at a sampling rate of 64 kHz, in 16-bit PCM WAV files. The user interface allowed the imitators to listen and compare the referent sound, record and play back an imitation. The imitators were alone during the recording session. They were instructed to provide an imitation in such a way that someone listening to them would be able to identify the sounds within the family. The imitators were instructed not to use any conventional onomatopoeia. There was a limit of five trials for each recording. We considered only the last trial, thus, resulting in 208 imitations in total.

### B. Swedish recordings

Four Swedish imitators (two male, 25 and 48 years old; two female, both 25 years old) took part in the experiment. All were professional improvisational actors, recruited through an agency and paid for their participation.

In total, the Swedish recordings comprised a total of 200 imitations, elicited using 50 referent sounds. The referent sounds were selected from the same three basic families as those selected for the French recordings: *basic mechanical interactions* (17 sounds), *manufactured products* (20 sounds), and *abstract* (13 sounds). For the Swedish data, the selection of referent sounds was also guided by the major articulatory mechanisms that they were likely to elicit, such that each of the major mechanisms would have adequate representation in the data.

The referent sounds were presented using a custom-made Max/MSP user interface, similar to the one used for the French recordings. The data were recorded in a soundproofed booth. The audio signal was recorded using a miniature boom microphone (DPA Microphones, model 4066) and a digital audio interface (RME, model Fireface UFX) and was recorded at a sampling rate of 48 kHz in 24-bit PCM WAV files. Some periodic sounds can be imitated with oscillation of other tissues than the vocal folds. To obtain some indication of the occurrence specifically of vocal fold vibration, an electroglottographic (EGG) signal (which measures the amount of contact between the vocal folds) was recorded as well, using a dual-channel EGG device (Glottal Enterprises model MC2-1). In addition, two video streams were recorded, one on a Canon Legria G30 at a frame rate of 50 fps, and the second on a Hero GoPro3 +at 100 fps. As in

the case of the French recordings, the user interface allowed the imitators to play referent sounds at will, as well as listen to their efforts at imitating the referent sounds.

A more detailed description of the recording procedure and the selection of referent sounds for both the French and the Swedish recordings is found in Lemaitre *et al.* (2015) and in Ternström and Mauro (2015).

### C. Annotations

The annotation of the combined database (408 imitations in total) was performed by two experienced phoneticians (co-authors P.H. and G.L.S.) using the software program ELAN (version 4.9.2, Brugman and Russel, 2004), an annotation tool that allows one to create, edit, visualize, and search annotations for video and audio data. As ELAN supports the display of speech and video signals, together with the corresponding annotations, it was possible to synchronize complementary signal sources (audio, video, and EGG) for more robust analysis of the data.

In the database, eight separate articulatory/phonatory variables were annotated by hand: airstream mechanism, vocal fold activity, epilaryngeal activity, velopharyngeal activity, lip manner of articulation, tongue manner of articulation, place of tongue constriction, and tongue shape. For the purposes of the present study, three main articulatory/phonatory categories were extracted from the database using scripted queries:

(1) *Vocal fold phonation*,
(2) *Supraglottal myoelastic vibration (SMV)*,
(3) *Turbulence*.

The choice of these three main categories was motivated from a previous exploratory analysis of the articulatory characteristics of the vocal sound imitations, made within the SkAT-VG project. This suggested that different combinations of these three main source mechanisms would suffice for a broad description of the majority of the imitated sounds. These mechanisms correspond to different kinds of modulations of the airflow coming from the lungs, namely, phonation, i.e., the vibration of the vocal folds, causing a periodic modulation of the glottal area; the vibration of other structures, above the glottis; and the creation of turbulence in the airflow, usually in the vicinity of a constriction or an obstacle somewhere in the vocal tract. Each of these categories was further divided into subcategories and extracted with the scripted queries from the original annotations. Note that these subcategories were not used directly for the final prediction, which was limited to the three main categories. The subcategories are presented here in order to describe the nature and origin of the different phonation types in more detail. These subcategories were used both in the annotation and extraction of examples, and for developing the features in Sec. IV.

(1) The "vocal fold phonation" category had seven subcategories:

- no vocal fold phonation (0),
- breathy voice (1),
- falsetto voice (2),
- modal voice (3),
- pressed voice (4),
- creaky voice (5),
- unspecified vocal fold phonation (6).

The numbers in parentheses refer to the coding of the audio excerpts as described below. The voice qualities associated with the subcategories breathy, falsetto, modal, pressed, and creaky phonation (1 to 5) are extensively described in the literature [cf., e.g., Laver (1980) for an overview]. In addition, we assigned category (6), "unspecified vocal fold phonation," to instances of fairly high pitched, quasi-periodic vibrations that did not fit any of the more established categories above and is not found in linguistic descriptions of voice quality. These were typically short (less than 150 ms) and occurred predominantly in association with the onset or offset of vocal fold phonation.

(2) The category "supraglottal myoelastic vibrations" (SMV) had eight subcategories:

- no vibration present (0),
- lax labial vibration (1),
- lax tongue tip vibration (2),
- lax uvular vibration (3),
- epilaryngeal vibration (4),
- velic vibration (5),
- tense labial vibration (6),
- tense dorsal vibration (7).

In the phonetic literature, the lax categories (1), (2), and (3) are referred to as "trills" that can be produced at different places of articulation (Ladefoged and Maddieson, 1996). Typically, trills are produced with a fairly lax stricture resulting in a cycle rate of 20–40 Hz. The subcategory "epilaryngeal" vibration (4) refers to constrictions in the lower pharyngeal region that can induce tissue vibration. In using the term epilaryngeal, we follow Moisik (2013, p. 91ff), highlighting the encompassing structure for this vibratory process rather than the exact way in which the stricture is made. Typically, the frequency of such epilaryngeal vibrations in speech ranges between 40 and 100 Hz (Moisik *et al.*, 2010; Moisik, 2013, p. 126ff). These factors, along with individual anatomical variations, contribute to the wide range of epilaryngeal vibration frequencies. The "velic" subcategory of vibration (5) refers to rare cases of ingressive sounds that set the velum into vibration. In effect, this means that the velum is set to vibrate while sucking air in through the nose, as in some forms of snoring. Finally, for labial and dorsal articulations, a tense stricture can be made that results in a higher frequency of oscillation (150–700 Hz) than the more lax strictures described above. These faster, tenser vibrations (which are not used as speech sounds in any language) were assigned the subcategories "tense bilabial" (6) and "tense dorsal" (7). Intermediate frequencies between lax and tense strictures were not encountered in the data, so the possible transitions from lax to tense strictures appear to be discrete.

(3) For the *turbulence* category, we defined eight subcategories:

- no turbulence (0),
- labial turbulence (1),
- turbulence with a grooved tongue anterior (*s*- and *sh*-like turbulence) (2),
- turbulence with a flat tongue stricture (*th*- and *kh*-like turbulence) (3),
- turbulence with a lateral tongue stricture (lateral turbulence) (4),
- glottal turbulence (h-like sound) (5),
- nasal turbulence (6),
- tissue-modulated turbulence (7).

This classification is based as much on the type of stricture as the place of stricture. The reason is that, in general, the type of stricture determines the character of the sound as much as does place. The "labial" subcategory (1) includes both bilabial and labiodental sounds. The subcategory referring to a "grooved tongue anterior" (2) corresponds to sibilant speech sounds (such as [s] and [ʃ], as in *sun* and *shun*), without detailing their place of articulation. The subcategory referring to a "flat stricture" (3) indicates articulations with a flat tongue constriction (no groove), which corresponds to speech sounds like [$\theta$] and [x] (as in *moth* and *loch*). The "lateral stricture" subcategory (4) refers to constrictions with a lateral (rather than central) outlet, corresponding to lateral fricatives in languages. "Glottal turbulence" (5) can be equated with the speech sound [h] as in *hat*. "Nasal turbulence" (6), which occurs as air exits the nostrils, is equivalent to voiceless nasals in languages. Last, we classified some dorsal articulations as having "tissue-modulated turbulence" (7). These were articulations in which a fluctuating constriction resulted in intermittent and irregular turbulence.

### D. Final extraction

The annotation procedure described above generated a list of segment data pointers into the original database files for each articulation subcategory. The final audio excerpts were extracted using a script in MATLAB, resulting in one audio file for each example. All sounding segments were extracted that had a combination of articulation annotations (in the scheme described above) that was a fit for one of the three main categories. Note that the three categories and their subcategories give rise to a large number of possible combinations in any given sound segment. Each file name was marked with a three-number combination (*SMV-phon-turb*) referring to each subcategory of the three articulatory categories as described above. For example, the combination (2-3-0) would signify an sound with modal voicing and without turbulence.

The analysis of slow SMV in particular demands a certain time window. The duration limit of the included segments was therefore set to 150 ms. This corresponds to three cycles of 20 Hz which is approximately the lower frequency bound for these vibrations. This also made the balance between the number of segments in the positive and negative categories in each category more even.

The extraction resulted in a total of 2689 audio segments of which 1242 were longer than 150 ms and thus kept for the modeling. The final distributions of the three data sets are provided in Table I. There is a reasonably even distribution of the number of segments in the positive and negative groups for phonation and turbulence. There are comparatively fewer SMV cases which result in a larger portion of negative segments in this class. This is a natural consequence of articulatory dependencies in the voice production. The number of segments varies across speakers with comparatively more segments for the French speakers.

## III. AUDITORY RECEPTIVE FIELDS TOOLBOX

The audio examples were analyzed using the ARF Toolbox implemented in MATLAB. The ARF Toolbox implements numerically a new mathematical framework for analyzing sounds with qualitative similarity to neural functions in the auditory pathway (Lindeberg and Friberg, 2015a,b). The model has not been derived primarily from available measurements or data about auditory neural functions. Instead it is an idealized mathematical model in the sense that it is primarily derived from a set of structural assumptions regarding auditory functions, for example, regarding covariance and invariance with respect to translations in time or frequency and glissando transformations. Specifically, idealized receptive fields from this model have been shown to closely approximate neural responses in both the inferior colliculus (ICC) and the primary auditory cortex (A1) in mammals (Lindeberg and Friberg, 2015a,b).

The first stage is to transform the audio signal into a time-frequency representation in terms of a multi-scale spectrogram. Its properties include logarithmic frequency bins, constant bandwidth and time-causal temporal processing. In subsequent stages, additional layers of receptive fields defining local areas in the time-frequency representation can be applied to the first-layer spectrogram. Depending on the shape and size of the receptive fields, different properties can be enhanced, such as onsets, partials, and formants (see Lindeberg and Friberg, 2015a,b). The ARF Toolbox was developed quite recently, and this is the first time it has been applied to a complex practical modeling problem.

TABLE I. The distribution of the number of segments in each of the three articulation categories. Numbers for each participant refer to the number of positive/negative segments.

| | Gender | Nationality | Total | Phonation | SMV | Turbulence |
|---|---|---|---|---|---|---|
| Positive total | | | | 698 | 300 | 705 |
| Negative total | | | | 543 | 941 | 536 |
| Speaker 1 | M | Swedish | 143 | 73/70 | 48/95 | 78/65 |
| Speaker 2 | M | Swedish | 81 | 45/36 | 23/58 | 41/40 |
| Speaker 3 | F | Swedish | 110 | 53/57 | 35/75 | 49/61 |
| Speaker 4 | F | Swedish | 87 | 47/40 | 62/144 | 43/44 |
| Speaker 5 | M | French | 206 | 120/86 | 36/152 | 128/78 |
| Speaker 6 | M | French | 188 | 115/70 | 48/70 | 72/116 |
| Speaker 7 | F | French | 223 | 127/96 | 10/213 | 183/40 |
| Speaker 8 | F | French | 203 | 118/85 | 57/146 | 111/92 |

### A. The first stage: Transformation from audio to spectrogram using the ARF Toolbox

According to the theory in Lindeberg and Friberg (2015a,b), a family of transformation methods can be used to produce a spectrogram from audio. We have chosen here to use the time-causal transformation using a series of first-order integrators (truncated exponentials or recursive filters in the discrete case). A continuous spectrogram $S_h$ is defined from a continuous signal $f(t)$ using a time-causal temporal window kernel $h_{comp}$ in the following way [Eqs. (55), (33), and (31) in Lindeberg and Friberg (2015a)]:

$$S_h(t, \omega; \mu) = \int_{t'=-\infty}^{\infty} h_{comp}(t - t'; \mu) f(t') e^{-i\omega t'} dt', \tag{1}$$

$$h_{comp}(t; \mu) = *_{k=1}^{K} h_{exp}(t; \mu_k), \tag{2}$$

$$h_{exp}(t; \mu_k) = \begin{cases} \frac{1}{\mu_k} e^{-t/\mu_k}, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{3}$$

For the discrete implementation in this study, we approximate the continuous kernel $h_{\text{comp}}$ by the composition of seven recursive filters with each layer of the form

$$f_{out}(t) - f_{out}(t-1) = \frac{1}{1-\mu_k} \left( f_{in}(t) - f_{out}(t-1) \right). \tag{4}$$

The filters are coupled in cascade with the temporal scale levels $\tau_k = c^{2(k-K)} \tau_0$ determined from a logarithmic distribution with distribution parameter $c = \sqrt{2}$ and related to the time constants $\mu_k$ according to [see Lindeberg and Friberg (2015a), section "Computational implementation"]

$$\tau_k = \sum_{i=1}^{k} (\mu_k^2 + \mu_k) \tag{5}$$

from which the time constants can be computed according to Eq. (151) in Lindeberg and Friberg (2015a). The resulting composed temporal window kernel is shown in Fig. 1 [see also Fig. 5 in Lindeberg and Friberg (2015a)]. The width of the kernel in terms of its standard deviation $\sigma_t$ was in the middle frequency range eight cycles of the center frequency of each bin. Thus, in this range, the frequency bandwidth was constant with respect to logarithmic frequency for quasi-stationary signals. In the upper and lower part of the frequency range, the standard deviation of the kernel was gradually flattened out to a constant value. In the lower range, this prevented the kernel from becoming unrealistically wide for low frequencies.

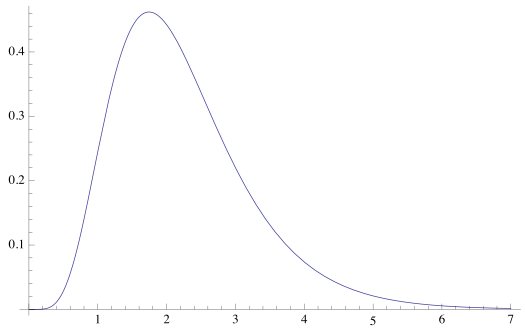


FIG. 1. (Color online) The time-casual kernel $h_{comp}$ used for the spectrogram transformation with $K = 7$ and $c = \sqrt{2}$.

The frequency bins were logarithmically spaced from MIDI note number 36 to 132 (approximately 65 Hz to 16.7 kHz as given by $f = 2^{(MIDI-69)/12} * 440$ Hz) with a resolution of 48 bins per octave. The position in time of each bin was time-compensated using the inflection point of the kernel as the reference.

Finally, the magnitude of the spectrogram $S$ was converted to sound level with a range from 0 to 60 dB normalized to the overall maximum value $S_{max}$,

$$S_{dB} = 20 \log_{10} \left( \frac{|S|}{S_{max}} \right) + 60. \tag{6}$$

In comparison with other time-frequency transforms, the unsymmetrical kernel used here gives a better sound onset response than a discrete Fourier transform using a symmetrical window such as a Hanning window. It has also been shown that this type of kernel, using another set of scale levels with constant $\mu_k$, is the same as the gamma-tone filter that is often used in auditory models (Lindeberg and Friberg, 2015a). Considering the constant bandwidth and logarithmic spacing of the frequencies, it is in this respect close to a constant-Q transform (CQT) (e.g., Brown and Puckette, 1992).

### B. The second stage: Transformations applied on the spectrogram using the ARF Toolbox

A smoothing operation over frequency and time is defined using a 2D spectro-temporal receptive field $T_{2D}$ applied on the sound level spectrogram [Eq. (83) in Lindeberg and Friberg (2015a)]

$$S_{filt}(t, \nu; \tau_f, \Sigma) = \int_{t'=-\infty}^{\infty} \int_{\nu'=-\infty}^{\infty} T_{2D}(t', \nu', \Sigma) \times S_{dB}(t - t', \nu - \nu'; \tau_f) dt' d\nu', \tag{7}$$

where $\Sigma$ is a spectro-temporal covariance matrix of the 2D spectro-temporal smoothing kernel of the form

$$T_{2D}(t, \nu; \Sigma) = g(\nu - vt; s)\, T_{1D}(t; \tau). \tag{8}$$

Here, $g$ is a Gaussian smoothing kernel over logarithmic frequencies $\nu$, $T_{1D}$ is a time-causal temporal smoothing kernel, $s$ is the log-spectral scale, $\tau$ the temporal scale, and $v$ is the glissando parameter that describes how fast logarithmic frequencies vary with time.

From such zero-order receptive fields, derivative based spectro-temporal receptive fields can in turn be defined according to

$$A(t, \nu; \Sigma) = \partial_t^{\alpha} \partial_{\nu}^{\beta} (g(\nu - vt; s)\, T_{1D}(t; \tau)), \tag{9}$$

where $\alpha$ denotes the order of temporal differentiation and $\beta$ the order of log-spectral differentiation. In this study, we did, however, not use the full flexibility of second layer receptive fields (9) and only the basic smoothing operation (7) with different time and frequency parameters.

## IV. FEATURES

A group of features was developed specifically for each of the three main articulation categories. For the phonation and turbulence category, we first developed an enhancement of the spectrum specifically targeted for each category using the ARF Toolbox. Then, we extracted potentially relevant features for each time frame. Finally, these frame features were combined using different statistics to form the final features used in the subsequent prediction. For the SMV category, we also used extraction techniques based directly on the audio waveform. The calculation of all features was implemented in MATLAB.

### A. Phonation features

The aim of the phonation features was to detect any regular harmonic signal considering also the case with a considerable amount of noise present. Therefore, we used a method that specifically enhances the fundamental frequency of a periodic and harmonic signal by adding frequency-translated copies of the spectrogram. This poses very few restrictions on the signal and it can, for example, also be applied to several simultaneous harmonic sources. For computing the phonation features we applied the following steps starting with the spectrogram $S_{dB}$ in Eq. (6).

*a. Removal of silence before and after sound.* Any silence in the beginning and end of the spectrogram was removed in the first stage. The spectrogram $S_{dB}$ was smoothed using a second stage filtering as described in Sec. III B. A discrete Gaussian kernel was used for both the time and frequency dimension with the standard deviation in frequency $\sigma_f = \sqrt{s} =$ 3 semitones, corresponding roughly to critical bands, and with a standard deviation in time $\sigma_t = \sqrt{\tau} = 0.01$ s.

A detection function $d_i$ was calculated for each frame $i$ in the smoothed spectrogram $S_f$ by taking the maximum sound level in each frame. A fixed threshold was defined at $-25$ dB below the maximum sound level for the whole example,

$$d_i = \max_j S_{f,i,j} - (\max_{i,j} S_f - 25). \tag{10}$$

The beginning and end of the initial spectrogram, in which $d_j$ was below zero, were removed.

*b. Whitening in spectral dimension using an ARF Gaussian filter.* A 1D receptive field corresponding to a smoothing filter across the frequency dimension was applied on the cropped spectrogram from the previous calculation, using a discrete Gaussian kernel with the standard deviation of $\sigma_f = 8$ semitones, resulting in the spectrogram $S_f$. For each time frame $i$ and frequency bin $j$, the filtered spectrum $S_{f,i,j}$ was subtracted from the cropped spectrogram $S_{dB,i,j}$ using an offset of 3 dB and a maximum range of 50 dB from the local maximum in the frame,

$$S_{w,i,j} = \begin{cases} S_{dB,i,j} - (S_{f,i,j} + 3), & S_{dB,i,j} > \max_j S_{dB,i,j} - 50, \\ 0, & S_{dB,i,j} \leq \max_j S_{dB,i,j} - 50. \end{cases} \tag{11}$$

An example of the resulting spectrogram $S_w$ is shown in Fig. 2, middle.

*c. Enhancement of harmonic fundamental frequency.* We assumed that the spectrum of the phonation part of the sound was perfectly harmonic and the remaining part of the audio consisted of some kind of noise. The enhancement of the harmonic fundamental was done by adding translated spectrogram copies according to the harmonic series. For example, the spectrum translated one octave down was added to the original spectrum. In this way, the fundamental was enhanced, adding the magnitude of the first partial. This will also add an extra partial one octave below. However, the impact of these extra partials will be small since they will not appear at the same frequency, and thus will not be additively enhanced. Accordingly, the resulting spectrogram $S_h$ was computed as a sum of $k$ translated copies of the whitened spectrum $S_w$,

$$S_{h,i,j} = \sum_k c_k S_{w,i,(j+l_k)}, \tag{12}$$

where $c_k$ are scalar constants and $l_k$ defines each translation in frequency according to the harmonic series. For the frequency translations that did not match the frequency sampling points, the spectrum was interpolated. One example of such a final harmonic enhancement is shown in Fig. 2 (bottom). As seen in the figure, the fundamental (the most red parts) is enhanced while the remaining partials are suppressed.

*d. Frame feature extraction.* The following six frame-based features were calculated from each time frame of the enhanced fundamental: Sound level (hf0_maxsl) and frequency (hf0_maxf0) of highest peak, sound level (hf0_max2sl) and frequency (hf0_max2f0) of the second highest peak, difference in sound level between the two peaks (hf0_maxsldiff), and mean sound level for all peaks except the highest one (hf0_meanrestsl).

*e. Final features across frames.* Statistical properties were calculated across the time-frames for all frame-based features in the preceding step. The statistics used for these final features were the *upper quartile* (_uqt), the *standard deviation* (_std), and the *mean of the absolute difference between frames* (mean absolute derivative in time) (_mva). They were calculated across the whole sound example. This resulted in a total of 18 $(6 * 3)$ features for the phonation

category. Note that the mean absolute derivative is also used as a feature in the image processing method surf (Bay *et al.*, 2008).

This method resembles the CPP previously used for estimation of breathiness and other voice characteristics (Hillenbrand *et al.*, 1994; Hillenbrand and Houde, 1996). The CPP estimates the amount of periodicity in the voice signal by measuring the peak prominence in the cepstrum. We use instead the summation of spectrograms to enhance the fundamental frequency. Then similar to the CPP, the periodicity (or the relative amount of an harmonic signal) will be high if there are prominent peaks in the resulting spectrum (e.g., as the bottom graph in Fig. 2). Also, instead of selecting one final measure, we chose several different possible features from the resulting spectrogram. The idea was to let the importance of these varying aspects be discovered by the machine learning algorithms in the subsequent processing.

### B. SMV features

The character of the supraglottal myoelastic vibrations and their relation to the resulting audio signal is largely unknown. Therefore, we used several different approaches for the feature extraction, using the time-domain signal, the instant sound level defined below, or the spectrogram as input. One challenge was that the signal could contain also phonation and turbulence at the same time. As specified in Sec. II C, the frequency of SMV vibrations sometimes overlaps with the frequency ranges of normal phonation. Our approach was to focus on the amplitude modulations resulting from SMV for the different frequency ranges as specified in the subcategories. That amplitude modulation is a strong effect in SMV seems intuitive given that myoelastic oscillations of the lips results in an amplitude modulation of other sounds produced by normal phonation in other parts of the vocal tract.

#### *1. Modulation filter bank*

As mentioned above, SMV sounds can be viewed as comparatively slow amplitude modulations of the turbulence and/or phonation. Therefore, we developed a modulation filter bank for specifically detecting amplitude modulation in the lower frequency range between 18 and 1000 Hz. The design can be viewed as a simplification of the auditory modulation filter bank proposed by Dau *et al.* (1997). The following steps describe the procedure.

*a. Instant sound level (ISL).* The ISL was computed from the RMS of the signal using a Hann window of 1 ms and a hop size of 0.1 ms. A subsequent high-pass filter with a cutoff at 15 Hz removed the DC component. As a result, the ISL had a sampling frequency of 10 kHz and a frequency range of 15 Hz–1 kHz.

*b. Spectrum.* An average spectrum was computed by averaging over a series of short-time Fourier transforms with a Hamming window of 1024 samples and an overlap of 512 samples.

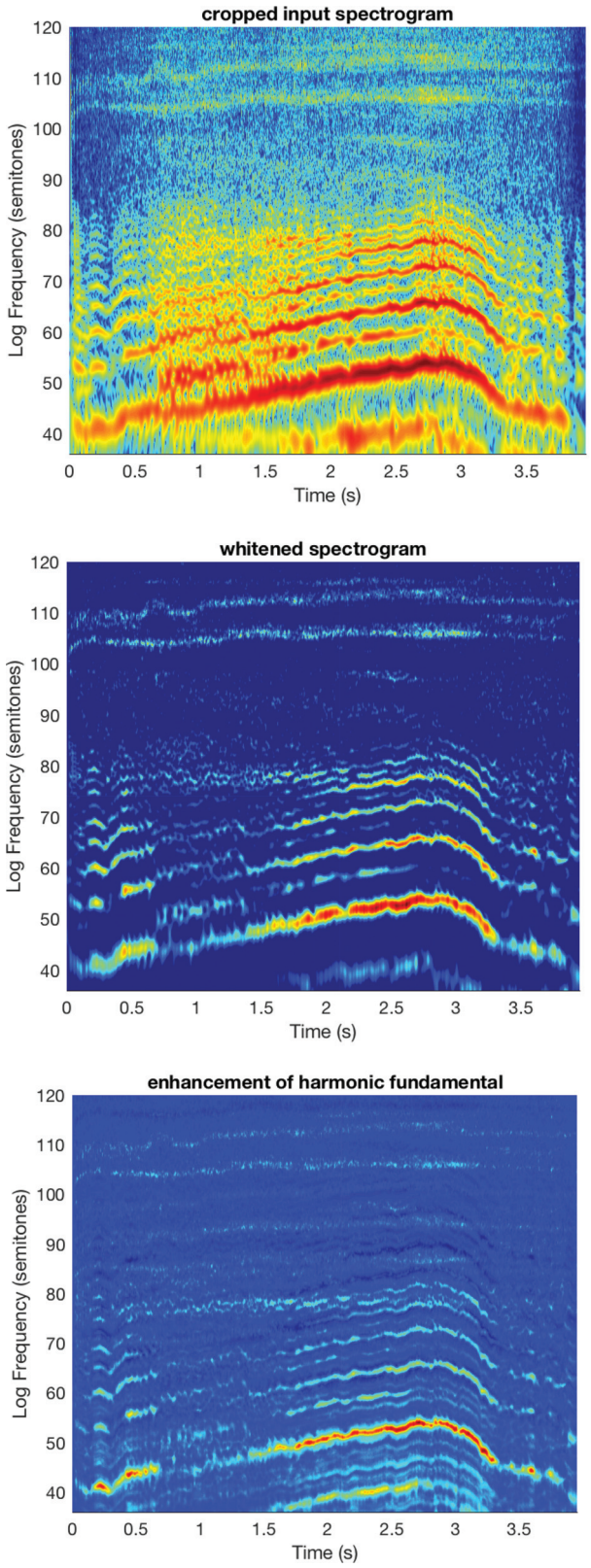


FIG. 2. The original spectrogram (top), the whitened spectrogram (middle), and the resulting enhancement of harmonic fundamental (bottom). The sound example is an imitation of an accelerating lorry containing both phonation and turbulence.

*c. Filter bank.* Six filter channels were defined according to the expected frequencies for the different type of vibrations as specified in Sec. II C. They were approximately logarithmically distributed with band 1, 18–35 Hz; band 2,

35–60 Hz; band 3, 60–110 Hz; band 4, 110–220 Hz; band 5, 220–500 Hz; band 6, 500–1000 Hz.

*d. Final features.* The maximum sound level in each band was used as the final six features (vibspecdb1–6).

The final signal processing design (e.g., computing the ISL) was determined from testing various solutions on a small selection of examples. The efficiency is particularly evident in the lower frequency range. An example is given in Fig. 3, which illustrates how the supraglottal myoelastic vibrations emerge in the averaged spectrum (calculated in step 2 above). Notice also that the vibrations can be observed in the upper/middle part of the spectrogram. A drawback is that the frequency of the vibrations needs to be rather stable during the whole example.

### 2. SMV features using vibrato extraction methods on the sound level waveform

The ISL described in Sec. IV B 1 was used as input. The ISL curve was filtered in three different frequency ranges. A method developed for vibrato extraction (Friberg *et al.*, 2007) was used for the extraction of the dominant AM frequencies in each frequency band. One advantage with this method is that it allows accurate detection of the frequency even if it varies on a cycle-to-cycle basis. The following steps were applied.

*a. Filtering.* The ISL signal was low pass and high pass filtered into three different bands corresponding to expected frequencies for these types of vibrations. Band 1, 15–40 Hz; band 2, 40–110 Hz; band 3, 100–250 Hz.

*b. Detect regular amplitude variations.* In each band, cyclic variations of the SL curve were detected using the three-point method suggested by Prame (1994) and implemented by Erwin Schoonderwaldt (Friberg *et al.*, 2007). Local peaks and troughs of the band filtered SL curve were detected with a simple peak-picking method. For each half-cycle $n$ (peak-trough-peak or trough-peak-trough) rate $R_n$ and extent $E_n$ were calculated using a three-point estimation involving the two adjacent peaks/troughs,

$$R_n = \frac{1}{t_{n+1} - t_{n-1}}, \tag{13}$$

$$E_n = \frac{1}{4}|A_{n+1} - 2A_n + A_{n-1}|, \tag{14}$$

where $t_n$ indicates the time instance of the peak/trough and $A$ is the filtered ISL. An example of the detected peaks and troughs is shown in Fig. 4.

*c. Final features.* There were four features calculated from the detected points for each of the three bands. The first two features consisted of the median rate $R$ and extent $E$ over all detected points (vibrate3–5 and vibext3–5). The third feature $LR$ was a calculation of the total length of detected variations, relative to the total length of the example (vibprop3–5). It reflects the relative duration of amplitude vibration in the example. The fourth feature was a combination of two features via multiplication, $LR * E$, reflecting the interaction between them (vibextprop3–5). In addition, two combinations of the extent values were computed using the maximum across band 1 and 2 (vibcomb34) and the maximum across band 1, 2, and 3 (vibcomb345). These last features reflect a coarser division of the frequencies into two regions. This resulted in a total of 4 features $*$ 3 bands $+$ 2 combination-features $=$ 14 features.

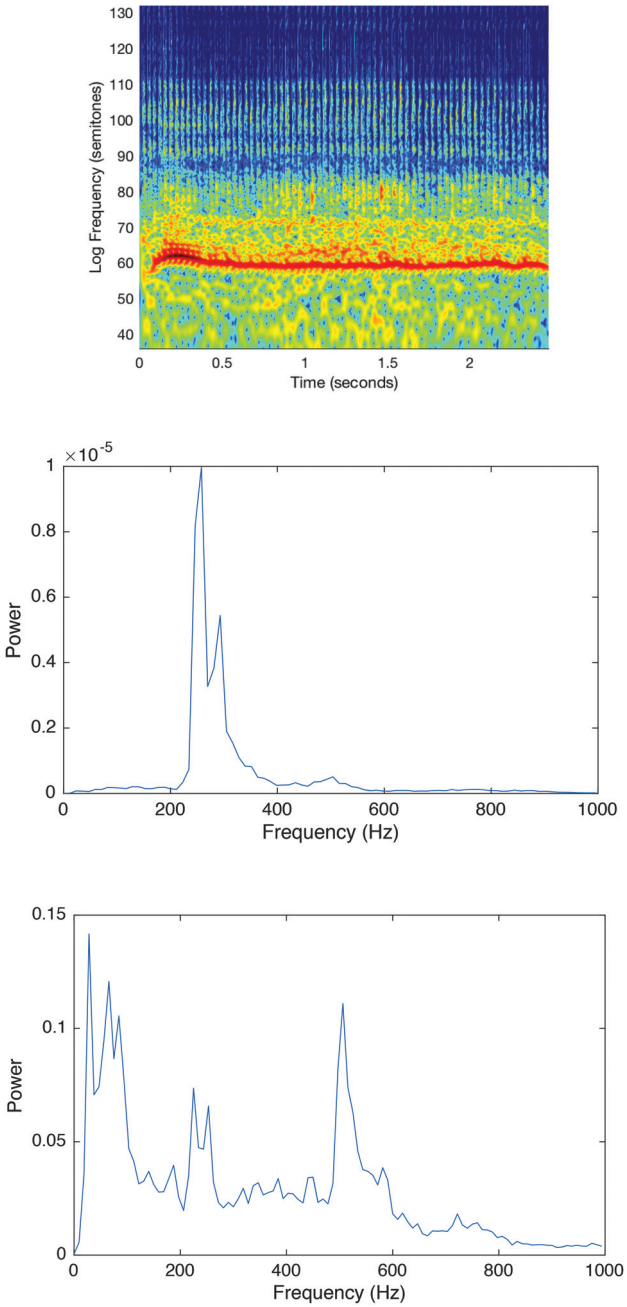


FIG. 3. Spectrogram (top), power spectrum of original audio signal (middle), and power spectrum of the ISL (bottom). The sound example is a vocal imitation of a lawn mower and contains both slow SMV and phonation. The averaged power spectrum of the original audio signal (middle) is dominated by the phonation frequency around 250 Hz (red line in spectrogram). The power spectrum of the ISL (bottom), shows the emerging low frequencies corresponding to the SMV, the comparatively lower amplitude for the fundamental frequency of the phonation, and the emergence of the second partial at around 500 Hz.

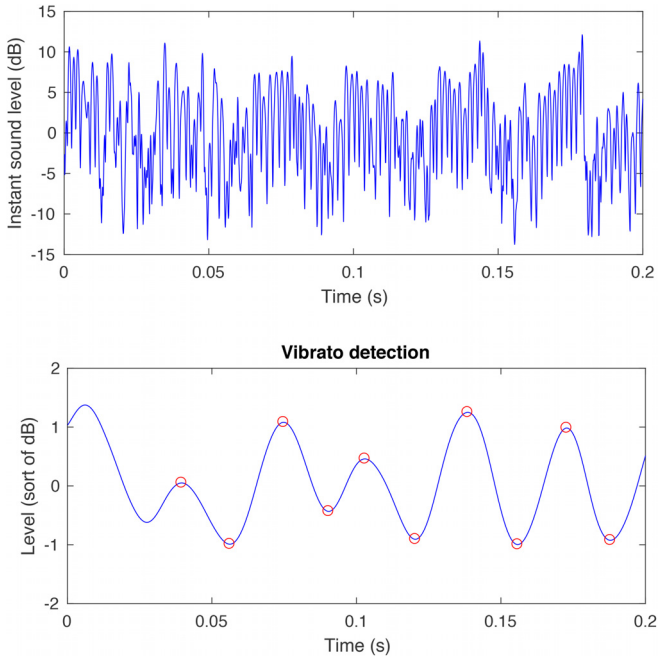


FIG. 4. (Color online) The ISL (top) and the detected peaks and troughs in the filtered waveform (bottom) for the lowest frequency band. The example is the first part of the one shown in Fig. 3.

This method is rather sensitive to a signal containing a mixture of frequencies that potentially can introduce errors in the resulting calculations. This is partly compensated for by using the median across several estimations. However, it is also potentially less sensitive to frequency variations within each band, since the calculations of the extent and rate are based on just a half-cycle of the modulation frequency.

### 3. SMV features using vibrato extraction methods on the spectrogram

It was evident that the supraglottal myoelastic vibrations were visible as cyclic variations in the upper part of the spectrogram (see top graph in Fig. 3). Therefore, we also implemented the vibrato extraction method using the spectrogram $S_{dB}$ in Eq. (6) as input. However, due to the resolution of the resulting signals fewer frequency bands were analyzed in this case. The following steps were applied.

*a. Extract sound level curve for upper spectrum.* The sound level curve of all frequencies above 1 kHz was calculated, using the median sound level across all frequency bins for each time frame in the spectrogram.

*b. Detect regular amplitude variations.* From the sound level curve, cyclic variations were detected using the same three-point method previously used for vibrato detection, as described in Sec. IV B 3 a.

In this case, we used a coarser division of just two bands addressing both the relatively slow vibrations by the tongue and lips (around 30 Hz) and the somewhat faster vibrations produced by various inner parts of the throat and tongue (around 70 Hz, but with a rather large span). For the slow variations, a low pass filter with a cutoff of 50 Hz was first applied and then the detection was made with a range of 15 to 40 Hz. For the fast variations the cutoff frequency was 220 Hz and the detection range was 40 to 200 Hz. This resulted in two sets of discrete detection points marking both each detected peak and trough for the two frequency bands. The rate $R$ and extent $E$ were calculated from Eqs. (13) and (14).

*c. Final features.* There were four features calculated from the detected points for each analysis. The first two features were extracted by computing the median across all detected points for the rate $R$ and extent $E$ (vibrate1–2 and vibext1–2). The third feature $LR$ was a calculation of the total length of detected variations relative to the total length of the example (vibprop1–2). The fourth feature was a combination of two features via multiplication, $LR * E$, reflecting the interaction between them (vibextprop1–2). This resulted in a total of eight SMV features.

### 4. SMV features using a difference function on the waveform

As the periodicity of sounds can be detected by correlating an audio file with itself in the time domain (autocorrelation), this strategy was also used for detecting the myoelastic vibrations. It was accomplished by implementing the average squared difference function (ASDF) as used in the YIN pitch detection algorithm [Cheveigné and Kawahara, 2002; see also Rao (2011)].

*a. Difference function.* The difference function $d_t$ is defined as the cumulative sum of the squared difference between each value $x(n)$ and the value at $x(n+\tau)$, where $n = 0,\ldots,\mathrm{N}-1$ is the index of a signal with N samples and $\tau$ the offset ranging from $\tau_{min}$ to $\tau_{max}$,

$$d_t(\tau) = \sum_{n=1}^{N-1} (x(n) - x(n+\tau))^2. \tag{15}$$

The input values in terms of $\tau_{min}$ to $\tau_{max}$, hop and block size $N$ were specified manually, and were tested with different values to ensure the best balance between speed and accuracy. The minimum frequency was chosen to be 20 Hz and the maximum frequency was set to 200 Hz. Hop and block size $N$ were set to 10 and 20 ms, respectively. Note that while the former parameter speeds up the calculation when it is increased, the latter speeds it up when it is decreased.

*b. Normalization and scaling.* The difference function $d$ was further scaled with the cumulative mean difference function (Cheveigné and Kawahara, 2002). The resulting $d'$ reduces the sensitivity to strong formants and reduces the sensitivity for peaks at multiple frequencies (Rao, 2011),

$$d'_t(\tau) = \begin{cases} 1 & \text{if } \tau = 0, \\ \dfrac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)} & \text{otherwise.} \end{cases} \tag{16}$$

*c. Frequency detection.* After having computed the difference matrix for a specific fragment, the lowest and second lowest minima of the specified frequencies were identified separately for each frame. The number of frames depends on the input parameters mentioned above. Figure 5 shows an excerpt from the output of the difference function $d'$ and the detected minima.

*d. Final feature extraction.* Several features were derived from the frequency minima values. For both the lowest minimum and second lowest minimum in each frame, the median, the mean of the absolute value of variation (derivative in time), and the standard deviation were computed (cor_med1–2, cor_mva1–2, cor_std1–2). Additionally, two global values were computed; the absolute value of the difference between the two median values (cor_diff), and the weighted amplitude of the lowest minima using a Gaussian curve, centered on the base frequency equal to 30 Hz (the most likely slow SMV frequency) (cor_gaus). This resulted in a total of eight features.

### D. Turbulence features

Since turbulence generates noise, the main idea was to estimate the noise part of the spectrum by removing the harmonic partials, i.e. the spectral peaks. The spectrogram $S_{dB}$ in Eq. (6) was used as input. The magnitudes of different frequency bands of the remaining noise spectrum were used for the final features. For the turbulence category we applied the following steps.

*a. Smoothing in time.* The spectrogram was smoothed in time using a second stage receptive field applied to the spectrogram. A discrete Gaussian kernel was used with a standard deviation of 30 ms.

*b. Estimation of noise spectrum using a smoothing filter.* A smoothing filter was specifically designed to remove the harmonic partials, taking into account the bandwidth of the spectrogram and the variation of the partial density across the spectrum. The filter has similarities to earlier median filtering and order statistics filtering methods, used for separating harmonic and percussive content (FitzGerald, 2010; Elowsson and Friberg, 2015). Instead of using the median for filtering, the 15th percentile was used. The filter window varied as a function of frequency according to a linearly interpolated break-point, using points at the logarithmic frequencies 15, 50, 96, 140 semitones (MIDI), and the corresponding logarithmic window size at 2, 15, 4, 3 semitones. The obtained spectral shape was further smoothed in the frequency domain, using a similar frequency-dependent filter with a Gaussian kernel. The percentiles, the window sizes and the break-points were chosen manually, trying to minimize the harmonic content while retaining most of the turbulence (noise).

*c. Estimation of spectral peaks using a smoothing filter.* The spectral shape following the peaks in the spectrum was estimated using the same filter and smoothing as in step 2 above. The only difference was that the filter used the 95th percentile instead.

An illustration of the resulting estimations of noise and spectral peaks at a specific time is shown in Fig. 6, and for the whole spectrogram in Fig. 7.

*d. Frame feature extraction (spectral bands).* The remaining spectral shape of the noise obtained in step 2 was divided both into seven octave bands (with boundaries at 36, 48, 60 72, 84, 96, 108, 120 semitones in MIDI units) and into two bands (above or below 1 kHz). For each band (and time frame) the median value across frequency was calculated (nosB2_1–2 and nosB7_1–7).

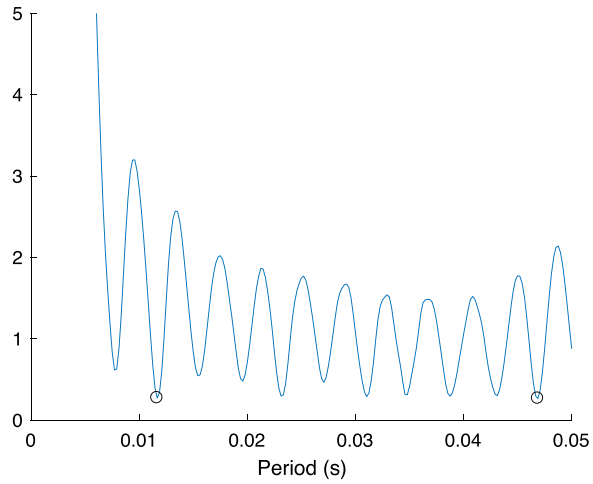


FIG. 5. (Color online) Example of the resulting $d'$ as a function of period length $\tau$ for one frame in the same example as in Fig. 3. The circles indicate the detected lowest and second lowest minima, respectively.

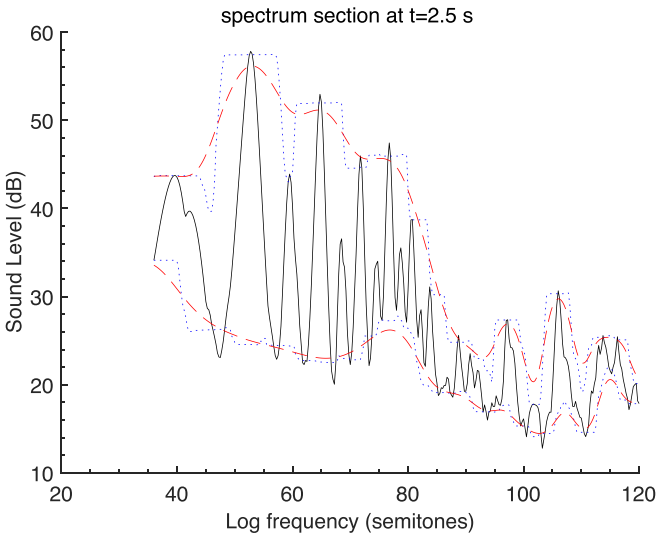


FIG. 6. (Color online) The smoothing filters with variable windows applied on a spectrum section at t = 2.5 s for the same sound example as in Fig. 2. The black line indicates the original spectrum, the dotted lines (blue in color print) are the percentile filtered spectra, and the dashed lines (red in color print) are the resulting spectra after Gaussian smoothing for the upper and lower estimation.

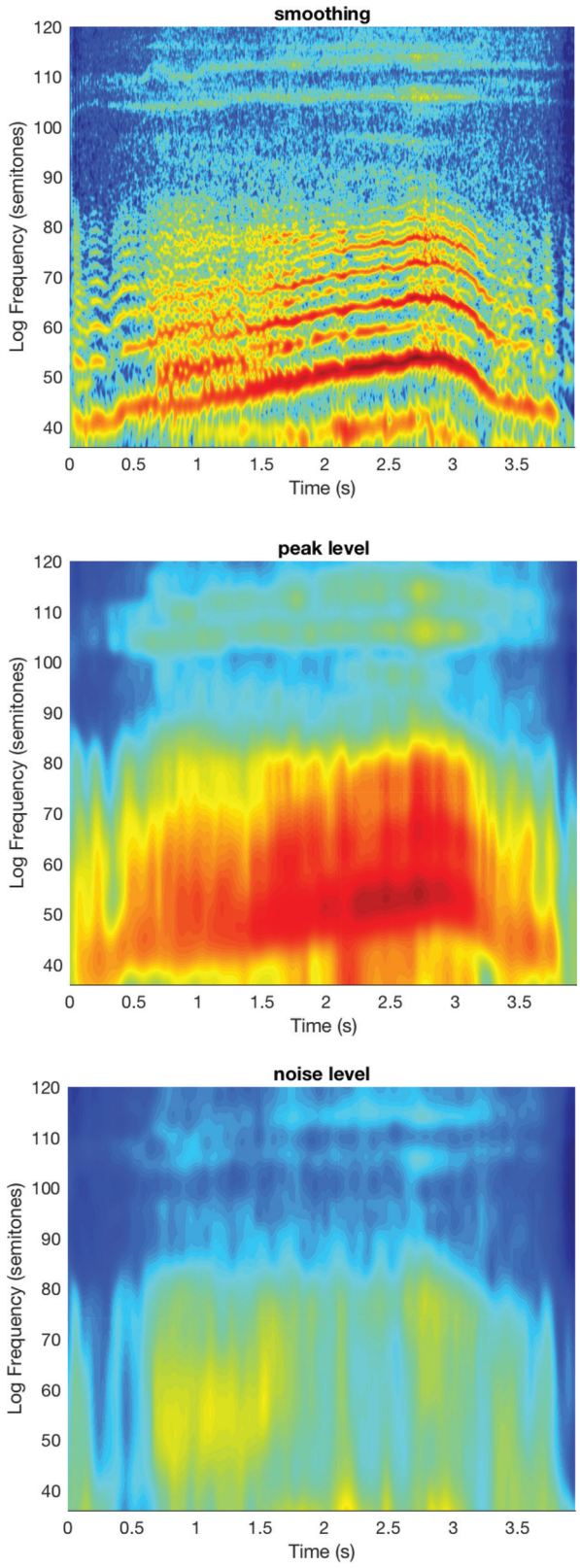


FIG. 7. An example of the resulting spectral shapes in the turbulence feature extraction. The smoothed spectrogram (top), the spectral shape of the peaks (middle), and the spectral shape of the noise after removal of narrow spectral peaks (bottom). Same sound example as in Fig. 2.

*e. Harmonics-to-noise ratio.* Using the final upper and lower estimation of the spectra obtained in step b, a measure of harmonics-to-noise was defined as the maximum distance between the two curves (nosH2N).

*f. Final features across frames.* The final features were calculated in the same way as for the phonation features. Thus, using the upper quartile (_uqt), standard deviation (_std), and the mean of the difference between frames (mean derivative in time) (_mva). This resulted in a total number of 30 (7 bands $*$ 3 stats $+$ 2 bands $*$ 3 stats $+$ 1 ratio $*$ 3 stats) features for the turbulence category.

## V. PREDICTION CATEGORIES AND METHODS

As discussed above, the current prediction focused on the three separate articulatory categories *phonation*, *turbulence*, and *SMV*. The prediction of these categories was a new challenge for which we found no prior examples in the literature. Each category can be active in a rather independent way. For example, SMV using the tongue and lips can be combined with both phonation and turbulence. Therefore, for each of the three categories, we made an independent classification/regression model. The ground truth was coded as 1 for the positive segments and 0 for the negative ones. Note that in this study we did not attempt to model the subcategories. See Table I for the distribution across categories and participants.

In the selection of the prediction methods we wanted to use both simple and more advanced models. Due to the relatively large number of features (84) in relation to the total number of cases (1242), we first applied partial least-square (PLS) regression. PLS regression attempts to minimize the number of independent features by a principal component analysis in combination with a linear regression (Geladi and Kowalski, 1986). The method can be used as an alternative to traditional linear regression when there are a large number of features. We used the PLS package in MATLAB for this computation. For the classification of the positive and negative category, the regression data were simply categorized as true (1) for values higher than 0.5 and otherwise false (0). The number of factors in the PLS regression was selected manually by choosing the minimum number that could still explain a major part of the cross-validated variation.

As a second method we applied support vector machine (SVM) classification (e.g., Smola, and Schölkopf, 2004). using the LIBSVM version 3.22 package for MATLAB (Chang and Lin, 2011). A radial basis function was used as the kernel and the parameters were set at their default values.

As a third method, we applied an ensemble of multilayer perceptrons (EMLP). This method was recently used to predict performed dynamics in a study that also had many features in relation to the number of cases (Elowsson and Friberg, 2017). Each network of multilayer perceptrons (MLPs) had the same topology and used the same input data. Since each MLP was randomly initialized, they will converge at different local minima. The ensemble of these networks will therefore act as a regularization technique that enhances generalization capabilities (Hansen and Salamon, 1990). In other words, when using the average prediction of these models, we can expect a better outcome than if we were to randomly choose one of them (Polikar, 2006). After initial testing, the following setup was chosen for each neural network (NN) of the ensemble:

- Each network had three hidden layers, with 15 neurons in each layer. This resulted in an architecture (including

input and output layer) of {84, 15, 15, 15, 1}. Each network was thus rather deep, although the number of neurons was still kept small.

- The non-linearities in the first two hidden layers were hyperbolic tangent (tanh) units, and the non-linearities for the last hidden layer were rectified linear units. The idea of a mixture of non-linearities within an ensemble of MLPs was previously used by Elowsson (2016). The output layer had a sigmoid activation function.
- The networks were trained with scaled conjugate gradient back propagation.
- Each network was trained for a maximum of 240 epochs (240 complete cycles with all training examples). Training was, however, set to stop if the gradient reached below $10^{-6}$.
- Each input feature was normalized within the range $\pm$ 1.

Two different cross-validation methods were used. The first one was the traditional tenfold method with 20 random permutations. The second was "leave-one-participant-out." Since there were a total of eight participants in this study, the training was performed on seven participants and the testing on the remaining one; and this was repeated for all participants. This would correspond more closely to a real-world case when a possible project prototype system is operated by a new user. However, due to the small number of participants, it is sensitive to individual variations, and therefore less reliable as an estimate.

## VI. RESULTS

### A. Correlations with ground truth

As a first test, the point-biserial correlation coefficients were computed between each feature and the ground truth. Table II displays the correlations for the phonation, SMV, and turbulence features, respectively. Due to the multiple testing, the significance values should be interpreted with some caution and should be viewed only as an overall indication of the correspondence. As seen in Table II, almost all features are correlated to some extent with at least one of the three ground truth measures. This makes it problematic to exclude any feature on the basis of the correlations. Note that even if one feature has a low correlation to the intended category, in most of the cases this feature has a higher correlation to the other categories. For example, the phonation feature hf0_maxsl_mva (see Table II) has a correlation of 0.08 to phonation ground truth but a correlation of 0.44 to turbulence ground truth. Thus, this feature can be potentially useful in the prediction of turbulence when all features are used.

The highest correlations are found for the phonation category in Table II with correlations up to $r_{pb} = 0.8$ indicating that these features are to a certain part capturing some of the

TABLE II. Correlations between phonation features and ground truth for the three articulation categories.

| Feature | Phonation | SMV | Turbulence |
|---|---|---|---|
| hf0_maxsl_uqt | 0.80[a] | −0.09[b] | −0.49[a] |
| hf0_maxsl_std | 0.74[a] | −0.05 | −0.17[a] |
| hf0_maxsl_mva | 0.08[b] | 0.09[b] | 0.44[a] |
| hf0_maxf0_uqt | −0.74[a] | −0.23[a] | 0.37[a] |
| hf0_maxf0_std | −0.73[a] | −0.01 | 0.41[a] |
| hf0_maxf0_mva | −0.76[a] | −0.08[b] | −0.55[a] |
| hf0_max2sl_uqt | 0.77[a] | −0.05 | −0.51[a] |
| hf0_max2sl_std | 0.66[a] | 0.01 | 0.06[c] |
| hf0_max2sl_mva | −0.36[a] | 0.19[a] | 0.27[a] |
| hf0_max2f0_uqt | −0.68[a] | −0.12[a] | 0.24[a] |
| hf0_max2f0_std | −0.56[a] | 0.06[c] | 0.44[a] |
| hf0_max2f0_mva | −0.81[a] | −0.04 | −0.53[a] |
| hf0_meanrestsl_uqt | 0.63[a] | −0.05 | −0.50[a] |
| hf0_meanrestsl_std | 0.61[a] | −0.03 | −0.29[a] |
| hf0_meanrestsl_mva | 0.48[a] | 0.21[a] | −0.47[a] |
| hf0_maxsldiff_uqt | 0.78[a] | −0.10[a] | −0.48[a] |
| hf0_maxsldiff_std | 0.75[a] | −0.01 | 0.10[a] |
| hf0_maxsldiff_mva | −0.25[a] | 0.16[a] | −0.51[a] |

[a]$p < 0.001$ significance level.
[b]$p < 0.01$ significance level.
[c]$p < 0.05$ significance level.

TABLE III. Correlations between SMV features and ground truth for the three articulation categories.

| Feature group | Variable | phonation | SMV | Turbulence |
|---|---|---|---|---|
| Modulation filter bank | vibspecdb1 | −0.52[a] | 0.21[a] | 0.21[a] |
| | vibspecdb2 | −0.52[a] | 0.37[a] | 0.25[a] |
| | vibspecdb3 | −0.44[a] | 0.39[a] | 0.20[a] |
| | vibspecdb4 | −0.20[a] | 0.27[a] | 0.13[a] |
| | vibspecdb5 | 0.42[a] | 0.20[a] | −0.23[a] |
| | vibspecdb6 | 0.65[a] | 0.10[a] | −0.42[a] |
| Vibrato extraction from spectrogram | vibrate1 | −0.28[a] | 0.21[a] | 0.12[a] |
| | vibext1 | −0.28[a] | 0.25[a] | 0.05 |
| | vibprop1 | −0.26[a] | 0.26[a] | 0.03 |
| | vibextprop1 | −0.22[a] | 0.28[a] | −0.03 |
| | vibrate2 | −0.26[a] | 0.17[a] | 0.19[a] |
| | vibext2 | −0.28[a] | 0.23[a] | 0.11[a] |
| | vibprop2 | −0.24[a] | 0.22[a] | 0.06[b] |
| | vibextprop2 | −0.22[a] | 0.10[a] | 0.20[a] |
| Vibrato extraction from ISL | vibrate3 | −0.39[a] | 0.23[a] | 0.17[a] |
| | vibext3 | −0.32[a] | 0.11[a] | 0.12[a] |
| | vibprop3 | −0.44[a] | 0.15[a] | 0.20[a] |
| | vibextprop3 | −0.34[a] | 0.12[a] | 0.13[a] |
| | vibrate4 | −0.43[a] | 0.24[a] | 0.18[a] |
| | vibext4 | −0.33[a] | 0.32[a] | 0.08[c] |
| | vibprop4 | −0.44[a] | 0.32[a] | 0.17[a] |
| | vibextprop4 | −0.33[a] | 0.37[a] | 0.06[b] |
| | vibrate5 | −0.22[a] | 0.11[a] | 0.23[a] |
| | vibext5 | 0.01 | 0.24[a] | −0.02 |
| | vibprop5 | −0.29[a] | 0.23[a] | 0.11[a] |
| | vibextprop5 | −0.03 | 0.27[a] | −0.07[b] |
| | vibcomb34 | −0.29[a] | 0.27[a] | 0.05 |
| | vibcomb345 | −0.04 | 0.21[a] | −0.03 |
| ASDF | cor_med1 | 0.60[a] | −0.16[a] | −0.42[a] |
| | cor_mva1 | 0.58[a] | −0.09[a] | −0.42[a] |
| | cor_std1 | 0.69[a] | −0.11[a] | −0.46[a] |
| | cor_med2 | 0.53[a] | −0.17[a] | −0.37[a] |
| | cor_mva2 | 0.56[a] | −0.15[a] | −0.40[a] |
| | cor_std2 | 0.65[a] | −0.15[a] | −0.45[a] |
| | cor_diff | 0.46[a] | −0.18[a] | −0.32[a] |
| | cor_gaus | −0.63[a] | 0.13[a] | 0.42[a] |

[a]$p < 0.001$ significance level.
[b]$p < 0.05$ significance level.
[c]$p < 0.01$ significance level.

TABLE IV. Correlations between turbulence features and ground truth for the three articulation categories. nosB2_1.–nosB2_2. indicate the two frequency band division, nosB7_1.–nosB7_7. indicate the seven octave bands, and nosH2N. is the harmonics-to-noise measure.

| Feature | Phonation | SMV | turbulence |
|---|---|---|---|
| nosB2_1_uqt | −0.51[a] | 0.28[a] | 0.27[a] |
| nosB2_1_std | −0.11[a] | −0.08[b] | 0.07[c] |
| nosB2_1_mva | −0.23[a] | 0.21[a] | 0.22[a] |
| nosB2_2_uqt | −0.75[a] | −0.02 | 0.50[a] |
| nosB2_2_std | −0.49[a] | 0.03 | 0.23[a] |
| nosB2_2_mva | −0.29[a] | 0.30[a] | 0.07[c] |
| nosB7_1_uqt | −0.52[a] | 0.24[a] | 0.27[a] |
| nosB7_1_std | −0.18[a] | −0.04 | 0.15[a] |
| nosB7_1_mva | −0.34[a] | 0.28[a] | 0.30[a] |
| nosB7_2_uqt | −0.44[a] | 0.29[a] | 0.21[a] |
| nosB7_2_std | −0.10[a] | −0.08[b] | 0.10[a] |
| nosB7_2_mva | −0.26[a] | 0.19[a] | 0.22[a] |
| nosB7_3_uqt | −0.49[a] | 0.30[a] | 0.26[a] |
| nosB7_3_std | −0.10[a] | −0.08[b] | 0.07[c] |
| nosB7_3_mva | −0.18[a] | 0.14[a] | 0.19[a] |
| nosB7_4_uqt | −0.61[a] | 0.21[a] | 0.38[a] |
| nosB7_4_std | −0.30[a] | 0.03 | 0.15[a] |
| nosB7_4_mva | −0.29[a] | 0.26[a] | 0.23[a] |
| nosB7_5_uqt | −0.66[a] | 0.11[a] | 0.41[a] |
| nosB7_5_std | −0.43[a] | 0.05 | 0.20[a] |
| nosB7_5_mva | −0.30[a] | 0.26[a] | 0.17[a] |
| nosB7_6_uqt | −0.69[a] | −0.05 | 0.42[a] |
| nosB7_6_std | −0.44[a] | 0.03 | 0.17[a] |
| nosB7_6_mva | −0.30[a] | 0.27[a] | 0.07[b] |
| nosB7_7_uqt | −0.76[a] | −0.16[a] | 0.58[a] |
| nosB7_7_std | −0.64[a] | −0.06[c] | 0.41[a] |
| nosB7_7_mva | −0.49[a] | 0.16[a] | 0.36[a] |
| nosH2N_uqt | 0.81[a] | −0.10[a] | −0.48[a] |
| nosH2N_std | 0.68[a] | −0.10[a] | −0.42[a] |
| nosH2N_mva | −0.82[a] | −0.11[a] | 0.45[a] |

[a]$p < 0.001$ significance level.
[b]$p < 0.01$ significance level.
[c]$p < 0.05$ significance level.

unique properties of the phonation category. Note also that the upper quartiles (…_uqt) obtain the highest correlations comparing the three statistical measures for each feature in most cases, thus, corresponding to the intended function of each feature.

In Table III we see that the different feature groups correlate reasonably with the SMV ground truth, although the maximum correlations in this case reach only about $r = 0.39$ (vibspecdb3). The filter bank features (vibspecdb1–6) and the vibrato extraction methods on the SL (vib…3–5) correlate positively for lower bands with SMV, and negatively with phonation, as expected. Here the discrimination of SMV from the other categories is less clear, since in many the cases the correlations are higher both for phonation and for turbulence than for SMV. For the autocorrelation features extracted from the time signal (cor_…), this could be expected, since the autocorrelation method YIN was originally developed for pitch detection; and since the autocorrelation was computed for the original waveform and not for the ISL.

For the turbulence features shown in Table IV, the correlations to the turbulence ground truth (rightmost column) vary considerably and reach $r = 0.58$ for the highest frequency band (nosB7_7_uqt). As for the SMV features, the turbulence features often correlate strongly but negatively with phonation. In this case, the harmonics-to-noise metric nosH2N_uqt correlates strongly with phonation, indicating that it is capturing the intended information.

TABLE V. Final prediction results for the phonation category.

| Cross-validation | PLS components | PLS accuracy (%) | SVM accuracy (%) | EMLP accuracy (%) |
|---|---|---|---|---|
| 10-fold | 6 | 96.4 | 96.6 | **96.9** |
| Leave-one-out | 5 | **96.1** | 95.9 | 95.9 |

## B. Classification of phonation

### 1. Overall results using all features

The results of the classification of the phonation category for the different cross-validations and methods are summarized in Table V. As shown in the table, all methods gave an overall classification accuracy above 95% for both cross-validations. The best PLS results were obtained with a modest number of PLS components (5–6). The differences between methods were quite small, indicating that the features were well able to capture the relevant acoustic properties for phonation versus non-phonation.

Figure 8 shows the results from the PLS regression, applied using six components and without cross-validation. Thus, this is the prediction output before the classification is performed. As seen in the figure, there are clearly two groups divided by the classification boundary at 0.5. Interestingly, the overall accuracy without cross-validation increased rather modestly to 96.8% (from 96.4 for cross-validation) indicating a small amount of over-fitting using this method.

### 2. Reduction of phonation features

The phonation features were extracted in order to catch different information in the pitch-enhanced spectrogram $S_h$ in formula (12). Most of them obtained a significant correlation with the ground truth, as shown in Table II. However, some of these features were found to correlate strongly with

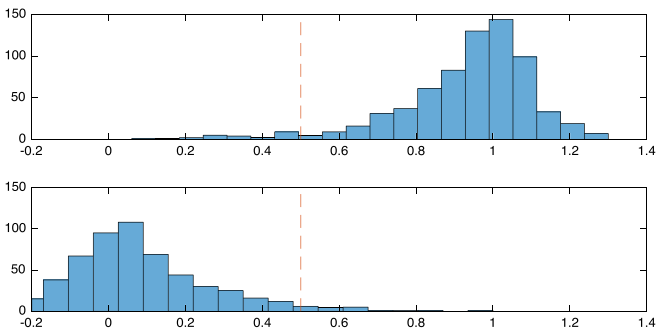


FIG. 8. (Color online) The output of the PLS regression using five components without cross-validation, for the phonation model. The upper histogram shows the distribution of the prediction for ground truth = 1 (phonation) and the lower histogram for ground truth = 0 (no phonation). The dashed line marks the classification boundary.

TABLE VI. Final prediction results for the SMV category.

| Cross-validation | PLS components | PLS accuracy (%) | SVM accuracy (%) | EMLP accuracy (%) |
|---|---|---|---|---|
| 10-fold | 8 | 86.0 | 88.9 | **90.9** |
| Leave-one-out | 5 | 83.8 | 85.9 | **87.2** |

each other within the current database. This was to a certain extent expected since they all were chosen for detecting phonation. The average pairwise correlation between all phonation features was 0.53, disregarding the sign. The highest correlations ($r = 0.95$ and 0.98) was found between the highest peak value (hf0_maxsl_uqt) and the difference in sound level between the highest and the second highest peaks (hf0_maxsldiff_uqt and hf0_maxsldiff_std). This situation can make a feature selection a bit arbitrarily and depend on the database. Nevertheless, an automatic procedure was applied using SVM and leave-one-participant-out. The independent contribution of each phonation feature was estimated by running the model without this feature. Then, all the features that contributed negatively (made the $R^2$ increase when the feature was omitted) were removed. This resulted in a set of eight features (hf0_maxf0_uqt, hf0_maxf0_std, hf0_max2sl_std, hf0_max2f0_mva, hf0_meanrestsl_uqt, hf0_meanrestsl_std, hf0_meanrestsl_mva, hf0_maxsldiff_std). For all the phonation features (18) the resulting $R^2 = 95.6$, thus a small decrease (0.3) in comparison with all features. For the selected features (8) the explained variation was even slightly higher with $R^2 = 96.0$. In conclusion we see that a subset of eight phonation features were effective and sufficient for predicting phonation.

### C. Classification of SMV

#### 1. Overall results using all features

The results of the classification of the SMV category for different cross-validations and methods are summarized in Table VI. Here the overall classification accuracy was slightly lower than for the phonation class, ranging from 84% to 91%. The EMLP method gave the best results and the differences between methods were larger. This indicates that the features still had some problems capturing the acoustical properties of SMV, as was indicated also by the correlations. This is not surprising, since these vibrations span across widely different types, each with different characteristic frequencies, as listed in Sec. II C.

Figure 9 shows the results from the PLS regression, applied using five components and without cross-validation. Here there is more overlap between the two groups. This result for the regression is not surprising, since the design of for example the modulation filter bank requires a machine learning model that can handle feature interaction. Also, the overall accuracy without cross-validation here increased rather modestly to 86.5%, indicating a small amount of overfitting using this method.

#### 2. Comparison of the different groups of SMV features

The correlation analysis presented in Table III indicated a rather weak coupling to supraglottal myoelastic vibrations,

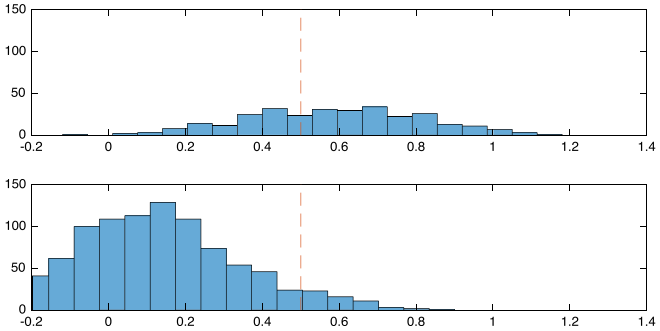


FIG. 9. (Color online) The output of the PLS regression using five components without cross-validation for the SMV model. The upper histogram shows the distribution of the prediction for ground truth = 1 (SMV) and the lower histogram for ground truth = 0 (not SMV). The dashed line marks the classification boundary.

and did not clearly indicate the feature group that might be the best candidate for prediction, although the modulation filterbank had the relatively largest correlations. In order to further investigate the differences between the SMV feature groups, we predicted the ground truth separately for each group, using both PLS and SVM classification. As seen in Table VII, the filterbank method obtained a slightly better accuracy for all methods. Note also that the filterbank group had the lowest number of features, indicating that these features were relatively more efficient. The difference in accuracy was 1.4 to 2.2 between the filter bank group and the ISL group.

The modulation filterbank features with SVM and tenfold cross-validation obtained an accuracy of 85%. Compared to the SVM method for the full features set (Table VI), the decrease in accuracy was about 4%, indicating that the filter bank features indeed capture some of the salient information within the whole feature set. For SVM and leave-one-participant-out this difference was even smaller and about 2%. This implies that a simplified model for SMV using only those six features can be implemented by a set of well-known straightforward signal-processing methods, including sound level and FFT computations.

### D. Classification of turbulence

#### 1. Overall results using all features

The results of the classification of the turbulence category for different cross-validations and methods are summarized in Table VIII. Here the overall classification accuracy was slightly lower than for the SMV category, ranging from 82% to 89%. The EMLP method gave the best result using tenfold cross-validation, while SVM obtained the best results for leave-one-out. The differences between methods were larger than for the phonation category, thus, it seems to perform in a way similar to the SMV category. This indicates that the features still had some problems capturing the acoustical properties of turbulence. Contrary to the SMV features, the turbulence features were rather straightforward to extract, and the intuitive impression was that they worked well. The relatively lower performance could possibly be attributed to the acoustic overlap between SMV and

TABLE VII. Comparison of the different SMV feature groups using PLS and SVM classification. The best accuracy (indicated in bold) was obtained by the filterbank features for both classification methods and cross-validations.

| | | | 10-fold cross-validation | | | Leave-one-out cross-validation | | |
|---|---|---|---|---|---|---|---|---|
| Feature group | Variables | No. of features | PLS components | PLS accuracy (%) | SVM accuracy (%) | PLS components | PLS accuracy (%) | SVM accuracy (%) |
| Modulation filter bank | vibspecdb1-6 | 6 | 3 | **81.4** | **85.1** | 4 | **80.0** | **83.9** |
| Vibrato extraction spectrogram | vib…1-2 | 8 | 5 | 78.3 | 77.7 | 4 | 77.9 | 77.4 |
| Vibrato extraction ISL | vib…3-5 | 14 | 3 | 79.2 | 83.5 | 3 | 78.2 | 82.5 |
| ASDF | cor_… | 8 | 2 | 75.9 | 77.0 | 2 | 75.8 | 75.3 |

turbulence features. Turbulence is simply the existence of noise in the signal. However, all the different types of SMV also generated noise although they were not annotated as turbulence. Thus, the detection of the SMV category needs to resolve these cases from the turbulence category by an interaction between all features. This could also explain why there is a relatively large difference between the methods. Note, that the PLS method uses a linear combination of features, and thus does not include any interaction in the model.

Figure 10 shows the results from the PLS regression, applied using six components and without cross-validation. Although the overall accuracy is comparable to the SMV case, the distribution indicates a better discrimination between the positive and negative groups in the figure. As in the previous cases the overall accuracy without cross-validation increased rather modestly to 86.1% indicating a small amount of over-fitting using this method.

### 2. Reduction of turbulence features

As for the phonation features we used SVM and leave-on-participant-out cross-validation for evaluating the different turbulence feature. The resulting explained variance using all turbulence features (30) obtained then an $R^2 = 80.4\%$. We then compared the two-band versus the seven-band features. Using only the two-band features (6), $R^2 = 72.8\%$. Using only the seven-band features (21), $R^2 = 81.1\%$. This indicated that the two-band features were not sufficient for detecting turbulence and also contributed negatively to the overall prediction and were therefore omitted. The independent contribution of each of the seven-band features (21) in combination with the harmonics-to-noise features (3) was finally estimated and all positive contributions were retained. This resulted in eight features (nosB7_2_uqt, nosB7_3_uqt, nosB7_4_uqt, nosB7_6_uqt, nosB7_7_uqt, nosB7_7_std, nosB7_7_mva, nosH2N_uqt) with an $R^2 = 80.7\%$. This feature selection corresponded well with the expectation. Most of the features averages using upper quartile (_uqt) were selected and there was a focus on the highest frequency band.

TABLE VIII. Final prediction results for the turbulence category.

| Cross-validation | PLS components | PLS accuracy (%) | SVM accuracy (%) | EMLP accuracy (%) |
|---|---|---|---|---|
| 10-fold | 6 | 84.6 | 87.4 | **88.5** |
| Leave-one-out | 6 | 81.6 | **83.5** | 83.1 |

## E. Final predictions for the reduced feature set

For the final prediction we included the reduced feature set from each classification. It included the eight phonation features (hf0_maxf0_uqt, hf0_maxf0_std, hf0_max2sl_std, hf0_max2f0_mva, hf0_meanrestsl_uqt, hf0_meanrestsl_std, hf0_meanrestsl_mva, hf0_maxsldiff_std), the six filterbank features from SMV (vibspecdb1–6), and the eight turbulence features (nosB7_2_uqt, nosB7_3_uqt, nosB7_4_uqt, nosB7_6_uqt, nosB7_7_uqt, nosB7_7_std, nosB7_7_mva, nosH2N_uqt), thus, totally 22 features. The final result for all three methods and the two cross-validations are shown in Table IX. As indicated by the numbers in parentheses, the difference between this feature set and all 84 features were in most cases rather small and both positive and negative. The prediction of phonation did not change much for the SVM and EMPL methods, while PLS decreased somewhat. Note that the prediction of phonation using only the selected eight features resulted in a similar accuracy ($R^2 = 96.0$ for SVM and leave-on-out) These results indicate that the prediction of phonation is quite stable, that a few features are sufficient, and that the prediction method is less important. The EMLP method obtained the best results in four out of six cases but the difference in comparison with SVM was in general rather small. The PLS method obtained mostly a decrease in accuracy in comparison with the full feature set. Also notable is that more PLS components were needed for the reduced feature set.

## VII. SUMMARY AND DISCUSSION

Using a set of features developed using extensions to the ARF Toolbox, the three different articulation categories, phonation, supraglottal myoelastic vibrations (SMV) and turbulence, were predicted using PLS regression, SVM, and EMLP. The model that performed best was in most cases the EMLP method and the classification accuracy was for ten-fold cross validation 96.9% for phonation, 90.9% for SMV, and 88.5% for turbulence for the full feature set (84 features). Note that this corresponds to the correlations 0.98, 0.95, and 0.94, respectively, between the ground truth and the prediction. Thus, despite the sometime rather low correlations for individual features, the models using machine learning could extend these results considerably, by combining the features and including interaction effects.

A reduced feature set of 22 features could predict phonation with a similar accuracy as the full feature set while the prediction of SMV and turbulence decreased approximately

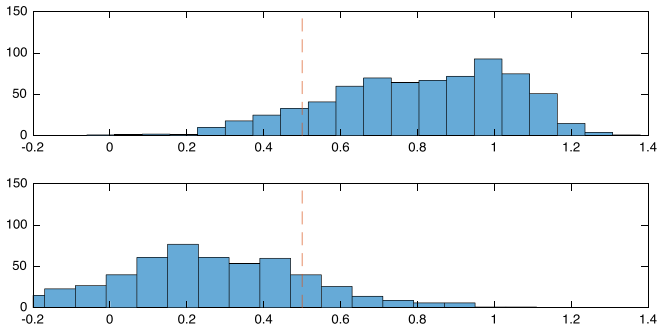

FIG. 10. (Color online) The output of the PLS regression using six components without cross-validation for the turbulence model. The upper histogram shows the distribution of the prediction for ground truth = 1 (turbulent) and the lower histogram for ground truth = 0 (not turbulent). The dashed line marks the classification boundary.

1%. Note that the feature reduction was using the results of the prediction in order to select the most important features. For phonation and turbulence feature selection we used SVM with leave-one-out and for SMV we used both SVM and PLS with both cross validation methods. This procedure is to a certain extent violating the cross-validation since the testing of new data is used for modifying the model. It could also possibly favor the methods used for the feature selection in the final prediction. Therefore, the full feature set could be considered as more unbiased relative to the current dataset.

The features derived from a modulation filter bank were able to predict SMV with an accuracy of 85% (tenfold cross-validation) using only six features. Thus, they can be used as a starting point for making a rather simple implementation of the model using only standard signal processing techniques.

One possible reason for the lower results for SMV category could be the unbalanced groups—this category contained proportionally more cases in the negative groups. This could possibly also be improved by further optimization of parameters both for the PLS and SVM method. However, the lower results for the PLS method compared to SVM and EMLP is likely due to the fact that it disregards any interactions between the features.

The lowest results were obtained for the turbulence category. We assumed *a priori* that turbulence should be strongly related to the amount of noise in the signal. Obviously, air turbulence will generate noise. However, the definition of turbulence in the annotations is a bit different. For example, supraglottal myoelastic vibrations without any extra sound source are not classified as turbulent, although there is usually a considerable amount of noise in the signal. A further comparison of the criteria for the annotations, as well as an analysis of the incorrectly classified examples in each category, seems to be an important path for future development.

TABLE IX. Final prediction results for all three main classes and three prediction methods using the reduced feature set of 22 features. Numbers in parentheses indicate the change in accuracy from the prediction using all the original 84 features (Tables V, VI, and VIII).

| Prediction class | Cross-validation | PLS comp. | PLS accuracy (%) | SVM accuracy (%) | EMLP accuracy (%) |
|---|---|---|---|---|---|
| Phonation | 10-fold | 6 | 95.8 (−0.6) | **96.8** (+0.2) | 96.7 (−0.2) |
| | Leave-one-out | 8 | 95.5 (−0.6) | 95.9 (0) | **96.1** (+0.2) |
| SMV | 10-fold | 12 | 85.1 (−0.9) | 88.8 (−0.1) | **89.8** (−1.1) |
| | Leave-one-out | 13 | 84.2 (+0.4) | **87.7** (+1.8) | 87.2 (0) |
| Turbulence | 10-fold | 11 | 82.9 (−1.7) | 86.7 (−0.7) | **87.5** (−1.0) |
| | Leave-one-out | 10 | 80.7 (−0.9) | 83.7 (+0.2) | **84.0** (+0.9) |

The current approach using a multitude of extracted features that are combined using machine learning seems to have a great advantage in comparison with the approach of developing one feature for describing, for example, phonation. This approach could be further extended to voice quality estimation and assessment.

The developed features have much in common with previously described features for characterizing voice quality used in several studies and analysis programs. In a future study, it would be interesting to compare the specific features developed in the current study with these established voice measures within the context of voice quality assessment.

## ACKNOWLEDGMENTS

This research was a part of the project SkAT-VG, by the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open Grant No. 618067. It was partly supported by the Swedish Research Council, Grant No. 2012-4685. We would also like to thank one anonymous reviewer for many insightful suggestions.

Alías, F., Socoró, J. C., and Sevillano, X. (**2016**). "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," Appl. Sci. **6**(5), 143.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (**2008**). "SURF: Speeded up robust features," Comput. Vis. Image Understand. **110**(3), 346–359.

Brown, J. C., and Puckette, M. S. (**1992**). "An efficient algorithm for the calculation of a constant *Q* transform," J. Acoust. Soc. Am. **92**(5), 2698–2701.

Brugman, H., and Russel, A. (**2004**). "Annotating multimedia/ multi-modal resources with ELAN," in *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*, developed at Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, NL, https://tla.mpi.nl/tools/tla-tools/elan/ (Last viewed September 26, 2017).

Burred, J. J., and Lerch, A. (**2004**). "Hierarchical automatic audio signal classification," J. Audio Eng. Soc. **52**(7/8), 724–738, available at http://www.aes.org/e-lib/browse.cfm?elib=13015.

Carding, P. N., Steen, I. N., Webb, A., Mackenzie, K., Deary, I. J., and Wilson, J. A. (**2004**). "The reliability and sensitivity to change of acoustic measures of voice quality," Clin. Otolaryngol. **29**(5), 538–544.

Chang, C. C., and Lin, C. J. (**2011**). "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Tech. (TIST) **2**(3), 1–39.

Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**(4), 1917–1930.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation, I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**(5), 2892–2905.

Elowsson, A. (**2016**). "Beat tracking with a cepstroid invariant neural network," in *17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pp. 351–357.

Elowsson, A., and Friberg, A. (**2015**). "Modeling the perception of tempo," J. Acoust. Soc. Am. **137**, 3163–3177.

Elowsson, A., and Friberg, A. (**2017**). "Predicting the perception of performed dynamics in music audio with ensemble learning," J. Acoust. Soc. Am. **141**, 2224–2242.
Elowsson, A., Friberg, A., Madison, G., and Paulin, J. (**2013**). "Modelling the speed of music using features from harmonic/percussive separated audio," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2013)*, pp. 481–486.
FitzGerald, D. (**2010**). "Harmonic/percussive separation using median filtering," in *Proceedings of DAFx-10*, Graz, Austria (September 6–10).
Friberg, A., Schoonderwaldt, E., and Juslin, P. N. (**2007**). "CUEX: An algorithm for extracting expressive tone variables from audio recordings," Acta Acust. united Acust. **93**, 411–420, available at https://www.ingentaconnect.com/contentone/dav/aaua/2007/00000093/00000003/art00010.
Geladi, P., and Kowalski, B. R. (**1986**). "Partial least-squares regression: A tutorial," Anal. Chim. Acta. **185**, 1–17.
Gorham-Rowan, M. M., and Laures-Gore, J. (**2006**). "Acoustic-perceptual correlates of voice quality in elderly men and women," J. Commun. Disorders **39**(3), 171–184.
Hansen, L. K., and Salamon, P. (**1990**). "Neural network ensembles," IEEE Trans. Pattern Anal. Mach. Intell. **12**, 993–1001.
Heman-Ackah, Y. D., Michael, D. D., and Goding, G. S. (**2002**). "The relationship between cepstral peak prominence and selected parameters of dysphonia," J. Voice **16**(1), 20–27.
Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). "Acoustic correlates of breathy vocal quality," J. Speech Lang. Hear. Res. **37**(4), 769–778.
Hillenbrand, J., and Houde, R. A. (**1996**). "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Lang. Hear. Res. **39**(2), 311–321.
Ladefoged, P., and Maddieson, I. (**1996**). *The Sounds of the World's Languages* (Blackwell Publishers, Oxford, UK).
Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge).
Lemaitre, G., Houix, O., Misdariis N., and Susini P. (**2010**). "Listener expertise and sound identification influence the categorization of environmental sounds," J. Exp. Psychol.: Appl. **16**(1), 16–32.
Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., and Susini, P. (**2016a**). "Vocal imitations of non-vocal sounds," PLoS One **11**(12), e0168167.
Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., and Susini, P. (**2016b**). "Vocal imitations of basic auditory features," J. Acoust. Soc. Am. **139**(1), 290–300.
Lemaitre, G., Scurto, H., Françoise, J., Bevilacqua, F., Houix, O., and Susini, P. (**2017**). "Rising tones and rustling noises: Metaphors in gestural depictions of sounds," PLoS One **12**(7), e0181786.
Lemaitre, G., Voisin, F., Scurto, H., Houix, O., Susini, P., Misdariis, N., and Bevilacqua, F. (**2015**). "A large set of vocal and gestural imitations," Deliverable 4.4.1 in the EC-project Sketching Audio Technologies using Vocalizations and Gestures (SkAT-VG), http://skatvg. iuav.it/wp-content/uploads/2015/11/SkATVGDeliverableD4.4.1.pdf (Last viewed September 5, 2018).
Lindeberg, T., and Friberg, A. (**2015a**). "Idealized computational models for auditory receptive fields," PLoS One **10**(3), e0119032.
Lindeberg, T., and Friberg, A. (**2015b**). "Scale-space theory for auditory signals," in *Proceedings of Scale Space and Variational Methods in Computer Vision (SSVM 2015)*, Vol. 9087 of Springer Lecture Notes in Computer Science, pp. 3–15.
Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., and Corthals, P. (**2009**). "Acoustic measurement of overall voice quality: A meta-analysis," J. Acoust. Soc. Am. **126**(5), 2619–2634.
Moisik, S. R. (**2013**). "The epilarynx in speech," Ph.D. thesis, University of Victoria, Department of Linguistics, Canada.
Moisik, S. R., Esling, J. H., and Crevier-Buchman, L. (**2010**). "A high-speed laryngoscopic investigation of aryepiglottic trilling," J. Acoust. Soc. Am. **127**(3), 1548–1558.
Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The timbre toolbox: Extracting audio descriptors from musical signals," J. Acoust. Soc. Am. **130**, 2902–2916.
Polikar, R. (**2006**). "Ensemble based systems in decision making," IEEE Circ. Syst. Mag. **6**(3), 21–45.
Prame, E. (**1994**). "Measurements of the vibrato rate of ten singers," J. Acoust. Soc. Am. **96**, 1979–1984.
Rao, V. M. (**2011**). "Vocal melody extraction from polyphonic audio with pitched accompaniment," Ph.D. thesis, Indian Institute of Technology Bombay, Department of Electrical Engineering, Bombay.
Smola, A. J., and Schölkopf, B. (**2004**). "A tutorial on support vector regression," Stat. Comput. **14**(3), 199–222.
Ternström, S., and Mauro, D. A. (**2015**). "Extensive set of recorded imitations," Deliverable D2.2.2 in the EC-project Sketching Audio Technologies using Vocalizations and Gestures (SkAT-VG), http://skatvg.iuav.it/wp-content/uploads/2015/01/SkATVGDeliverableD2.2.2.pdf (Last viewed September 5, 2018).