Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis

Albert Rilliard, Christophe d'Alessandro, Marc Evrard

Paradigmatic variation of vowels in expressive speech: acoustic description and dimensional analysis

Albert Rilliard*
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France &
Universidade Federal do Rio de Janeiro, CNPq, Brazil

Christophe d'Alessandro
Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7190,
Institut Jean Le Rond d'Alembert, 4 place Jussieu, F-75005 Paris,France

Marc Evrard
LIMSI, CNRS, Université Paris Sud, Université Paris-Saclay, F-91405 Orsay, France †
(Dated: October 26, 2023)

Acoustic variation in expressive speech at the syllable level is studied. As emotions or attitudes can be conveyed by short spoken words, analysis of paradigmatic variations in vowels is an important issue to characterize the expressive content of such speech segments. The corpus contains 160 sentences produced under seven expressive conditions (Neutral, Anger, Fear, Surprise, Sensuality, Joy, Sadness) acted by a French female speaker (a total of 1120 sentences, 13140 vowels). Eleven base acoustic parameters are selected for voice source and vocal tract related feature analysis. An acoustic description of the expressions is drawn, using the dimensions of melodic range, intensity, noise, spectral tilt, vocalic space, and dynamic features. The first three functions of a discriminant analysis explain 95% of the variance in the data. These statistical dimensions are consistently associated with acoustic dimensions. Covariation of intensity and F0 explains over 80% of the variance, followed by noise features (8%), covariation of spectral tilt and F0 (7%). On the basis of isolated vowels alone, expressions are classified with a mean accuracy of 78%.

PACS numbers: PACS: 43.71.Bp; 43.70.Fq; 43.71.Es

I. INTRODUCTION

A. Aims of the study

In everyday speech, emotions, attitudes or other types of expressions can be conveyed by even the shortest words. Variations of acoustic features in monosyllabic words or interjections allow us to express, e.g., fear, joy or anger.

Investigating acoustic variation in expressive speech at the syllable level is an important issue to understand human communication or to design automatic processing systems. This has rarely been studied so far [but see 5, 44, 55, for works on a similar time-frame] since two main approaches have dominated studies of the acoustic variations in expressive speech. On the one hand, expressive variations have been studied on the syntagmatic axis, in terms of patterns or contours along successive syllables, phrases, or sentences [54]. This approach is related to prosodic variations [intonation, rhythm, intensity, and voice quality parameters like the shape parameter R_d or the Normalized Amplitude quotient NAq: 3, 14, 28]. On the other hand, multidimensional voice quality analyses of expressive speech are usually based on long-term averages of acoustic quantities [e.g., 34, 60]. The number of dimensions is generally large, with redundant parameters. Short-term variations, as well as fine phonetic details, are lost in the process. Both approaches can hardly address the question of expressive variation in syllable-sized segments.

In the present work, paradigmatic variations for syllable-sized segments are studied. Vowels segments are chosen instead of full syllables. With a reduced set of phonetic classes compared to syllables, each vowel class presents more occurrences in a corpus, and syllabic nuclei in French are based on a vowel. Finally, as vowels are voiced segments, voice quality and fundamental frequency analyses can be performed in a comparable way.

A first aim of the study is an acoustic description. A reduced set of acoustic-phonetic dimensions (like vocal effort, melody, tension, noise, supraglottal, and dynamic features) is proposed to characterize vocal expressions. A second aim is to unveil the main acoustic dimensions for the discrimination of these expressions, according to the statistical organization of the acoustic space.

The remaining of this introduction presents a review of the literature on acoustic variations in expressive speech, which will serve as a basis for discussing our results. In a first section, the phonetic bases for acoustic analysis of expressive speech are reviewed. Then, the acoustic dimensions of voice linked to the encoding of emotions and vocal expressivity are reviewed.

* albert.rilliard@limsi.fr
† Current address: Toyota Technological Institute, Computational Intelligence Laboratory, 2 Chome-12-1 Hisakata, Tenpaku, 468-8511 – Nagoya, Japan

B. Phonetic descriptions of voice quality

Laver [48] uses *voice quality* as "the characteristic auditory colouring of an individual speaker's voice, and not in the narrower sense of the quality deriving solely from laryngeal activity." (p. 1). He thus distinguishes between a set of changes originating at the glottis and another linked to variable conformations of the vocal tract. He also proposes a general *tension* setting that affects both laryngeal and supralaryngeal settings of voice quality. This scheme has been proven effective to perceptually analyze possible phonetic variations in voice [e.g., 12, 49, 53].

The effects of the glottal source on voice are linked to pitch and loudness, and to the existence of noise. Two types of noises are described: additive noise linked to breath, and harsh noise linked to irregular vibrations of the folds (originating from creak or harsh voices). Thus, following d'Alessandro [2], the source component of voice quality can be summarized as a production of voice with several modes of regular vibrations of the vocal folds; the modes are linked to varying vocal folds' tension and conformations [see 39]. They are possibly produced with additive noise (breathy voice) or irregularities in the folds vibrations, leading to the perception of harsh voices [see 31, for the description of these aperiodicity types].

The effects of the vocal tract on voice quality may be summarized as (i) changes in the tract length, (ii) changes in the vocal cavity width, and (iii) nasalization. Changes in the vocal tract length are typically obtained by changing either the shape of lips [protruded/stretched—cf. 59, for a discussion about lip stretching and vocal tract length] or the position of the larynx (raised/lowered). Changes of the width of the vocal tract are obtained by a change in the configuration of articulators—the most notables being a fronted/retracted position of the tongue body, a contracted/expanded pharynx, and an open/closed jaw position.

One of these voice quality settings is barely observed alone in spontaneous speech. For example, changes in one articulators position are correlated to (or compensated by) changes in another: the jaw typically influences the vertical position of the tongue and lips. Moreover, Laver's general dimension of tension links laryngeal and supralaryngeal effects. Tension in the voice is linked to the muscular tension in the speaker, which relates to the notion of vocal effort, as described by Liénard & Di Benedetto [51] and Traunmüller & Eriksson [71]. An important muscular tension produces a tense voice, which is generally characterized by a raised larynx, constricted pharynx, tense and thin vocal folds, mobile jaw and lips. Such tense voices tend to be acoustically characterized by a high pitch, high loudness, more high-spectrum energy, and a raised first formant (F_1). On the opposite, lax voices are produced with a low larynx, larger pharynx, breathy phonation, and restricted movements in the jaw and lips. These settings generate a voice with a low pitch, low energy and low high-spectrum energy, with additive noise, a lower F_1, and a reduced formant space [2, 48]. The tension parameter is particularly interesting because it links various dimensions of voice quality to a single explicative factor [45]. It is to be related to the Effort Code proposed by Gussenhoven [35] and to the "4^{th} prosodic dimension" described in Campbell & Mokhtari [14].

The various characteristics of voice quality and expressive or emotional voices are carried by the acoustic signal. Analyses of acoustic parameters are generally based on the source-filter model of voice production [27]. A large number of algorithms for voice quality analysis have been proposed [11]. Inversion techniques try to determine the parameters of such models [3, 20, 28, 36]; most of these approaches focus on the source aspect of voice quality. Some works also propose a resynthesis framework to evaluate the perceptual salience in the resulting signal of controlled changes in the model's parameters [16, 32, 33, 46].

C. Vocal dimensions of emotional expressions

Scherer's Component Process Model (CPM) predicts physiological changes induced by the appraisal process of external stimulus' novelty, valence, and relevance for an individual. The CPM predicts physiological responses induced in individuals by appraisal processes and motivational changes. These responses are essentially linked to the pleasantness of the stimulus, its obstructive or conducive nature, the coping potential of the individual facing the stimulus, and the behavioral response selected to face that stimulus [e.g., a display of power, of submission, of withdrawal— 65, 66]. According to this model, positive events lead to lip spreading and vocal tract widening; a conducive stimulus may trigger a relaxation, thus lowers energy and fundamental frequency (F_0), and leads to a lax voice—while a pleasant stimulus may induce a higher power response, and notably a higher energy (linked to a tenser voice). Negative stimuli may induce a muscular contraction, thus a narrower vocal tract, tenser vocal folds, but depending on the individual's coping potential, it may result in various expressive changes. A withdrawal behavior induces a hypotonic vocal tract, with very low F_0 and energy. Displays of power in situations of control lead to higher energy with a relatively low F_0, while situations of low power where control is possible lead to high pitch and a comparatively low energy.

These predictions address changes linked to emotional variations and to their correlated expressions. One can find within this description clear links between the phonetic aspects of voice quality, the emotional motivation of the observed changes, and their perceived effect on the hearer [for details on perception, see 10].

Studies aiming at retrieving emotions in speech and voice are based on long-term averages of large inventories of acoustic quantities [e.g., 37, 42, 47, 60, 74]. Most

of these approaches have classification aims; typically, a few categories (or dimensions, appraisals) of emotions are proposed, extracted from psychological models of emotions [e.g., 21, 65, 77]. These works target the accuracy of their learned recognition models, which function as black boxes, and do not often report on the relative weight of parameters to classification accuracy [meanwhile see 42]. Some researchers propose approaches involving an explicit descriptive target: e.g., Banse & Scherer [6] or Goudbeek & Scherer [34] aim at validating the predictions of the CPM. The acoustic parameters used in these approaches, being extracted from an inversion process, or calculated on spectral (or cepstral) representations, are measured on short-term (typically 30 ms or pitch synchronous) frames. Extracted values are then averaged over utterances. Thus, short-term variations, as well as fine phonetic details, are lost in the process.

Works and predictions made for emotional expressions shall be distinguished from expressive changes in voice for and during spoken communication that—if undoubtedly related—are also produced under various other constraints, an important one being the intelligibility of the message.

D. Encoding of spoken expressivity

The expressions performed during spoken communication are controlled, intended [*voluntary* in 57, terms], and participate in the speech act [56, 75]. Hurley [40] describes a model that explains the communicative skills and their cultural encoding on the basis of imitation, deliberation and mindreading capabilities of humans, which include understanding the behavior of self and another, particularly emotional expressions and intentions. Self-representations of emotional expressions enable individuals to deduce, from a perceived voice, the emotion felt (or displayed) by their interlocutor [57]—emotions the listeners may then label according to their own experience and using the verbal categories available in their language [76]. This process of reproducing voice changes that are carrying symbolic values is at the core of verbal expressivity (in both production and perception). Theoretical codes have been proposed, derived from observations of the constraints imposed by the vocal apparatus on voice, to express various speech acts.

Ohala proposes the *Frequency Code*, which assumes that changes in the voice pitch may be used to signal the speaker's size and strength [59]. The conventionalization of this code within social and linguistic communication systems explains the expressions of, e.g., imposition vs. more submissive speech acts. The predictions of the Frequency Code match those of Scherer [65, 66] for high and low power displays—including the original use of smile (lip-corner retraction) as a submissive display [cf. 59, p. 332–335, and 65, p. 3465]. Gussenhoven [35] added an *Effort Code*, named after situations requiring a stronger articulatory effort. It is linked to displays such as showing enthusiasm or authority. An increased effort leads to wider pitch movements and hyperarticulated speech, which is consistent with the tension setting described by Laver [48]. These codes give hints at the conventionalization process that evolves symbolic signals into the abstract shapes embedded in languages [50, 59]. Acoustic changes do not only arise from the control of the glottis parameters. Léon [50, p. 77–79] describes a strategy used to produce charming voice involving both a breathy phonation and a fronted articulation, the latter having a symbolism of younger voice, linked to the perception of a shorter vocal tract. Expressive voices are thus intricately linked with physiological constraints and articulatory processes imposed on the speaker's vocal tract.

The article is organized into four parts. The corpus and the selected acoustic parameters are presented in Section II. Section III describes the main acoustic dimensions observed in the data. In section IV, the data main statistical dimensions are detailed. The results of section III and IV are compared and discussed in Section V, and compared to the results reported in this introduction.

II. CORPUS AND METHODS

A. Corpus design and recording

The corpus used for this study has been initially designed, recorded, and assessed for expressive synthesis. It is a relatively large phonetically balanced corpus, containing the same phonetic material (160 sentences covering all the phones and diphones of French) under seven emotive conditions (coined Neutral, Anger, Fear, Joy, Sadness, Sensuality, and Surprise—cf. infra). To the best of our knowledge, there was no comparable corpus (complete phonetic coverage, seven expressions for the same sentences) available in French [see 17, or 22, for surveys on available databases].

As our aim was not to study the emotional phenomenon in its complexity, but rather to study acoustic dimensions in expressions, acted expressions seemed appropriate. Naturally occurring expressions present gradual variations across the spectrum of affects, while acted data presents a caricature of prototypical expressions [15]. The simplicity of expressive labeling (one of the dominant burden in natural corpora) and a reproducible set of variations across sentences have been preferred. This choice may reduce the scope of our finding in the domain of emotional expressions, but make them more reliable as far as the acoustic description of voice quality is concerned [for a discussion on the usefulness of variation among corpus, see 73].

A female speaker of Parisian French (professional actress, aged 31 y.o.) was recorded reading sentences under seven expressive conditions. The expressivity corresponding to each base was presented via scenarios featuring typical situations. The speaker was free to interpret them in her own way but was asked to stay consistent

across each respective base. The 160 sentences were selected for phonetic diversity, as a part of a 1402 sentences corpus (the 160 first sentences of the complete set present the better phonemic coverage). Phonetic coverage was obtained from an open text database of French, using a greedy algorithm [24]. Acoustic (AKG 414-XLII condenser microphone) and electroglottographic (EGG, Glottal Enterprises EG2-PCX2) signals were recorded in a recording booth (48kHz, 16 bits digital format). Recording level was calibrated before recording using a Brüel & Kjær acoustical calibrator; sound pressure level was then corrected for recording level.

Recordings were segmented in sentences. Phonemic and syllabic boundaries were obtained using grapheme-to-phoneme conversion and a forced-alignment procedure, and subsequently stored in *Praat* TextGrid files [9]. Phonemic boundaries, particularly pronunciation differences between expressive situations (e.g., schwa deletion or insertion) were manually checked and corrected. The complete corpus contains 160 sentences for each of the seven expressive conditions, thus 1120 utterances. It corresponds in average to 4350 phonemes per base, including about 1877 vowels (from 1867 up to 1881), for a total of 13140 vowels that form the basis of the present work.

B. Elicited expressive variations

The expressive conditions have been selected to maximize the elicited acoustic variation, according to the methodology used for the GEMEP corpus [7], based on Scherer's CPM [65]. The speaker received the following instructions about the emotional state represented by each expression (the English labels used in the paper are given here before a more detailed description, original labels were in French):

1. "Neutral" (French *neutre*): a neutral declarative reading of the sentence, with no specific expressivity; the speaker just provides a piece of information (neutral valence, arousal, and power).

2. "Anger" (*colère*) corresponds to GEMEP hot anger/rage and has a negative valence, a high arousal, with a display of power.

3. "Fear" (*peur*) corresponds to GEMEP anxiety/worry and has a negative valence, a low arousal, without power display, but seeking control.

4. "Joy" (*joie*) corresponds to GEMEP joy/elation and has a positive valence, a high arousal, in a situation of control (high power).

5. "Sadness" (*tristesse*) corresponds to GEMEP sadness/depression and has a negative valence, a low arousal, in a situation of withdrawal (low power).

6. "Sensuality" (*sensualité*) to some extent corresponds to GEMEP tenderness and was built on the "charming voice" of Léon [50, pp. 77-79]; it has a positive valence, a low arousal, low power display, with a display of proximity to the listener.

7. "Surprise" (*surprise*) corresponds to GEMEP surprise, and has a neutral valence, power but aroused by the novelty of a stimulus.

The expressions actually perceived by listeners were evaluated through a perception test reported in Evrard et al. [25]. In summary, the correct recognition rate for Anger, Sadness, and Surprise is over 90%. Joy and Sensuality are respectively recognized with rates of 85% and 88%, and Joy shows some confusion with Sensuality. Fear has a recognition rate of only 68% and is mostly confused with Surprise (16%).

C. Base acoustic parameters

As discussed in the introduction, both source-related and tract-related variations are measured. These variations are represented by a set of eleven acoustic parameters, which are measured for each of the 13140 vowels:

- **Fundamental frequency** (F_0): F_0 measures the frequency of vibration of vocal folds, and is the closest acoustic estimate of the perceived pitch on voiced sounds [62, 69]; it is expressed in semitones relative to 1 Hz (ST) (scale close to perception, 58). F_0 is measured on the EGG signal [which is used to give reference values of F_0, e.g., in 18], using a dedicated algorithm in the COVAREP toolbox [19] proposed by Henrich et al. [38]. In about 4% of the observations, the EGG-based measurement fails (the lack of a correct EGG signal may be due to, e.g., a displacement of the electrodes); in those cases, F_0 was estimated from the audio signal through *Praat* default pitch detection algorithm. Raw estimations of F_0 were made each 5 ms, then the median of observations on each vowel was kept.

- **Inter-vocalic difference of** F_0 (ΔF_0): it represents the difference of pitches between two adjacent vowels (i.e., an estimation of the pitch movement between a syllable and the preceding one); ΔF_0 is expressed as the difference of median F_0 values measured on the two successive vowels. It is expressed in ST. ΔF_0 is zero for the first syllable in an utterance, negative if the current syllable is lower than the preceding and positive otherwise.

- **A-weighted intensity** (INT_A): weighted measure of the signal intensity that approximates the perceived loudness, expressed in decibel (dB). The INT_A measure is related to the perception of vocal effort [52] and is similar (albeit not F_0-dependent) to the "spectral emphasis" measure

used in Traunmüller & Eriksson [71]. It was measured (on the calibrated speech signal) using a dedicated *Praat* script [following 41]; raw measurements of INT_A were made each 10 ms, then the median of observations on each vowel was kept.

- **Vowel-to-vowel duration** (*V-to-V*): the *V-to-V* measure, introduced by Barbosa [8], is a measure of speech rhythm comparable in size to the syllable and calculated as the duration between the beginnings of two adjacent vowels; duration is standardized to account for phoneme intrinsic duration [13]. The smoothing based on a five-unit window described in Barbosa [8, p. 727] was not applied here. Raw measures of duration are extracted from the semi-automatic segmentation of sentences.

- **Harmonic-to-Noise Ratio (HNR) based on a fixed-frame periodic-aperiodic decomposition** (*PAP*): the ratio of energy in the harmonic component to the noise component, both being estimated using d'Alessandro et al. [1]'s decomposition algorithm. The latter is based on a fixed-size frame—effectively separating both additive noises (e.g., from turbulences at the glottis) and irregularities of the vocal folds vibrations (i.e., structural noise linked to jitter) in the noise component of the decomposed signals. This measure thus estimates the amount of aperiodic noise, compared to the periodic component, and expresses it in dB. The decomposition was performed using a local MATLAB® implementation of d'Alessandro et al. [1]'s algorithm.

- **HNR based on a pitch-scaled periodic-aperiodic decomposition** (*NOISE*): similar to the *PAP* measure, but using a pitch-scaled frame to determine the decomposition, Jackson & Shadle [43]'s algorithm separates the additive noise from the periodic part of each vocal fold's vibration cycle. This measure thus estimates the importance of additive noise in a periodic signal and expresses it in dB. The measure was performed using Jackson & Shadle [43]'s routines. Both measures of HNR use the F_0 estimation given by the EGG as an input. In both cases, the HNR was estimated comparing the relative intensity levels of the periodic and aperiodic signals using a MATLAB® script, with a 5 ms time step; the median of these raw HNR measurements over each vowel was kept.

- **Spectral tilt** (H_1-A_3): the difference between the amplitude of the first harmonic (H_1) and the amplitude of the third formant (A_3) gives an estimation of the spectral tilt [36]; it was computed using a 10 ms step with a *Praat* script using F_0 and F_3 estimations. Raw measurements in dB were then standardized for vowel influence (calculating the z-score for each vowel type across expressive bases); the median value over each vowel was kept.

- **First three formants** (F_1, F_2, F_3): to observe variations in supralaryngeal settings, the first three resonances of the vocal tract were estimated from the speech signal, using *Praat* "burg" procedure. The parameters used were those recommended for French female speakers by Gendrot & Adda-Decker [30, detecting five formants with a maximum frequency of 5500Hz for all vowels but the /u/, for which the maximum was set at 5000Hz; the default *Praat* time step was used]. The raw values in Hz were converted into mel scale [63] and also standardized for each vowel type (the mel and standardized values are used). The median of measurements (in mel or standardized) observed on each vowel was kept.

- **Vowel centralization** (*Centr.*): the degree of vowel centralization is an estimation of hypo-hyper articulation. It was measured as the Euclidean distance between a vowel median (F_1,F_2) values and the median (F_1,F_2) values for all the schwas in the corpus (taken as an estimation of a central articulation for the speaker), and expressed in mel.

D. Completeness of the corpus

In some cases (e.g., devoicing), some measures are undefined (e.g., F_0, H_1-A_3) and thus fail to return valid values. For about 4% of the vowels, at least one in eleven measurement procedures yield an undefined value. Table I shows that the Sensual expression is much more affected. Undefined values mostly affect H_1-A_3 and the two HNR measurements. This issue may be linked to a breathy voice quality, which affects F_0 measurements, a fundamental basis of these measures. The remaining data for the Sensual voice still amounts to 1469 vowels displaying valid measurements for all acoustic parameters.

TABLE I. Percentage of vowels having at least one undefined value, in each expressive condition.

Neutral	Anger	Fear	Joy	Sadness	Sensuality	Surprise
0.1%	0.2%	0.1%	0.3%	2.5%	21.9%	0.8%

The acoustic description of expressions is drawn according to the following acoustic dimensions: (i) INT_A and F_0, which represent the primary prosodic parameters, fundamental frequency related to pitch, and intensity related to vocal effort; (ii) H_1-A_3 and F_1, which represent the main features of spectral changes, linked to vocal effort and the voice lax/tense dimension; (iii) *PAP* and *NOISE*, which represent aperiodicities in the voice; (iv) the supraglottal features, which represent vocal tract shapes; and (v) the dynamic parameters (ΔF_0, *V-to-V*).

III. ACOUSTIC DESCRIPTION

All the acoustic parameters have been measured for a single speaker. The raw values represent only this speaker and could hardly be extended to others speakers as such (e.g., F_0 is highly gender-dependent). However, parameter covariations represent both the individual vocal strategies of the speaker (idiosyncratically or driven by linguistic, cultural, or phylogenetic constraints) and general physiological constraints. Therefore, general tendencies can be derived from this data and associated with results found in the literature.

A. Intensity and F_0

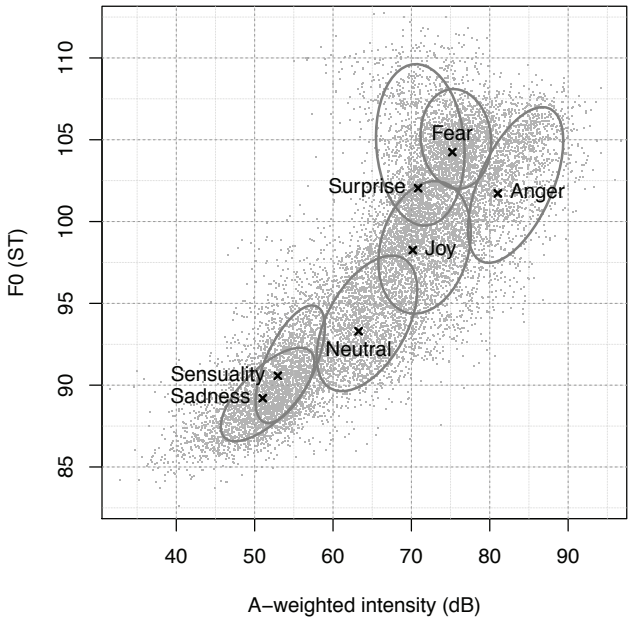

FIG. 1. Plot of each vowel on the INT_A and F_0 axes. The means observed for each expression are indicated by 'x'. Ellipses indicate the dispersion around the mean of half the data for a given expression.

Figure 1 presents a plot of the ($INT_A \times F_0$) values observed for all vowels, together with the mean position of each expression, and with ellipses showing half of the distribution around these means. It shows the spread of expressions along these two acoustic dimensions and the overlaps between categories. The shape of the distribution over the whole corpus illustrates the production constraints that link both quantities. Increased expressive arousal [34], expressed via a greater vocal effort, is related to an increased intensity from Sensuality and Sadness up to Anger, through Neutral voice and Joy. Voice intensity and F_0 are known to covary: Titze & Sundberg [70] proposes "an 8-9 dB increase in SPL per octave of F_0 rise" (p. 2946), while Liénard & Di Benedetto [51] found a 0.75 correlation between F_0 and amplitude measurements. A 0.83 correlation is observed on this corpus between A-weighted intensity and F_0. Thus, for this set of expressive conditions, and this speaker's specific strategy, the changes in F_0 are to a great extent explained by an increase in vocal effort—which can be linked to expressive arousal. Notably, changes in F_0 independent of INT_A are also observed: Fear and Surprise have higher F_0 for their respective level of intensity, as compared to the line going from Sensual up to Anger voice.

B. Tension setting

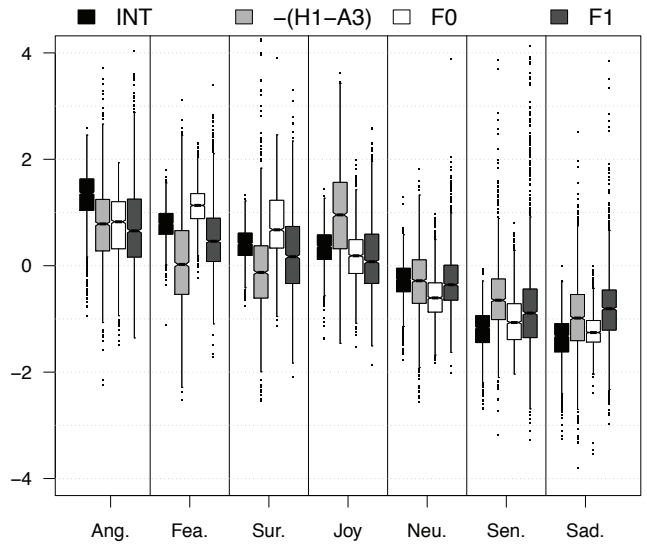

FIG. 2. Box plots of intensity (black), H_1-A_3 (light grey), F_0 (white), and F_1 (dark grey) values expressed in z-scores, for each expression ranked in decreasing order of intensity. Note that the opposite of $H_1 - A_3$ is plotted here, to make similarities and differences more visible as this measure is negatively correlated to other parameters.

In addition to intensity and F_0, two measures are related to the general tension setting described by Laver [48]: the spectral tilt $H_1 - A_3$, linked to the glottal component of tension, and F_1 linked to the supraglottal one. The $H_1 - A_3$ and INT_A measures have a negative correlation of -0.48, while F_1 and F_0 show a 0.59 correlation. Figure 2 shows the distributions of these four parameters (the opposite of $H_1 - A_3$ is plotted to account for its negative correlation), across expressions; all parameters are expressed in z-scores to be comparable. Distributions of F_1 and $H_1 - A_3$ show a wider spread within each expression, and their central values are less distinct than those of intensity and F_0; these parameters are thus less distinctive (in an information theoretic approach). The

F_1 parameter is essentially correlated to the changes in intensity and F_0. The $H_1 - A_3$ parameter shows different tendencies, for some expressions. Conversely to the trend observed for intensity, $H_1 - A_3$ shows higher values for Fear and Surprise (lower position on fig. 2), compared to its relatively low values observed for Joy (higher position on fig. 2). Note that the observation for Joy could be an effect of lip spreading (smile) or rounding. The contrasted changes in spectral slope for Fear and Surprise can also be related to changes in F_0, which are less correlated to intensity changes. Such F_0 changes could have been produced by tenser vocal folds.

C. Noise

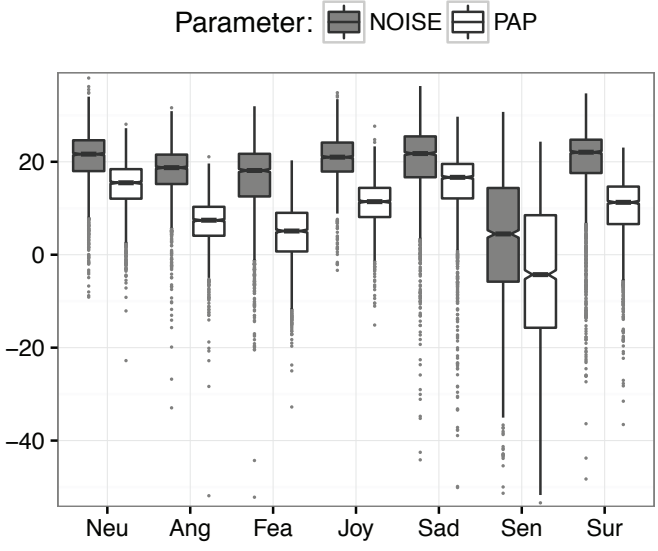

FIG. 3. Distributions of *NOISE* and *PAP* for each expression.

Figure 3 shows the distributions of the two HNR measures across expressions. The Sensual voice is produced with the highest level of aperiodicities (lesser HNR values); these aperiodicities being observed on both *PAP* and *NOISE*, one may conclude they are mostly due to additive noise (i.e., a turbulent air stream through the glottis). The *NOISE* measure did not detect much additive noise in the other expressions; on the contrary, the *PAP* measure does find some aperiodicities (that should thus be structural) for the expressions of Fear and Anger (and to a lesser extent Joy and Surprise). The importance of structural noise observed in these expressions is reduced compared to that of additive noise in Sensuality—with more than 9 dB difference between the median *PAP* measurements for Fear and Sensuality. The additive noise measured on the Sensual voice is to be linked to its breathy phonation, and it may give a distinctive cue to separate Sensual and Sad voices, otherwise comparable: see their position on the *intensity* $\times$ F_0 plane of fig. 1 and 2. The structural noise observed for Fear may be linked to short-term variations of F_0 in this expression, which can be related to the harshness described by Laver [48], associated with tension.

D. Supraglottal features

The acoustic measures linked to supraglottal settings are primarily the first three formant frequencies. A measurement of vowel centralization (on the $F_1 \times F_2$ plane) was computed, as well as the rhythm at the level of syllables (the *V-to-V* parameter), which is related to hyper or hypo-articulation. Changes in mean formant values, once corrected for vowel influence, show little changes related to expressive variations: the greatest changes have been already illustrated with F_1, which correlates to a global tension setting (see figure 2). A detailed view of formant changes based on their raw mel values (i.e., not vowel-standardized), taking into account each vowel class, gives some clues about their variation across expressions. Figure 4 presents the distribution of oral vowels, obtained for each expression, and based on the first three formants. Each plot represents the median position of formants of each vowel, obtained from Neutral speech (in gray) and compared to a given expression (in black). One can observe the articulatory tendencies induced by expressive constraints, through their acoustic byproducts.

Not surprisingly, given preceding observations, the larger divergences from the Neutral setting is observed along the F_1 dimension. Expressions with higher intensity/pitch (Fear, Anger, Surprise) have their vocalic triangle shifted towards higher F_1 values. Conversely, expressions with lower intensity/pitch (Sadness and Sensuality) show a tendency towards lower F_1—but this lowering of F_1 is constrained by the vowels' degree of opening (/a/ are more affected than /e/, while there is no effect on /i/ and /u/ along F_1). These changes in F_1 may be related to the degree of jaw opening movements [26], which would be correlated to F_1, with a floor (or palate) effect for high vowels (as the vocal tract cannot be completely closed while producing vowels). These relations are also supported by the description of the general tension setting by Laver [48], which is related to wider jaw movements.

Expressions also affect the F_2 and F_3 measures—but these changes are even more dependent on vowels. Along with the F_2 dimension, the largest departures from Neutral positions are observed for the Sensual and Joyful voices. However, whereas the F_2 of most vowels is affected by the Sensual voice, the expression of Joy affects mostly the rounded vowels (/u, o, ɔ, ø, œ/), with the exception of /y/. On the F_3 dimension, increased values are observed for Sensuality and Surprise (for back vowels), but decreased values are observed for Joy, once again for rounded vowels but the /y/. Such patterns of formant changes in Joy and Sensual voices may be explained by a constrained smile, and a fronted articulation, respec-

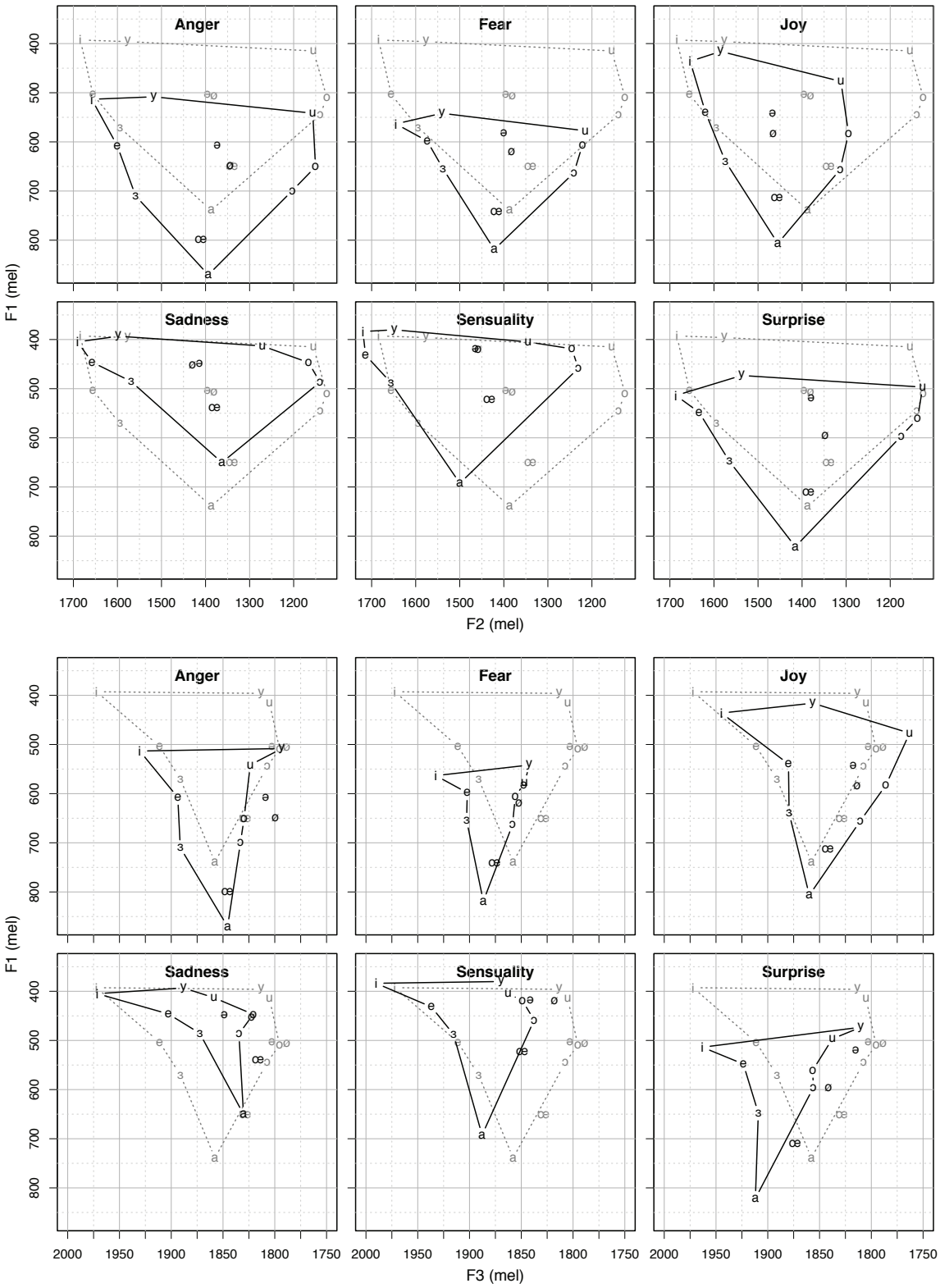

FIG. 4. Vocalic spaces obtained by plotting the median values of (F_1,F_2) (top) and (F_1,F_3) (bottom) of each vowel class in each expression.

tively. Constrained smiles (corresponding to a spread lip ar-

ticulatory gesture) can be heard [68], and smiles are typical of positive expressions (if not restricted to them): a constrained smile freezes labial articulation, thus diminishing the protrusion/labialization of rounded vowels. The effect of labialization on formants depends on the vowel constriction point, with back vowels differing from anterior ones [67]. Stevens [67] predicts an F_2F_3 peak for a labialized /i/ (p. 291f—a labialized /i/ is an /y/) that corresponds to an important F_3 and a (comparatively small) F_2 lowering. The constrained smile in the Joy expression of this corpus, by reducing the degree of labialization of the /y/, leads to an unchanged F_2 and a raised F_3, compared to the more rounded Neutral /y/. On the contrary, Stevens [67] predicts lowered F_2 for rounded back vowels (fig. 6.22, p. 293): the smile of Joy has thus the opposite effect of not lowering F_2 for these vowels (hence the higher F_2 in Joy compared to Neutral speech)—and to lower their third formant. In the case of Sensual voice, formant changes (raised F_2 and F_3) mostly affect back vowels—as if the speaker uses a fronted articulation.

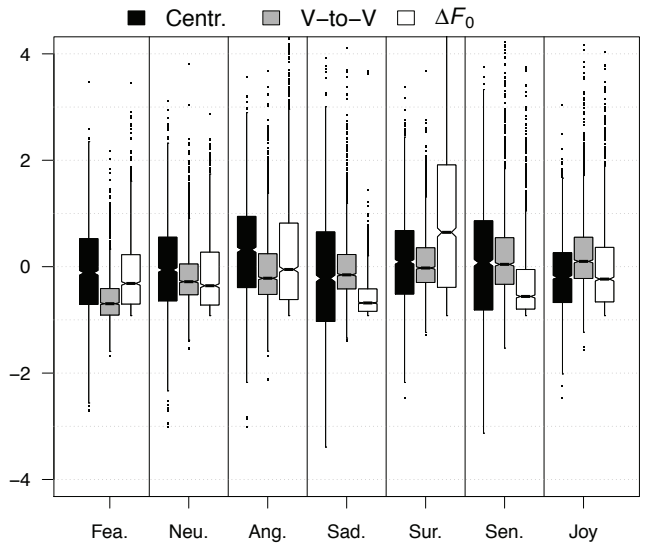

FIG. 5. Box plots of *Centr.* (black), *V-to-V* (light grey), and ΔF_0 (white) values expressed in z-scores, for each expression ranked in increasing order of duration (*V-to-V*).

E. Dynamics

Another effect of expressive voices is observed on vowel distributions: the surface enclosed by the polygons linking outer vowels shrinks for some expressions (Fear, Joy, Sadness), indicating a tendency to hypoarticulated speech (as compared to vowels spread in the Neutral speech). This effect is linked to the centralization measure (figure 5). Among the seven expressions, the observed hypoarticulation may be explained by several factors: by smaller vocal tract opening in Sadness (reduced jaw movements), by the constrained lip aperture in Joy, or because of a faster elocution. Faster speech (as compared with Neutral) is typically observed here for Fear (see the *V-to-V* measure in figure 5), which is the only expressive situation exhibiting shorter durations than the Neutral base.

F_0 differences from vowel to vowel give hints about the importance of pitch changes during speech production in different styles (see the ΔF_0 measure in figure 5). The most extreme variations of this measure are observed for Surprise, with a median above 4 ST. These extreme variations correspond to the large modulations of pitch reported for this expression in French [29]. A ΔF_0 above 2 ST is also observed for the expression of Anger [see the large standard deviation reported for F_0 by 34]. On the contrary, both Sad and Sensual voices show few modulations, with a median ΔF_0 below 1 ST.

F. Summary

From this description of the variations of each parameter across expressions, a distinctive description of these seven expressive conditions could be drawn. The Neutral, declarative, condition serve as a reference to describe the "expressive" ones. Two expressive conditions tend to show lower arousal, while four show higher arousal than the Neutral expression; this asymmetry may explain the positions of some acoustic cues in the Neutral condition that do not fit on the mean of observed values. The Neutral set was performed with intensity and F_0 (and their F_1 and $H_1 - A_3$ correlates) slightly below (above for $H_1 - A_3$) the median of observed values. This also applies to F_0 dynamic. On the contrary, Neutral voice shows high harmonicity, with the highest observed HNRs; it shows short duration and low levels of centralization (note that the context of the recordings may involve a clear articulation).

Compared to the Neutral base, the Sadness and Sensuality conditions are characterized by lower means of intensity, F_0, F_1 (especially for open vowels), and high spectral tilt. Sensuality is singularized by its high levels of aperiodicities (low HNRs due to breathy phonation), which contrast with the modal voice of Sadness. Sensuality is also produced with a fronted articulation, marked by higher F_2 for back vowels. Sadness is characterized by its high level of vocalic reduction (hypoarticulation linked to the lowest *Centr.* measure), and its flat F_0 contour (lowest ΔF_0).

On the other side of the arousal range, the expressions of Anger, Fear, Joy, and Surprise are characterized by high energy levels—with Anger exhibiting the highest, followed by Fear, Surprise, and Joy. These expressions also show higher F_0 levels than the Neutral condition. Comparatively to its measured energy, Fear has high F_0, higher than Anger—as in the case of Surprise when compared to Joy. Fear and Surprise are thus pro-

duced with a rise of pitch greater than the changes that may be explained by an increased vocal effort. Levels of $H_1 - A_3$ follow (negatively) intensity changes, except for Joy, which shows a rise in high-spectrum energy that may be explained by the increased energy in the third formant linked to an audible frozen smile. This smile also explains the increased F_2 measures in Joy for back rounded vowels and the decrease of F_3 for the /y/ vowel. These constraints on articulation may explain the centralization observed in the vocalic triangle of Joy, which is also characterized by an increased rate of melodic variations (higher ΔF_0). Fear, and to a lesser extend Anger, are produced with the highest levels of irregularities in the measured F_0 (lowering of *PAP* not explained by additional noise), which may be linked to high muscular tension—Fear being also produced with the fastest rate (which may account for some hypoarticulation). While Anger has the broadest formantic space, leading to the interpretation of wider jaw opening [26] and pharyngeal expansion [65]. Finally, Surprise is the expression having the highest ΔF_0, which may partly explain the structural noise observed in the *PAP* measure for this expression.

IV. ACOUSTIC DIMENSIONS

In the preceding section, expressive variation has been analyzed using a set of acoustic parameters, according to sets of voice quality settings. Descriptions of the expressions have been sketched. In this section, intrinsic dimensions of the acoustic space are investigated through an inferential statistical analysis.

A. Discriminative power of acoustic parameters

A Linear Discriminant Analysis (LDA) was run on the (standardized) acoustic measures, using R DiscriMiner library [64] to find the discriminatory power of parameters (alone or in combination) separating the proposed set of expressions [72]. Since the procedure does not accept missing values, the analysis is based on a restricted set (see table I for details). First, a preliminary descriptive analysis was run to measure the discriminatory power of each acoustic parameter, regarding the expressions in this corpus. Results are presented in table II. All parameters have a significant discriminatory power ($p < 0.01$), even when their contributions account to only a small amount of the variance. The intensity and F_0 measures carry the greatest share of information with canonical correlations above 0.7, followed by their covariates, $H_1 - A_3$ and F_1 (which explain a smaller part of the variance). Measures of noise (primarily *PAP*), ΔF_0, and duration follow. Other measures did not show an important global discriminatory power, but may still be of interest to discriminate specific cases.

An LDA procedure was then run using a cross-validation procedure (based on 10 groups) to evaluate the accuracy of the model in classifying individual vowels into these seven expressive categories, on the basis of the given acoustic measures. The resulting confusion matrix is presented in table III. Globally, the model achieves a 78% classification score. Note that this classification measure is obtained on vowels—not on full sentences.

TABLE II. Predictive power of the acoustic parameters, measured as canonical correlations through an LDA.

Parameter	INT_A	F_0	$H_1 - A_3$	F_1	*PAP*	*NOISE*
can. correl.	0.85	0.76	0.41	0.38	0.34	0.26
Parameter	ΔF_0	*V-to-V*	F_2	F_3	*Centr.*	
can. correl.	0.18	0.14	0.06	0.06	0.03	

TABLE III. Percentage of vowels from each expressive base (rows) classified as one of the 7 expressive categories (columns) by the LDA model built on the acoustic parameters.

Original	*Predicted*						
	Neut.	Ang.	Fear	Joy	Sad.	Sens.	Surp.
Neutral	81	0	0	6	4	4	4
Anger	2	84	8	5	0	0	1
Fear	0	6	82	4	0	0	7
Joy	8	3	4	71	0	0	14
Sadness	5	0	0	0	90	6	0
Sensuality	9	0	0	0	22	68	0
Surprise	6	2	10	17	0	0	65

The best classification rates are achieved for Fear, Sadness, and Neutral (over 80%), while others expressions show rates between 60% and 70%. Three expressions exhibit consistent confusions—over chance level (14%): Sensuality is mistaken for Sadness, while Surprise and Joy show mutual confusions.

B. Discriminant functions and acoustic dimensions

The combinations of base acoustic parameters forming the discriminant functions (*DF*) highlight the use of parameters to discriminate expressions (table IV presents the correlations between the variables and the *DF* of the descriptive model). The first *DF* (which explains more than 80% of the variance) opposes vowels that have low intensity, F_0, and F_1; these low values are related to increased spectral tilt. This *DF* is reminiscent of the dimensions of effort and tension described in the preceding section. The second *DF* (8% of the variance) is linked to both measures of noise: it separates voice produced with high levels of noise (and here typically Sensuality, with additive noise) from the others. The third *DF* (7% of the variance) selects vowels where F_0 covaries with $H_1 - A_3$,

TABLE IV. Correlation between the acoustic variables and the six *DF*; percentage of variance explained by each *DF*.

	DF_1	DF_2	DF_3	DF_4	DF_5	DF_6
F_0	-0.88	-0.07	-0.45	-0.07	-0.02	0.08
ΔF_0	-0.30	-0.09	-0.09	-0.29	0.72	-0.24
INT_A	-0.98	-0.03	0.05	0.13	-0.00	0.07
PAP	0.04	-0.87	0.31	0.24	0.06	-0.05
NOISE	-0.18	-0.76	0.21	0.15	-0.05	0.19
$H_1 - A_3$	0.58	0.00	-0.40	0.43	0.38	0.17
V-to-V	0.12	0.03	0.29	-0.64	0.28	0.49
F_1	-0.66	-0.08	-0.05	0.10	-0.07	0.29
F_2	0.13	0.24	0.03	-0.36	-0.16	-0.49
F_3	0.03	0.20	-0.27	-0.15	0.26	-0.02
Centr.	-0.10	0.08	0.05	0.09	0.31	0.33
Prct.	80.8	7.9	6.7	2.7	1.8	0.1

without changes in intensity—thus a voice produced with pitch changes that are not induced by intensity. This function typically separates Fear and Surprise from Joy and Anger productions. The first three eigenvalues explain about 95% of the variance; the remaining three *DF* thus discriminate particular cases. The fourth *DF* is linked to lengthening, and thus separates the expressions of Joy, Surprise and Sensuality (which show the most pronounced lengthening) from Fear and Neutral (spoken faster, especially Fear). The fifth *DF* is correlated to ΔF_0, which is, as already mentioned, mostly typical of Surprise [29]. The last *DF* links centralization and a lowering in F_2, which corresponds to supraglottal change in the articulation that have been related to either smile or an anteriorization in articulation (see section III).

C. Frequency code and effort code

F_0 is certainly a major cue to expressive voice variations, but fully understanding its production requires taking into account vocal effort, which is intimately linked to vibration mechanisms [39, 70]. As these analyses show, it is the combination of both intensity and F_0 that gives the most comprehensive distribution of affects. The possibility for the speaker to control F_0 independently of intensity is illustrated by the LDA third *DF*. Thus, the speaker exerts a control either on the intensity (via an increased effort) or on F_0. This result supports the Frequency Code [59] and the Effort Code [35]: concepts that discuss different types of F_0 changes. The Effort Code explicitly links F_0 excursions to the exertion by the speaker of a greater effort. In our data, F_0 changes spanning from Sensuality up to Anger are typically induced by an increased effort—and linked to a higher arousal in Scherer [65] terms [see also 34]. The explicit control of F_0 only (which is typically the case of Fear opposed to Joy in this corpus) would in that view be related to the Frequency Code, and the observed changes support its predictions on emotional expressions [an F_0 rise linked to a lack of control over the situation: 59]. In such cases, the measures of F_0 and $H_1 - A_3$ bring the most interesting cues, knowing the arousal level [34].

D. Role of back cavity

A covariation of INT_A and F_0 with F_1 is observed. F_1 is related to the resonance of the back cavity of the vocal tract—at least for French oral vowels [4]. Changes in overall muscular tension modify the vertical position of the larynx and the muscular contraction of the vocal apparatus; Laver [48] linked it to higher F_0. Increased degree of articulation is linked to wider jaw opening [26]. Such changes affect the length and width of the back cavity—and thus the first formant value. Given the observed changes in F_1 for the expressions exhibiting a high F_0, both phenomena (raised and constricted larynx, lowered jaw) are likely to play a role in F_1 changes. A wide jaw opening is expected for Anger and Surprise, and affects the whole vocalic space along F_1. The expression of Fear is produced here with a very fast rhythm, which is not compatible with large movements of the jaw: higher F_1 are observed in this case, but with a reduced span between high and low vowels. Smaller changes in F_1 across the vocalic space may be related to reduced articulatory movements, and higher overall F_1 values may be linked to high tension, as well as a wider mean opening of the jaw.

If F_1 changes are mostly a subproduct of arousal, the main expressive space is thus constructed by the speaker on two aspects of voice production: vocal effort and pitch control. Distinctions between expressions confused on this plane (i.e., the expressions with low or high arousal) are made thanks to the use of others, secondary, acoustic cues.

E. Noise dimension

One of the important secondary cues is the presence of noise in the voice. In this corpus, noise is mostly due to a breathy phonation and can be observed in the Sensual voice. Breathy phonation for Sensual voice follows the cliché of B. Bardot's *charming voice*, as described by Léon [50, p. 77–79]. The acoustic measures of noise allow a separation between Sensual and Sad voices that are otherwise similar in their F_0 and intensity values. Other types of noise may have been observed: typically a presence of vocal fry, which could have led to a different expressivity, but it is not observed in this data, thus we will not speculate [for definition or an example of its expressive use see, e.g., 31, 61]. Léon's description also cites a fronted articulation (for its symbolism of "little girl"); fronting that is observed in this corpus (increased F_2 for back vowels), together with a lowering of intensity. The

performance recorded in this corpus thus matches classical descriptions of Sensual voice in French females.

Both Fear and Anger also received high levels of structural aperiodicities that may be correlated with the high levels of muscular tension (at the glottis) necessary for such productions, whatever their otherwise different characteristics. This tension gives, as a byproduct, a more trembling voice that may convey a negative valence [cf. a similar conclusion for high arousal in 34]. Note that the amount of noise in Anger is lower than the one found in Fear; but noises in both voices still share the same explanation of a greater muscular tension. The differences between Fear and Surprise are linked to their intonation patterns: important inter-syllabic changes are performed in Surprise, while the intonation of Fear is reaching a ceiling and presents a flat contour. The expression of Joy, with an intensity level comparable to Surprise, has a much lower spectral slope (close to the one observed for Fear). This shallow slope seems unlikely to be due to vocal effort alone, as intensity is much lower for Joy than Fear. The median A_3 (critical for the $H_1 - A3$ measure), is given in table V. Joy is performed with the highest A_3 values. One may expect that the smiling gesture, typical of this expression, may enhance A_3. Such a high energy in the high spectrum is coherent with the description of naturally occurring smiling speech, perceived as high-pitched [23]. The four high-pitched expressions are also, to some extent, distinguished by the changes they induced in the formants—in F_1 for the most aroused, and also F_2 and F_3 in the case of Joy.

TABLE V. Median values of A_3 (in dB) for each expression.

Neutral	Anger	Fear	Joy	Sadness	Sensuality	Surprise
18	28	17	33	3	10	26

F. Supraglottal changes

Secondary cues such as structural noise, melodic changes, and formantic values does not have such a prominent discrimination role as intensity and F_0 at a global level, notably because their distributions over expressions show larger overlaps. These overlaps arise from different reasons. It has been shown that the formantic changes linked to expressive variation are not systematic across vowels. They are linked to vowel categories (open or not, rounded or not) and changes in formants related to expressions may occur in opposite directions, depending on the vowel (e.g., opposite F_2 change for /y/ vs. /u/ in Joy, for the same lip spreading reason).

Measurements of vowels centralization, in addition to being difficult to obtain without a good knowledge of the speaker's voice, are not conclusive in this work. Other measures such as the vowel compact/diffuse and grave/acute dimensions [63] have been tested but, as linear combinations of the direct formant measures, they do not lead to different solutions (higher F_1 increases compactness, higher F_2 decreases it, raised formants increased the acute measurement).

Changes in the vocal tract shall still have an important impact on the perception of these expressions, but a comprehensive use needs a more complex statistical model.

V. DISCUSSION AND CONCLUSIONS

A. Paradigmatic acoustic discrimination & perception

The results do replicate previous descriptions in the literature [6, 34, 51, 71], and they also do follow predictions made by several theoretical models of speech production or affect [2, 35, 48, 59, 65, 67, 70]. We believe the acoustic parameters presented here adequately reflect the presented set of voice quality dimensions—the use of which will have to be replicated and extended across speakers, genders, languages, and along finer and more varied expressive variations.

This work is a quest for the main dimensions in paradigmatic changes for vowels in the expressive speech of a given speaker. The base dimensions are measurable acoustic parameters, for both the voice source and vocal tract components of the speech production model.

In a first part, expressions were analyzed in a top-down approach, using acoustic dimensions such as intensity and tension. In a second part, a bottom-up approach unveiled the hidden dimensions in the acoustic space through statistical analyses. Additionally, a statistical analysis allows assigning weights to the relative significance of intrinsic dimensions.

An overall score of 78% of good classifications is obtained by the LDA, on the basis of isolated vowels only; listeners presented with full sentences do achieve an 86% of good recognition [25]. This indicates that paradigmatic vocal quality analysis is a relevant approach since isolated vowels in sentences carry an important part of the expressive content.

Recognition of individual expressions also differs between the automatic classifier and human listeners. According to results presented in Evrard et al. [25], the expression of Fear in this corpus is the most difficult to recognize (with 68% of correct identification by listeners); on the contrary, it is one of the expression with the highest classification score by the LDA (82%). A reverse pattern is observed for Joy, well recognized by listeners (85%), but which receives a 71% classification score by the LDA model. Automatic and perceptual categorization processes also show differences in their confusions patterns. Joy is confused with Sensuality and Surprise by listeners, while its acoustic patterns are close to Surprise only.

The two tasks—perceptual identification and acoustic

classification—are obviously different, and such a comparison is limited. Perception involves a top-down information processing strategy; Joy and Sensuality share conceptual features (like a positive valence) that are not interpreted by such a bottom-up only model. There may be cues to valence in the acoustic signal [e.g., the acoustic characteristics of smile in these two cases, the presence of noise in Fear—see 34]. However, using this knowledge to link the Joy and Sensuality categories that are otherwise very different acoustically (regarding intensity and pitch), would require a much more complex (and realistic) statistical model than the one presented here. Listeners do also possess a detailed representation of human voices (having notably a full knowledge of the phonetic properties of sounds) that allows them to grasp changes in vowel quality in a holistic way. The presented LDA analysis lacks a comprehensive representation of the speaker's vocal ability (i.e., vowel categories are not part of the model). The model is also based on a single speaker and obviously cannot extrapolate to others—especially not to male voices.

Such differences underline the gap between acoustic measurements and high-level analysis of voice expressivity by human listeners. The perceptual results support the importance of supraglottic changes, and primarily smile. The presence of a smile may explain the perceptual confusion between Joy and Sensuality. On the other side, the LDA categorization, despite all its limitations linked to a single speaker corpus and a limited expression set, does show that a paradigmatic approach has merits. The statistical model did not aim at providing a classification of emotional expressions, but rather at (i) showing that sets of vocal qualities (compounds of voice quality settings found in an expressive voice) may be acoustically differentiated at a reasonable rate (78%)—thanks to paradigmatically measured acoustic parameters, and (ii) extracting the acoustic dimensions used to segregate such types of voice qualities.

B. Weighted dimensions and acoustic description

An interesting outcome of these results is the relative weights of acoustic dimensions. The first statistical dimension explains a substantial part of the variance in the acoustic space (more than 80%), being a combination of few parameters: INT_A, F_0, F_1, $H_1 - A_3$. This is in good agreement with Figure 1, where the seven expressions, which present seven types of compound voice qualities, show a clear organization in the (INT_A, F_0) plane. This result suggests that, for this corpus, the most important acoustic cues for characterizing expressivity in isolated vowels are given by intensity and melodic height. This is reminiscent of the arousal dimension of emotional models, and of the frequency code and effort code of communicative models.

The second statistical dimension corresponds to noise measurements (see Figure 3 for acoustic description). It is primarily beneficial on this data for the distinction between Sadness and Sensuality (for additional noises), and between Neutral and Anger, Fear or Surprise (for structural noises).

The third statistical dimension corresponds to changes in spectral tilt that are independent of intensity. Such changes are thus not related to vocal effort, but rather to vocal tract induced spectral changes or to intensity-independent variations of F_0. An important feature of vocal tract settings in this set of expressions is related to facial mimic, like smile (i.e., lips spreading that increases F_3 amplitude) present in Joy. Covariations of $H_1 - A_3$ and F_0 are typical of Fear and Surprise, with important F_0 changes that are not explained by effort.

Other dimensions are more difficult to interpret and are of lesser importance. In summary, the four dominant acoustic parameters are F_0, INT_A, PAP and $NOISE$, and $H_1 - A_3$. Other parameters play a role in the obtained discrimination results, but it seems difficult to interpret their specific contributions. It has to be noted that other types of voice quality variations exist, and combinations of them that could depend on other types of expressions, such as, e.g., strategic choices by speakers or speaker characteristics. Typically, there is no nasality settings in this corpus.

Let's also note that F_0 was measured on the EGG signal, which enhances its reliability; such a signal is not commonly available, but robust pitch detection algorithms are now widely available. Meanwhile, such algorithms may fail on some types of voice quality, as it failed in 20% of cases for the breathy sensual voice. Harsh or creaky voices may also have proven challenging. The very fact that F_0 detection fails, or vocalic segment are found unvoiced, may also be used as a voice quality parameter (i.e., using the percentage of unvoiced vowels or unvoiced frames as a measure of non-modal voices).

C. Conclusion

The results obtained in this work are based on the analysis of only one female speaker, in one language, with a restricted set of acted expressions. In this case, it seems that the paradigmatic variation in vowels is consistent across expressions. More work is needed to take into account cross-speaker variation, across languages and genders, and for naturally occurring expressions (which shall exhibit more subtle change patterns).

The results obtained are in good agreement with earlier descriptions of expressive voice variations, made on several speakers, and based on different sets of long-term acoustic measurements [6, 34]. It also supports the predictions of various models and descriptions of speech and voice expressivity, by explanatory measures for acoustic changes along vocal effort [35], pitch [59], emotions [65], or smiling speech [23, 68]. The covariation of the two main acoustic cues (F_0 and intensity) follows the descriptions of the literature [51, 70, 71]. Those outcomes, along

with an efficient descriptive model of acoustic change obtained from the measures used in this work, reinforce the strength of this description.

These results demonstrate that paradigmatic variations of vowels in expressive speech are highly consistent. This approach seems more detailed and accurate than long-term average analysis [e.g., 34]. It is a local short-term method that is complementary to prosodic or syntagmatic voice quality analyses.

ACKNOWLEDGMENTS

Part of this work has been conducted in the framework of the ADN T-R project funded by Région Ile-de-France (convention No 111012235). The authors would like to express their gratitude to the two anonymous reviewers for their constructive comments.

[1] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," IEEE Transactions on Speech and Audio Processing **6**(1), 12-23 (1998).

[2] C. d'Alessandro, "Voice source parameters and prosodic analysis". In Sudhoff, S., Lenertova, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., and Schlieer, J. (Eds.), *Methods in empirical prosody research* (Berlin : Walter de Gruyter, 2006), 6387.

[3] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," Speech Communication **18**(2), 131–138 (1996).

[4] L. Apostol, P. Perrier, and G. Bailly, "A model of acoustic interspeaker variability based on the concept of formantcavity affiliation," The Journal of the Acoustical Society of America **115**(1), 337–351 (2004).

[5] J. A. Bachorowski, and M. J. Owren, "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," The Journal of the Acoustical Society of America **106**(2), 1054–1063, 1999.

[6] R. Banse, and K. R. Scherer, "Acoustic profiles in vocal emotion expression," Journal of personality and social psychology **70**(3), 614–636 (1996).

[7] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," Emotion **12**(5), 1161–1179 (2012).

[8] P. Barbosa, "From syntax to acoustic duration: A dynamical model of speech rhythm production," Speech Communication **49**(9), 725–742 (2007).

[9] P. Boersma, and D. Weenink, "Praat : doing phonetics by computer" (version 5.3.32)[computer program]. retrieved October 17, 2012.

[10] T. Brosch, G., Pourtois, and D. Sander, "The perception and categorisation of emotional stimuli: A review," Cognition and Emotion **24**(3), 377–400 (2010).

[11] E. H. Buder, "Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990", in Voice Quality Measurements, R.D. kent and M.J. Ball, Editors, Singular Publishing Group, 2000, 119–244.

[12] Z. Camargo, and S. Madureira, "Voice quality analysis from a phonetic perspective: Voice Profile Analysis Scheme (VPAS) Profile for Brazilian Portuguese," In Proc. Speech Prosody 2008, Campinas, Brazil, 57–60 (2008).

[13] N. Campbell, "Automatic detection of prosodic boundaries in speech," Speech communication **13**(3), 343–354 (1993).

[14] N. Campbell, and P. Mokhtari, "Voice quality the 4th prosodic dimension," Proc. of International Congress of Phonetic Sciences, Barcelona, Spain, 2417–2420 (2003).

[15] N. Campbell, "Getting to the heart of the matter; Speech as the expression of affect. rather than just text or language," Language Resources and Evaluation **39**(1), 109–118 (2005).

[16] D. G. Childers, and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," The Journal of the Acoustical Society of America **90**(5), 2394–2410 (1991).

[17] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modelling using neural networks," Neural networks **18**(4), 371–388 (2005).

[18] A. de Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," The Journal of the Acoustical Society of America **111**(4), 1917–1930 (2002).

[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," In Proc. ICASSP, Florence, Italy, 960–964 (2014).

[20] B. Doval, C. d'Alessandro and N. Henrich, "The spectrum of glottal flow models," Acta acustica united with acustica **92**(6), 1026–1046 (2006).

[21] P. Ekman, and D. Cordaro, "What is meant by calling emotions basic," Emotion Review **3**(4), 364–370 (2011).

[22] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition **44**(3), 572–587 (2011).

[23] C. Émond, "Les corrélats prosodiques et fonctionnels de la parole perçue souriante en français québécois spontané (Prosodic and functional correlates of perceived smiling speech in spontaneous Quebec French speech)," Université du Québec à Montréal (2013).

[24] M. Evrard, "Synthèse de la parole à partir du texte : Des phonostyles au contrôle gestuel pour la synthèse paramétrique statistique (Text-to-Speech synthesis: from phonostyles to the gestural control of statistical parametric synthesis)," Ph.D. Thesis, Paris-Sud University, 2015.

[25] M. Evrard, S. Delalez, C. d'Alessandro, and A. Rilliard, "Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis," In Proc. 16th INTERSPEECH 2015, Dresden, Germany, 3370–3374 (2015).

[26] D. Erickson, A. Suemitsu, Y. Shibuya, and M. Tiede, "Metrical structure and production of English rhythm," Phonetica **69**, 180-190 (2012).

[27] G. Fant, *Acoustic Theory of Speech Production* (Haag: Mouton, 1960).

[28] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," Speech, Music and Hearing Quarterly Progress and Status Report (STL-QPSR) **36**(2-3), 119–156 (1995).

[29] I. Fónagy, "Languages within language: An evolutive approach," Amsterdam: John Benjamins Publishing (2001).

[30] C. Gendrot, and M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German," In Proc. Interspeech, Lisbon, Portugal, 2453-2456 (2005).

[31] B. R. Gerratt, and J. Kreiman, "Toward a taxonomy of nonmodal phonation," Journal of Phonetics **29**(4), 365–381 (2001a).

[32] B. R. Gerratt, and J. Kreiman, "Measuring vocal quality with speech synthesis," The Journal of the Acoustical Society of America **110**(5), 2560–2566 (2001b).

[33] C. Gobl, and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," Speech communication **40**(1), 189–212 (2003).

[34] M. Goudbeek, and K. R. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," The Journal of the Acoustical Society of America **128**(3), 1322–1336 (2010).

[35] C. Gussenhoven, *The phonology of tone and intonation* (Cambridge: Cambridge University Press, 2004).

[36] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," The Journal of the Acoustical Society of America **101**(1), 466–481 (1997).

[37] A. Hassan, and R. I. Damper, "Classification of emotional speech using 3DEC hierarchical classifier," Speech Communication, **54**(7), 903–916 (2012).

[38] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," The Journal of the Acoustical Society of America **115**(3), 1321–1332 (2004).

[39] N. Henrich, C. dAlessandro, B. Doval, and M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," The Journal of the Acoustical Society of America **117**(3), 1417–1430 (2005).

[40] S. Hurley, "The shared circuit model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading," Behavioural and Brain Sciences **31**, 1–58 (2008).

[41] IEC 61672-1, *Electroacoustics - Sound level meters - Part 1: Specifications* (Geneva: International Electrotechnical Commission, 2013).

[42] A. I. Iliev, M. S., Scordilis, J. P., Papa, and A. X. Falco, "Spoken emotion recognition through optimum-path forest classification using glottal features," Computer Speech & Language, **24**(3), 445–460 (2010).

[43] P. J. Jackson, and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," IEEE Transactions on Speech and Audio Processing **9**(7), 713–726 (2001).

[44] G. Kochanski, E. Grabe, J. Coleman and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," The Journal of the Acoustical Society of America **118**(2), 1038–1054 (2005).

[45] G. Kochanski, "Prosody beyond fundamental frequency". In Sudhoff, S., Lenertova, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., and Schlieer, J. (Eds.), *Methods in empirical prosody research* (Berlin : Walter de Gruyter, 2006), 89122.

[46] J. Kreiman, M. Garellek, G. Chen, A. Alwan, and B. R. Gerratt, "Perceptual evaluation of voice source models," The Journal of the Acoustical Society of America **138**(1), 1–10 (2015).

[47] O.W. Kwon, K., Chan, J., Hao, and T. W. Lee, "Emotion recognition by speech signals," In Proc. 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, 125-128 (2003).

[48] J. Laver, *The phonetic description of voice quality* (Cambridge: Cambridge University Press, 1980).

[49] J. Laver, "Phonetic evaluation of voice quality". In Kent, R. D. and Ball, M.J. (Eds.), *Voice quality measurement* (San Diego: Singular Publishing, 2000), 37-48.

[50] P. Léon, *Précis de phonostylistique: parole et expressivité (Manual of phonostylistics: speech and expressivity)* (Paris: Nathan, 1993).

[51] J. S., Liénard, and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," The Journal of the Acoustical Society of America **106**(1), 411–422 (1999).

[52] J. S. Liénard, and C. Barras, "Fine-grain voice strength estimation from vowel spectral cues," In Proc. INTERSPEECH, Lyon, France, 128-132 (2013).

[53] J. Mackenzie-Beck, "Perceptual analysis of voice quality: the place of vocal profile analysis". In Hardcastle, W.J. and Mackenzie-Beck, J. (Eds.), *A figure of speech: a festschrift for John Laver* (Mahwah: Lawrence Erlbaum, 2005), 285-322.

[54] H. Mixdorff, and H. R. Pfitzinger, "Analysing fundamental frequency contours and local speech rate in map task dialogs," Speech Communication **46**(3), 310–325 (2005).

[55] C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet, "Discriminating expressive speech styles by voice quality parametrization," In Proc. 16th International Congress of Phonetic Sciences, Saarbrcken, Germany, 2081–2084 (2007).

[56] J. A. de Moraes, and A. Rilliard, "Illocution, attitudes and prosody: A multimodal analysis," In Raso, T. and Ribeiro De Mello, H. (Eds.), *Spoken Corpora and Linguistic Studies* (Amsterdam: John Benjamins Publishing Company, 2014), 233–270.

[57] I. R. Murray, and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," The Journal of the Acoustical Society of America, **93**(2), 1097–1108 (1993).

[58] F. Nolan, "Intonational equivalence: an experimental

evaluation of pitch scales," In Proc. of the 15th International Congress of Phonetic Sciences, Barcelona, Spain (2003).

[59] J. J. Ohala, "The frequency code underlies the sound symbolic use of voice pitch". In Hinton, L., Nichols, J. and Ohala, J. J. (Eds.), *Sound symbolism* (Cambridge: Cambridge University Press, 1994), 325-347.

[60] V. Petrushin, "Emotion in speech: Recognition and application to call centers," In Proc. of Artificial Neural Networks Intelligence Engineering, St Louis, USA, 7–10 (1999).

[61] R. K. Podesva, "Phonation type as a stylistic variable: The use of falsetto in constructing a persona," Journal of sociolinguistics **11**(4), 478–504 (2007).

[62] M. Rossi, "Interactions of intensity glides and frequency glissandos," Language and speech **21**(4), 384–396 (1978).

[63] S. Rvachew, K. Mattock, L. Polka and L. Ménard, "Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian French," The Journal of the Acoustical Society of America **120**(4), 2250–2259 (2006).

[64] G. Sanchez, "DiscriMiner: Tools of the Trade for Discriminant Analysis". R package version 0.1-29, http://CRAN.R-project.org/package=DiscriMiner (2013).

[65] K. R. Scherer, "Emotions are emergent processes: they require a dynamic computational architecture," Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1535), 3459–3474 (2009).

[66] K. R. Scherer, "The nature and dynamics of relevance and valence appraisals: Theoretical advances and recent evidence," Emotion Review **5**(2), 150–162 (2013).

[67] K. N. Stevens, "Acoustic phonetics," Cambridge, MA: MIT Press (1998).

[68] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," Perception & psychophysics **27**(1), 24–27 (1980).

[69] E. Terhardt, G. Stoll, and M. Seewan, "Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions," The Journal of the Acoustical Society of America **71**(3), 671–678 (1982).

[70] I. R. Titze, and J. Sundberg, "Vocal intensity in speakers and singers," The Journal of the Acoustical Society of America **91**(5), 2936–2946 (1992).

[71] H. Traunmüller, and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," The Journal of the Acoustical Society of America **107**(6), 3438–3451, (2000).

[72] W.N. Venables, and B.D. Ripley, *Modern Applied Statistics with S* (Fourth edition. Berlin: Springer, 2002).

[73] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," Journal of Phonetics **48**, 1–12, (2015).

[74] Y. Wang, and L. Guan, "An investigation of speech-based human emotion recognition," In Proc. 6th IEEE Workshop on Multimedia Signal Processing, Siena, Italy, 15–18 (2004).

[75] A. Wichmann, "The attitudinal effects of prosody, and how they relate to emotion," Proc. ISCA Tutorial and Research Workshop on Speech and Emotion, Newcastle, U.K. (2000).

[76] A. Wierzbicka, "Defining emotion concepts," Cognitive science **16**(4), 539-581 (1992).

[77] M. Yik, J. A. Russell, and J. H. Steiger, "A 12-point circumplex structure of core affect," Emotion **11**(4), 705–731, (2011).