Acoustic and perceptual profiles of american english social affective expressions

Donna Erickson, Albert Rilliard, Ela Thurgood, João Antônio de Moraes, Takaaki Shochi

JoSS Journal of Speech Sciences

DOI: 10.20396/joss.v13i00.20015

ACOUSTIC AND PERCEPTUAL PROFILES OF AMERICAN ENGLISH SOCIAL AFFECTIVE EXPRESSIONS: A CASE STUDY

ERICKSON, Donna[1]*
RILLIARD, Albert[2]
THURGOOD, Ela[3]
DE MORAES, João Antônio[4,5]
SHOCHI, Takaaki[6]
1Haskins Labs, New Haven, CT, USA – ORCID: https://orcid.org/0000-0003-2283-9432
2Université Paris Saclay, CNRS, LISN, Orsay, France
3California State University, Chico, CA, USA
4Universidade Federal do Rio de Janeiro, RJ, Brazil
5CNPq, Brazil
6Université Bordeaux 3, Bordeaux, France

Abstract: *This pilot study reports on acoustic and perceptual profiles of two American female speakers' productions of six American English social affective expressions: Authority, Declaration, Irritation, Sincerity, Uncertainty and walking on eggs (WOEG) as spoken in the linguistic sentence frame, Mary was dancing. The acoustic profile describes the prosodic characteristics (F0, intensity, duration, voice quality and tonal targets) of the utterances as a whole, as well as the voice quality characteristics of the nuclear stress syllable in the utterances. The perceptual profiles describe (1) listeners' 3-dimensional VAD emotional ratings, i.e., Valence, Arousal, and Dominance, of the utterances and (2) listeners' auditory impressions of the nuclear stress syllable. Multifactorial Analyses (MFA) were applied to examine the relation between the prosodic characteristics and the VAD scales, and also the relationship between voice quality measurements on the nuclear stress vowel and auditory perceptions. The prosodic MFA results indicate that for these two American English speakers, a soft / noisy voice, with weak harmonics and irregular rhythm with pauses and hesitations, as in the expressions of Uncertainty and WOEG, is perceived by listeners as accommodating (not Dominant) and not positive (negative Valence). Loud, tense voices with energy in the upper frequencies, as in the expression of Irritation, are perceived as Aroused. Expressions of Authority, Declaration, and Sincerity tend to have comparatively regular rhythm and relatively flat intonation. The MFA analysis of voice quality measurements and auditory perceptions suggests that Normalized Amplitude Quotient (NAQ) may indeed be a good estimate for tense voice due to glottal closing behavior, Cepstral Peak Prominence (CPP), a good estimation for strong non-noisy harmonics, Peak Slope, a good estimate of spectral related tense voice, and Hammarberg Index, for distribution of spectral energy, i.e., strong or weak energy in the upper frequencies. But these measures do not completely account for the auditory judgments of voice quality. An interpretation of the acoustic and perceptual profiles of the social affective expressions is discussed in terms of theoretical codes (Frequency Code and Effort Code) proposed for explaining symbolic relations between vocal characteristics and pragmatic affective processing in spoken human communication.*
Keywords: Social affective expressions; Prosody; Voice quality; Valence Activation Dominance ratings; Auditory perceptions, Acoustic measurements; Social codes.

*Corresponding author: ericksondonna2000@gmail.com

1 Introduction

Over the past several decades, the acoustic and perceptual characteristics of emotional speech have been studied (see e.g. Erickson 2005 for a review). Studies of various emotions, such as anger, happiness, sadness, etc., have reported that, for instance, anger as well as happiness are characterized by increased loudness and heightened F0, sadness by decreased loudness, lowered F0, and longer duration, etc. (e.g., Scherer et al. 1991; Williams and Stevens 1972). Note, however, that "emotion" is complex in a number of ways, including labeling of emotions-- e.g., hot anger is different from cold anger; the former is characterized by high F0, e.g., Bänziger & Scherer (2005), the latter by low F0, among other things, e.g, Johnstone & Scherer (2000). Also, active grieving, characterized by high intensity and high F0, is different from sadness, characterized by low intensity and low F0 (e.g., Erickson et al. 2006). Moreover, acted emotions have different acoustic characteristics from spontaneous emotions (e.g., Erickson 2005; Erickson et al. 2006; Jürgens et al. 2011).

An approach to examining acoustic and perceptual characteristics of emotion is to focus on social affective expressions (SAE). Emotional expressions are not examined per se, but rather, the focus is on how individuals in a language-specific culture interact with each other on a daily basis. These expressions have their roots in underlying physiological emotions (e.g., Damasio 1998), but are tailored to be accepted within the cultural settings of the speakers (e.g. Rilliard et al. (2017). Some of the early studies of social affective expressions include e.g., Rilliard et al. 2009, 2012, 2013, 2014, 2017; Shochi et al. 2009, 2023; de Moraes et al. 2012. These studies tended to focus on the perceptual aspects of social affective expressions, reporting that some expressions were fairly universally recognized while others, such as e.g. irony and seduction, were not. Niebuhr (2014), reporting on irony in German, observed that ironic expressions tended to have lowered intensity, longer duration, and lower and flatter F0. Similar findings were reported by Moraes et al. (2012) showing that irony has lengthened duration on stressed syllables. Work by Erickson et al. (2002) suggests that acoustic characteristics of irony/sarcasm may be culturally / language dependent: Sarcasm is not frequently used by Japanese (personal experience of the first author, https://japanintercultural.com/free-resources/articles/stay-away-from-irony-with-the-japanese/), but when it is, it is characterized by final high rising F0 (Yanagida 2002) while English sarcasm often is produced with a final drop in F0 (Erickson et al. 2002). Work by González-Fuente et al. (2015) report "that visual information produced after ironic sentences is a key factor in the identification of the speaker's ironic intent" (p. 26); Mauchand et al. (2020) reported that depending on the linguistic content, an ironic expression can be perceived as positive or negative. Seduction in Japanese tends to be produced with a higher F0, while in English, French, or Brazilian Portuguese, it has a lower F0 (e.g., Rilliard et al. 2017; Erickson et al. 2023). For more discussion of what makes a voice sound sexy or attractive, see Weiss et al. (2021).

Politeness expressions are also included in SAE, and again language-cultural differences abound. Loveday (1981) observes that both male and female English speakers raise their f0 for expressing politeness, while in Japan, female speakers raise their F0 even more, while the F0 of male speakers doesn't change that much. Nadeu and Prieto (2011), looking at polite speech in Catalan, observed that increasing F0 alone does not necessarily increase the degree of perceived politeness; rather contextual and gestural, specifically facial information, is important. Similar findings about the importance of head and face movements for recognizing politeness/impoliteness are reported by Shochi et al. (2023). Rilliard et al. (2012) also discussed the acquisition of polite expressions by Japanese children, showing that by the time they are 9 or 10 years old, they have adult-like recognition of politeness categories. A comprehensive study by Rilliard et al. (2014) showed that non-Japanese listeners (French, American and Brazilian

Portuguese) have about a 60% global recognition rate of Japanese polite/non-polite speech. Caballero et al. (2018) report that compared with politeness, impolite utterances have lower F0 and slower speech rate. See also Culpepper et al. (2003) for a discussion of polite and impolite expressions.

Work by Székely et al. (2017) reports that uncertain expressions tend to be softer, with elongation of function words and filled pauses. Ward & Hirschberg (1985) suggest that uncertainty is conveyed by a fall-rise intonation contour. Complaining speech by French speakers has been studied by Mauchand and Pell (2021): increased F0 and HNR, together with less shimmer and jitter were found to be acoustic characteristics of a complaining voice.

This pilot study presented here pursues the earlier work by Rilliard and colleagues to examine fine details of how our voices change in daily conversations and how these changes affect the listeners' perceptual assessments of socially induced communicative functions. In this case study, we examine data from this existing speech corpus of highly controlled, high-quality recordings of utterances from expressive dialogues of two speakers with very similar socio-cultural backgrounds that also share comparable communicative performances, as cross-checked by perception tests (see e.g. Rilliard et al. 2013). Details of the speech corpus are explained in the Methods section below. In our previous studies, we examined perceptual groupings of a subset of SAE produced and well-recognized by U.S. English speakers. In Rilliard et al.(2017), we reported on the interaction between normalized values of F0 and intensity of 8 L1 speakers (5 females, 3 males) of U.S. American speakers' productions of 11 attitudes that were well-perceived by U.S. listeners (N= 35). F0 and intensity characteristics were observed as follows: surprise was produced in a loud and high-pitched voice; Irritation with the loudest voice, more than 10 dB higher than declaration; declaration with a voice in the mid-range of F0 and intensity; seduction, authority, contempt, obviousness and irony, with lower F0 but higher intensity than declaration; politeness, sincerity and walking-on-eggs with a higher pitch than declaration, where walking on eggs had the lowest intensity. In commenting on the interaction of F0 and intensity in SAE, the authors suggested that "using a lower pitch for expression of dominance (here typically *authority* and *contempt*) is reminiscent of the predictions of the Frequency code (23)" (Rilliard et al. (2017), p. 38; also see Rilliard et al. (2009, 2013) for some qualitative acoustic observations about social affective expressions of US English).

Nevertheless. no real in-depth acoustic or perceptual analyses of these data have been performed. This paper is an attempt to remedy this by drawing some connections between acoustics, production, and perception of attitudinal expressions that we had not previously explored before. To this aim, we choose to do a case study: the dataset used here is only a small part of a larger corpus, as we aim to dig into the complex relationships between the acoustic characteristics of speech signals and the socio-affective expressions they carry through rapid and subtle changes in voice qualities. Of course, a small, speaker-specific sample only presents some facets of such a complex behemoth, but these fine phonetic details, observed through various lenses (acoustic parameters, voice quality settings, perceptual evaluations), bring insights into the important relationship and limitations to take into account so to avoid bias while describing and understanding social communications. We thus focus on finding links between perception and acoustics, and speaker-specific differences as a secondary thing. In the results and discussion sections, we first look at overarching patterns for different types of affective expressions. The acoustic analysis, as described below, is much more in-depth than previously conducted, including F0, intensity, pitch accents, and voice quality assessments; the perceptual analysis involves the 3-Dimension Valence-Arousal-Dominance emotional approach, as described below. As a secondary thing, we then look at speaker-specific things to motivate future research

directions or to try to have preliminary explanations regarding trading relations in terms of prosodic cues.

By way of more explanation of how SAE is distinct from emotions per se, we note that as we interact with others in a variety of situations to communicate with one another, our voice changes in a number of ways—we speak with a louder or softer voice, or more slowly or more quickly, or our voice pitch goes up or down, or our tone of voice becomes more tense or less harsh. The social constructs speakers make as they communicate with others are referred to here as "social affective expressions" (expressing politeness or doubt, marking surprise or irony); they may correspond to what is designed in the literature as *prosodic attitudes* (Uldall 1960; Fónagy and Bérard 1972; Wichmann 2000) and *illocutions* (see Couper-Kuhlen 1986; Mello and Raso 2011 for discussions). Note that these attitudes are not emotional expressions, such as anger, grief, joy, etc., as mentioned above. They are related to basic emotions, but social attitude expressions have undergone a conventionalization due to a number of factors, including the culture in which they are expressed. Hence, voice changes in social affective expressions vary according to social-awareness factors such as sensitivities to these nuances of social and cultural interactions, our individuality (as a person of a given gender, age, social and dialectal background, education, professional profile, social role, for example), and our personality (e.g., an extroverted person may speak with a louder voice and more pitch variations; Erickson et al. 2018), etc.

Such voice changes that occur with social affective expressions are related indirectly to basic emotions (e.g., Scherer et al. 1984) in that feeling more tense or more relaxed will affect our voice: muscular tension in the body affects the way we breathe, and consequently the quantity of air passing through our vocal folds; it affects the tension in the vocal folds, and thus their vibration patterns, etc. (for detailed predictions, see Scherer 2009a, 2009b). How aroused we feel can thus affect the acoustic manifestations of our expressions, which in turn not only affects how aroused the voice sounds (Arousal) but how positive/pleasant (Valence) or how assertive/dominant (Dominance) the voice is (Osgood et al. 1975; Goudbeek and Scherer 2010). Note four dimensions, not only three, are necessary to account for variation across emotional expressions, adding an *unpredictability* dimension (Fontaine et al. 2007) that – if it proved particularly important for analyzing socio-affective expressions (Rilliard and de Moraes 2017)– is of lesser relevance in the subset that will be studied here.

The Valence Arousal Dominance (VAD) percepts are manifested acoustically: loud, high-pitched voices are perceived as aroused/excited (e.g., Juslin and Laukka 2001; Scherer 2003; Bänziger and Scherer 2005; Goudbeek and Scherer 2010; Schmidt et al. 2016; Erickson et al. 2020), as also faster (Goudbeek and Scherer 2010) and higher pitched voices (Mozziconacci 1998; Schröder et al. 2006). Schmidt, Janse, and Scharenborg (2016), however, report that loudness is a more important cue than F0 for arousal. Tense/non-breathy voices are more aroused than breathy voices (Anikin 2020; Erickson et al. 2020a). Tense voices are also heard as more dominant (Anikin 2020), as are voices with high and "less noisy and flatter spectrum" (Goudbeek and Scherer 2010, p. 1332), as well as louder, faster speech with a wide range of F0 (Geng et al. 2020). As for valence, duration and spectral cues are important (Juslin and Laukka 2001; Scherer 2003); a steep spectral slope is an acoustic correlate of positive valence (Goudbeek and Scherer 2010). Breathy voices with steep spectral slopes are more positively rated than tense voices with sustained energy in the upper frequencies (Anikin 2020). Goudbeek and Scherer (2010) report that arousal explains most of the change in acoustics and that valence and dominance cues are dependent on the level of perceived arousal.

Whether a voice/speech utterance is perceived as having a positive or negative valence is a question that is receiving increased interest in the speech research field. Voices with positive valence tend to be more attractive, sexy, or charismatic (see, e.g., Weiss et al. 2021). Moreover,

valence cues seem to be culturally dependent: German listeners find high-pitched voices unpleasant (Schmidt et al. 2016); Japanese and Mandarin listeners prefer a high-pitched, less breathy voice, while Brazilian Portuguese listeners prefer a lower-pitched, breathy voice (Erickson et al. 2020a). For related work concerning cross-cultural similarities and differences in emotion recognition, please see e.g., Russell (1994); Elfenbein and Ambady (2002a, 2002b, 2003); Barrett (2006); Hareli et al. (2015); Laukka and Elfenbein (2021).

Beyond emotional expressions, vocal variations rooted in biological characteristics may also play a role in social affective expressions. Ohala's Frequency Code (Ohala 1983, 1984, 1994) links lower pitch to a larger body (based in large part on studies of voices of other mammals, and thus, the motivation for the derived symbolism of low voice for dominance, including speech acts such as assertions and questions. An extension of such a biological code has been proposed by Gussenhoven, notably with the Effort Code (Gussenhoven 2004), which links greater vocal effort to more dominant behavior, among other predictions. These assumedly biologically-motivated codes manifest themselves in social communicative interactions, and will be referred to as "social codes," codes that are employed by speakers, along with emotion-motivated vocal changes, to communicate socially in daily life situations using conventionalized social affects (see de Moraes, 2011, for a discussion).

Additionally, and importantly, another source of variation in social affective expressions is due to our unique set of glottal source and vocal tract configurations, which results in different articulatory strategies to produce the expression and, consequently, may derive different acoustic manifestations for the same communicative act. Moreover, there is not necessarily a single acoustic parameter nor even a unique constellation of acoustic parameters that describe these social affective expressions, although, in the end, speakers manage to produce recognizable social affects (see Fitch et al. 1980 for a similar argument at another analysis level).

To better understand links between articulatory strategies and social affective expressions, we need to find a way to explain the relation between articulation, perception, and the acoustic manifestation. Here in this paper, we try to enter the details of what two speakers are doing to express a small set of expressions and to show how it may be complex and how it may be explained. A naïve assumption might be that speakers, at least of the same language/cultural background, would use the same or similar cues to express the same social affect, e.g., that American English speakers would speak with louder, lower-pitched voices to express Irritation, which would be heard by American listeners as aroused, dominant, and not very likeable. However, the situation is much more complex, especially when we add in voice quality changes. With regard to voice quality, currently, there are several models for assessing it, notably that by John Laver (1980), who described voice quality as involving long-term articulatory settings that impact the general sounding of one's voice without necessarily affecting the perceptual access to phonemic categories. Such auditorily perceived voice qualities have different settings in terms of source (vocal fold configuration) and filter (supra-laryngeal) configurations, plus a general tension setting (that applies to both source and filter). Clinicians, in assessing pathological voice quality, employ a system called Voice Profile Analysis (VPA), which is based on Laver's work and is currently a widely used method for assessing voice quality (e.g., Camargo and Madureira 2008; Camargo et al. 2019; San Segundo et al. 2019).

Acoustic analysis of speech has come up with many different parameters for examining changes in the signal that relate, at least in part, to auditory voice qualities (for a review centered on voice source, see d'Alessandro 2006). For instance, an index of breathy voice quality–described by Laver (1980, p. 132) as "By comparison with modal voice, the mode of vibration of the vocal folds is inefficient, and is accompanied by slight audible friction"–that is commonly

used is the difference between the amplitude of the first harmonic (H1) and that of the next harmonic (H2), potentially with a correction for the possible influence of vocalic resonances (Hanson 1997); however, this measure works best for oral low pitched /a/ vowels. Indices of good harmonicity (strong harmonics, little noise) includes Harmonic to Noise Ratio (HNR; e.g., d'Alessandro et al. 1998; Jackson and Shadle 2001) and Cepstral Peak Prominence [CPP] (Hillenbrand et al. 1994; Hillenbrand and Houde 1996), where large values indicate strong harmonics, and conversely lower values indicate higher noise levels, comparatively to the energy in the harmonic part (but note different types of noises may be indiscriminately mixed; see Rilliard et al. 2018 for a discussion). The so-called "Hammarberg" index (Hammarberg et al. 1980; Banse and Scherer 1996) is an estimation of the spectral slope (linked to vocal effort; Liénard and Di Benedetto 1999; Liénard 2019). High values of this index indicate steep spectral slope, linked to low effort and possibly breathy phonation; low values indicate more energy in the upper frequencies, and are typical of higher vocal effort. For estimates of the tense-lax vocal dimension, two popular measures are (1) Normalized Amplitude Quotient (NAQ), related to glottal closing behavior, where larger values indicate tenser voice (Alku et al. 2002), and (2) Peak slope (PS, Kane and Gobl 2011), which was designed to estimate spectral tilt. This is not by any means an exhaustive list. The point, however, is that none of these parameters completely captures what breathiness is, or what tenseness is. This is to a large part due to the many factors affecting voice quality, (e.g., rate of vocal fold vibrations, phonation mode, glottal closure amplitude, pharyngeal narrowing, vowel quality, etc.) and notably the highly complex role of supraglottal tract, so that such parameters have relations with various aspects of articulatory voice qualities as described in, e.g., Laver (1980) or Esling et al. (2019), and as appears clearly from d'Alessandro (2006) Table 2 (p.78).

To date, as far as we know, no in-depth study of voice quality has been done for the social affective expressions examined here. Since voice quality during the course of a single utterance changes with segmental and prosodic make-up, in this paper, in addition to an examination of F0, intensity, and duration for the entire utterances, we did an in-depth examination of acoustic and auditory cues for the nuclear stress vowels in the first syllable of the word dancing.

In this case study of a few expressive samples from two female speakers, we make a detailed phonetic analysis of the various dimensions of affective expressions, especially targeting voice quality since the audio recordings are of high quality. If the phonetic data is not generalizable to other speakers, it does show that for these particular expressions and for the strategies selected by these two speakers, the same communication goal is achieved in ways that are similar yet different. For the perception aspect of the study, we asked groups of listeners to rate the Valence-Arousal-Dominance dimensions and also the auditory aspects of voice quality in these speakers' social affective expressions. The assumption is that such perceptual judgments apply at a more generalizable level, at least to American listeners, for these specific stimuli. In this paper, we are trying to describe the complex relationships between the social expressive strategies of these two women of very similar age, geographic location and cultural background and their reception by perceivers.

2 Methods

2.1 Recording procedure and corpus

As mentioned in the Introduction, the social affective expressions, such as e.g. authority, irritation, uncertainty, etc., are part of a larger corpus; the purpose was to help second/foreign language students successfully communicate in that language/culture since the phonetic implementations

of these expressions vary across languages and cultures (e.g., Rilliard et al. 2014; Mixdorff et al. 2018; Idemaru et al. 2019; Shochi et al. 2020). The recorded data include sixteen social affective expressions produced by various language learners, e.g., English speakers learning Japanese, Japanese speakers learning French, etc. For details on the motivation of the sixteen social affective expressions and a description of some of the differences observed within and across different languages/cultures, the reader is referred to Rilliard et al. (2013, 2017). Dialogue specifications were given in terms of the relative hierarchical levels of the interlocutor, their social relation and the communication aim of the speakers. Thus, the two speakers (the experimenter and the experimentee) relied on the dialogue description to play a short scenario that led to the target utterance. Typical social affective expressions were collected during such interactive dialogue tasks in which the recorded speaker played a role with the experimenter, such that the dialogues ended with the speaker saying a target utterance ("Mary was dancing"). As produced during an interactive (albeit scripted) dialogue task with given communication goals, the social affects were not defined by their labels but by the communication situations, which makes them more easily comparable across speakers and cultures, as they do not rely on the interpretation of a given label or its translation (see Wierzbicka 1992).

The recording took place at Waseda University (Japan) in sound-prepared facilities using an Earthworks QTC1 omnidirectional microphone placed one meter from the speaker's mouth, and with its recording level calibrated using a Brüel & Kjær acoustical calibrator; the microphone was plugged into a Panasonic AG-AC160 video camera recording in AVCHD format. The resulting videos files were later edited, the sound level corrected for varying recording levels, and individual sentences were stored in individual mp4 files.

Since not all speakers are equally efficient in their verbal communication skills (notably depending on their personality, training, etc., see Erickson et al. 2018; Niebuhr et al. 2019), and in order to ascertain whether a speaker successfully expressed a particular affect (i.e, politeness, sarcasm, seduction/flirtation, arrogance, etc.), perceptual evaluations of their performances by first language speakers were carried out (Rilliard et al. 2013). The results showed interspeaker differences in how well they could convey to listeners the intended social affective expressions.

We selected well-recognized attitudes from the USA English corpus, as judged by American listeners (Rilliard et al. 2013). As described in this study, the judgments were made by 17 subjects (7 females and 10 males), mean age 25), all native AmericanEnglish speakers; they listened to 256 stimuli (8 speakers performing 16 attitudes with two sentences) and rated the performance of the speaker in expressing the targeted attitude, on a 1 to 9 scale. Based on the results, six well-perceived attitudes produced by two female speakers on the target utterance "Mary was dancing" were selected. Thus, in this pilot study, we focus on twelve stimuli. The interest in our current study is not the capacity of listeners to identify or label the targeted speech acts, but to raise questions as to how speakers may use various vocal cues to form a panel of expressions, how their strategies matched the theoretical social codes proposed in the literature (and which of these codes), and how the individual vocal characteristics of speakers led them to use different strategies for these expressive aims.

As for the speakers, one speaker (S6) was the first author of the paper; the other speaker (S3) was a friend of S6; both had considerable experience teaching English as a second language in Japan, and were, therefore, able to express well-recognized social affective expressions which could be used to help teach learners of English how to express American-style social affective expressions. The speakers exchanged roles, with S6 playing the role of the experimenter first and then S3 playing the experimenter. The fact that the two speakers were friends probably lent reality

to the various dialogue situations. At the time of the recording, S3 was in her early 50's and S6 in her mid 60's. Both speakers had a certain amount of training as soprano singers. Also, both speakers had lived in Japan, S6 somewhat longer than S3. Both speakers were Caucasian, roughly of the same generation, and shared a similar mid-western cultural background. S3 was born in Minnesota, and S6 in neighboring South Dakota.

The six target social expressions examined here are listed in alphabetic order: Authority, Declaration, Irritation, Sincerity, Uncertainty, and "Walking On Eggs" (WOEG). A note about Declaration: this is considered a neutral SAE, to function as a type of baseline. The situations corresponding to the six expressions are shown in Table 1. Speaker A is the target speaker who performs the expressive utterance, while speaker B is the experimenter.

Table 1: Characteristics of the dialogues used for the recordings, with some contextual elements about the dialogue, the relations between the two speakers, and the place where it took place.

Label	Situation	Dialogue
Authority	Speaker B asks what Mary was doing, and speaker A answers with an authoritarian tone in order to influence Speaker B or to impose his view or wish. Speaker A is in charge of the dance school. Speaker B and Mary are working in this dance school as teachers. The scene is at the dance school.	B: Do you know what Mary was doing? A: Mary was dancing.
Declaration	The speaker A gives the information [Mary was dancing] without any personal perspective. Speakers A & B are colleagues, same age. The scene is at a Starbucks Coffee shop.	B: What was Mary doing when you arrived? A: Mary was dancing.
Irritation	You (speaker A) are sitting next to speaker B. Speaker A answers "Mary was dancing" expressing his irritation towards his interlocutor (B), since it is the third time speaker A is answering the same question and his interlocutor B doesn't pay attention. Both speakers A & B are almost same age and know each other. The scene is at public place.	B: What was Mary doing last night? A: Mary was dancing. (I already told you 3 times before. Are you deaf or what?)
Sincerity	Speaker A's chief didn't find Mary last night. Mary needs to improve her skills more for the next competition which is coming very soon. You were with her yesterday night, and you know she was dancing very hard. Speaker B is chief of the ballet school which speaker A belongs to. Mary is Speaker B's colleague. The scene is at the dance hall.	B: What was Mary doing? A: (I tell you the truth) Mary was dancing.
Uncertainty	Conveying "I think that…(slow) Mary was dancing, but I am not 100% sure, when I got there, it was very crowded, and I couldn't see very well." Speaker A expresses uncertainty about the information he is giving. Speakers A & B both are colleagues, same age. The scene is at Starbucks coffee.	B: What was Mary doing when you arrived? A: Mary was dancing
WOEG	Speaker B, who is speaker A's chief, has been looking for Mary since yesterday. Mary needed to send an important document to Speaker B by last night, but she didn't. Instead, she was dancing. You were with Mary, and you are afraid to tell on her. But you realize you have to tell the truth, even though you don't want to.	B: Where was Mary last night? I called her several times last night, but she didn't answer the phone. She needed to send me the document as you know! I know both of you were together last night. What was she doing?

	Speaker B is chief of the section which speaker A belongs to. Mary is Speaker B's colleague. The scene is at Speaker B's office.	A: (Huuh.) Mary was dancing.

2.2 Acoustical analyses

First, the stimuli were segmented at the phonemic level using Praat (Boersma and Weenink 2022) TextGrids, and a series of acoustic measurements were estimated from the signal with a 10ms time step. The following ones were estimated thanks to Praat algorithms:

- The voice's fundamental frequency (F0, expressed in semitones re. 1Hz) was measured using Praat's "ac" algorithm and hand-corrected using Praat's Pitch object; both the voicing decision and the choice of the F0 candidates were checked; better candidates were selected to fit pitch perception in case of problems. Estimating the rate of vocal fold closure, F0 is widely used to study intonation and is seen as the main acoustical correlate of perceived vocal pitch in voice, albeit pitch is influenced by many factors (e.g., Rossi 1978; Ohala 1994; Bishop and Keating 2012).
- Signal intensity, expressed in dB, was estimated using Praat default parameters (i.e., the "pitch floor" parameter was set to 100Hz; recall the recording level was calibrated during the recordings, so the measurements shall be comparable across speakers). Intensity is a direct relation to the perceived loudness of speech (with a power function; see Stevens 2000, p. 225), and shall also be related to the perceived pitch (Niebuhr et al. 2020).
- The Cepstral Peak Prominence (CPP; Hillenbrand et al. 1994; Hillenbrand and Houde 1996) was estimated using Praat implementation of CPP (Maryn and Weenink 2015). CPP estimates, in the cepstral domain, the difference in energy between the harmonic structure of speech (its voiced or periodic component, measured at the cepstral peak) and the noise component (its aperiodic part, estimated through a regression over the cepstrum). CPP was built to robustly detect aperiodicities (including breathiness) in pathological voices as opposed to modal and more periodic voices; it is "robust" as it did not require F0 to estimate this periodic/aperiodic difference, which is a difficult measurement for some pathological voices.
- The Hammarberg index (Hammarberg et al. 1980; Banse and Scherer 1996) was measured as the difference in peak energies of the 0-2 and 2-5kHz bands, following the procedure described in Hammarberg et al. (1980); it is expressed in dB. This index gives an estimation of the proportion of spectral energy in the lower part of the spectrum (0-2kHz) compared to the higher part (2-5kHz), which is linked to the spectral slope and thus shall oppose low effort, breathy, or hypofunctional voices (that have few energy in the high-frequencies) to hyper-functional, or high effort, tense voices (that have comparatively more energy above 2kHz).

  Using the COVAREP toolbox (Degottex et al. 2014), two more voice quality parameters were estimated that are described in the literature as estimates of the tense-lax vocal dimension:

- The Normalized Amplitude Quotient (NAQ; Alku et al. 2002). This measurement is an estimation of the closing quotient (CQ) of glottal flow models: it represents the relative duration of the closing phase during one cycle of vocal folds vibration, which differs according to the folds' tension. Breathy voices have slower fold movements, thus higher CQ, while tense voices have more rapid closing, hence smaller CQ. The AQ estimate is

based only on measures in the amplitude domain from the glottal flow waveform and its derivative – thus it is thought to be more robust than methods requiring the determination of precise events in the time domain. It is then normalized (hence the "N" of NAQ) according to the duration of a fundamental period. NAQ is derived from the glottal waveform and thus requires inverse filtering to be applied to the speech signal to remove the contribution of the vocal tract transfer function.

- The Peak Slope parameter (PS, Kane and Gobl 2011). It was developed to measure voice quality differences along a breathy-modal-tense continuum and thus returns an index representing a kind of spectral slope, but measured on the wavelet transform. It fits a regression line through the peaks of each wavelet scale and returns the slope of this line. This (negative) slope is expected to be shallow (close to zero) for breathy voices and steeper for more tense examples.

Our research is based on a prosodic model where syllables are the basic prosodic unit (e.g., Fujimura 2000), the nuclear stress syllable in the utterance has the most salience (e.g., (Erickson and Niebuhr 2023), and an utterance is a composite of the syllables. Hence, we examine the acoustic characteristics of the sentence, the syllables, and the nuclear stress syllable.

Focusing on the complete sentence, the values of these parameters were summarized, taking into consideration the sentence's mean and standard deviation on vocalic segments only. Since many factors affect voice quality during speaking (i.e., vowel quality, intensity, F0, etc.), we also focused on one vowel only, the vowel that has the most stress in the utterance, which is, therefore, the most salient and most likely to carry social affective information. Specifically, we measured the acoustic attributes for the vowel in the final nuclear-stressed syllable of the utterances, the /æ/ in dancing. Since, even within this single vowel, voice quality parameters and F0 change, we considered the median values of the initial, medial, and final third of the vowel for each acoustic parameter.

Duration parameters were also estimated from the phonemic segmentation: (i) the duration of the complete sentence, (ii) the mean and standard deviation of syllabic duration in each sentence, and (iii) the duration of the nuclear stress syllable.

Finally, for intonation patterns, a ToBI analysis (Mainstream American English) was used to describe the phrase boundaries and pitch accents (Pierrehumbert 1980; Silverman et al. 1992; Beckman and Ayers 1997). The annotations were done by the first author, trained in ToBI working with The Ohio State University group.

2.3 Perceptual analysis

The perceptual aspects of the affective expressions were approached from two perspectives: First, listeners were presented with complete "Mary was dancing" sentences they had to rate on the 3-Dimensional Valence-Arousal-Dominance (VAD) scales; next, to address voice quality, the vowels in the nuclear stress syllables of the six paired utterances were presented to a second group of listeners who were asked to auditorily rate six voice quality dimensions. The details of each perception test are described below.

2.3.1 Valence Arousal Dominance

The listener ratings of the VAD perception tests have been reported in Erickson et al. (2022).

Based on performance ratings obtained in Rilliard et al. (2013), the twenty best-performed social affect for the target sentence "Mary was dancing" were selected. Both speakers produced social affective expressions labeled as "Authority, Declaration, Irritation, Sincerity, Uncertainty,

Walking On Eggs." In addition, S3 produced "Admiration, Arrogance, Contempt, Irony, Obviousness" and S6 produced "Politeness, Surprise, Seductiveness." (the best performances of each speaker were selected.) So, among these twenty stimuli, twelve were paired in terms of social affects across the two speakers: these are the six targeted social affects of this article.

Fourteen listeners (12 f, 2 m) at a northern California university listened to the twenty social affective utterances as part of a classroom exercise. The listeners, all fluent speakers of English with no known hearing disorders, were presented with the utterances using a PowerPoint presentation in a classroom; the sounds were played over a loudspeaker. That not all listeners had English as their first language, that no other demographics were collected, and that the sounds were played over a loudspeaker are limitations of this study. The listeners were asked to rate their impressions of each utterance in terms of how excited, how positive, and how assertive the speaker sounded. Specifically, after hearing each utterance, they were to mark on their answer sheet using a scale of 1 to 5, where 5 indicated very excited or very positive or very assertive, and 1 indicated very calm or very negative or very accommodating. These three laymen definitions were taken as terms referring to the emotional dimensions referred to here as Valence Arousal and Dominance, and that are defined as follows by Goudbeek & Scherer (2010): "Valence refers to the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation leading to an emotion. Arousal represents the degree of alertness, excitement, or engagement produced by the object(s) of emotion […]. Potency/control has a long but more controversial history in emotion research referring to the individual's sense of power or control over the eliciting event" (Goudbeek and Scherer 2010, p. 1322). Dominance in this work refers to what Goudbeek & Scherer (2010) call Potency/Control.

Each slide had one utterance sound file along with an illustration of little mannikins (taken from the work by Xue et al. 2018) portraying a series of facial expressions ranging from calm to excited, or negative to positive, or accommodating to assertive (see Erickson et al., 2022, for more details). Each utterance was presented to the listeners 3 times in block order; the first time, they were to rate how excited the utterance was; the second time, how positive the utterance was; and the third time, how dominant it was. Although the listeners were asked to rate each utterance on the three scales, a few answers are missing (3 for arousal, 4 for valence, 5 for dominance).

2.3.2 Auditory voice quality impressions

As was mentioned in the introduction, voice quality is perceptible by listeners, but not easily describable in auditory terms. One difficulty with a perceptual evaluation of voice quality is its rapidly varying nature: indeed, within the utterances, it was not necessarily constant. Another problem is that not everyone uses the same terms to describe voice quality (Henrich et al. 2008). Thus, in order to address these two difficulties, we created a basis for comparison of voice quality ratings for the listeners: we asked listeners to rate the /æ/ in dancing in relation to prototypical examples of a nasalized /æ̃/ vowel produced with varying voice qualities: breathy, falsetto, voice loudness, voice pitch, tenseness, and twangy-ness.

Exemplar vowels were used because, in real speech, voice qualities are mixed, e.g., F0 and loudness tend to covary such that a voice with a higher F0 is generally heard as louder also. Using exemplar vowels for which the voice quality was more controlled affords a better window into ascertaining which acoustic parameters best match the auditory percepts. A professional soprano singer trained in the Estill voice method (Steinhauer et al. 2017) made the recordings. Because of her training, she was able to isolate and control to a large degree supralaryngeal (tongue, velum, jaw, mid pharynx, low pharynx, larynx height, etc.) and laryngeal (vocal fold thickness, arytenoid approximation, thyroid tilt, cricoid tilt, etc.) articulations to produce the six types of auditory voice

quality dimensions. Thus, she was able to produce a twang-quality exemplar, for instance, by narrowing the mid-pharyngeal area (see e.g., Perta et al. 2021; Erickson et al. 2022b, for an articulatory and acoustic description of twang). She recorded six pairs of exemplar /æ̃/ vowels (producing the word dancing). The six pairs of voice qualities were the following:

- Breathy // Not breathy
- Falsetto // Modal
- Loud // Soft
- High // Low
- Tense // Not tense
- Twang // No twang

The voice qualities on the left side are the "non-normal" targets, and "normal" on the right side, with "normal", or "modal" voice quality as was defined by Laver (1980), with the exception of the "Loud–Soft" and "High–Low" pairs.

For the perception test, the nuclear stress vowel of "dancing" (/æ̃/) was extracted from each utterance in both the social affective stimuli and the exemplar stimuli. In recording the exemplar sounds, the professional singer listened to all the utterances. She determined which utterances were most breathy, most twangy, most tense, etc. The most breathy utterances were the WOEG expressions for both S3 and S6; thus, the breathy exemplar sound was produced to match the breathiness of these expressions (note that the WOEG expression may also carry other voice qualities). The Irritation expressions were twangy, and so the twangy exemplar sound was based on these expressions. For the high–Low pitch exemplar, the singer produced a high exemplar sound at 350 Hz and a low one at 100 Hz, based on the F0 range of S3 and S6.

Using the "psytoolkit" online interface (Stoet 2010, 2017), listeners were asked to use headphones and listen to the twelve /æ̃/ vowels extracted from the affective expressions, and then to indicate, using a nine-point Likert scale, how close each sound was to the exemplar sounds, e.g., breathy or non-breathy, etc. The instructions defined the different voice qualities as follows:

> *As an example, for Breathiness, sound Y represents a very breathy sound; sound X represents a very non-breathy sound. Your job is to indicate if sound Z is more like the breathy Sound Y or the non-breathy Sound X.*
>
> *For Twangy voice quality—one sound is more like a Nashville country singer or a Texas drawl or a wicked witch cackle.*
>
> *For Tense voice quality—one sound is spoken with a more intense voice, as opposed to a more relaxed voice.*
>
> *For Falsetto voice quality—one sound is more like expressing surprise, one more like normal speech.*
>
> *For High pitched voices—one sound is high-pitched as opposed to a low-pitched sound.*
>
> *For Loud voices—one sound is loud while one sound is soft.*

Eighteen listeners, all North American English speakers, completed the test. The listeners belonged to an online song circle, and so had some musical awareness, but not necessarily musical training. If they didn't understand exactly the meaning of a voice quality term, they were encouraged to ask for more information. They completed the test online, using headphones. The age of listeners ranged from 35 to 76, with a mean age of 58, thus, similar in age to the S3 and S6 speakers. No listener reported hearing problems. No other demographic information was collected, which is a limitation of the study. About the layman's descriptions of the auditory voice qualities, especially "twangy"—this may have introduced some limitations due to the various cultural references it may evoke in some listeners; however, since all the listeners were musicians,

familiar with country music to some extent, we feel that any possible pejorative connotations of "twang" were minimized.

2.4 Statistical analysis

The dataset at hand is based on a set of six social affective expressions produced by two speakers: we are here interested in the speaker-specific strategies, so a descriptive step will be presented, based on the observed value or on measures of central tendency and dispersion of the acoustic parameters collected from the complete sentences or from the nuclear-stressed vowels (see details above).

As a second step, the perceptual evaluations of these expressive productions by groups of listeners will be detailed using regression models to sort out which factors have a significant and important effect on their answers. We hypothesize that the perceptual results would be relatively stable over a larger population of listeners, at least with similar language backgrounds, thus the inferential approach.

The final step aims at observing potential links between these two description levels (production and perception), comparing acoustic measurements obtained on the complete sentences with the VAD results, and acoustic measurements on the stress vowels with the auditory perceptions of voice quality. To that aim, we used a Multiple Factor Analysis (MFA; see Pagès and Husson 2001; Husson et al. 2017), that reduces the dimensionality of each data table based on several sub-tables with heterogeneous data types. For complete sentences, three sub-tables were considered: the VAD scales, the acoustic parameters, and the duration measures (corresponding to the divisions in Table 2 where the raw numeric values are given); for nuclear stress, two sub-tables were considered: the six voice quality scales and the measurement on each third of the nuclear-stressed vowels (it corresponds to divisions in table 3). The MFA then helps in observing how each dataset is related to the different stimuli (the table's rows). This is essentially a descriptive tool that allows an organized comparison of complex multidimensional datasets. From each MFA output, based on the projection of the table's rows along the first principal axes of the MFA, a hierarchical clustering algorithm was applied (following Husson et al. 2017). We used this algorithm to better represent the similarities and differences between the data points (through the hierarchical construction of the dendrogram) and to extract the more relevant differences in applying clustering to the results. The reader may usefully read Abdi and Williams (2010) for the interpretation of the output tables.

Table 2: for each speaker (Spk S3 and S6), and for each social affect expression (SAE), values used for the MFA analysis on complete sentences. For perceptual data: mean z-scores on the V, A, and D scales. For the production parameters observed over the sentences, mean and standard deviation (m/s) are reported respectively on the first and second line for each parameter, when applicable (the lines presenting standard deviation are shaded): (m/s) fundamental frequency (F0, ST), (m/s) intensity (Int, dB), (m/s) cepstrum peak prominence (CPP, dB), (m/s) normalized amplitude quotient (NAQ), (m/s) peak slope (PS), (m/s) Hammarberg index (HI, dB). For the duration parameters (measured in seconds): duration of sentence (Dsent), (m/s) duration of syllables (Dsyl), duration of nuclear stress syllable (Dnss).

Spk	S3	S3	S3	S3	S3	S3	S6	S6	S6	S6	S6	S6
SAE	Auth	Decl	Irri	Sinc	Unce	Woeg	Auth	Decl	Irri	Sinc	Unce	Woeg
A	-0.085	-0.018	1.377	-0.090	0.027	-0.408	-0.351	-0.658	1.372	-0.370	-0.646	-0.097
V	0.168	0.634	1.114	-0.015	-0.816	-1.140	0.424	0.387	0.159	0.442	-0.778	-0.686
D	0.884	0.391	0.464	0.549	-0.544	-1.067	0.365	-0.267	1.101	-0.213	-0.895	-0.991
F0	90.7	90.6	93.3	87.8	91.5	89.7	94.5	92.5	97.8	94.1	91.6	93.5

	1.53	2.45	4.40	2.80	3.47	3.20	1.44	1.09	3.52	2.06	1.50	5.02
Int	65.2	62.8	67.1	61.6	56.9	57.0	67.9	66.1	72.9	66.2	62.4	59.8
	5.25	3.66	5.16	3.61	6.82	3.11	2.46	2.40	3.81	4.74	3.20	5.32
CPP	21.8	21.1	20.6	18.5	16.2	16.6	22.3	19.6	20.6	17.4	17.2	13.0
	4.60	3.03	3.95	4.43	5.33	4.19	3.67	4.19	3.95	2.46	4.87	4.30
NAQ	0.081	0.088	0.161	0.088	0.078	0.049	0.120	0.065	0.122	0.125	0.078	0.086
	0.0398	0.0339	0.0552	0.0330	0.0409	0.0379	0.0284	0.0402	0.0573	0.0521	0.0415	0.0442
PS	-	-	-	-	-	-	-	-	-	-	-	-
	0.5145	0.5217	0.5119	0.5053	0.4630	0.4878	0.5035	0.5189	0.4720	0.4847	0.5087	0.5021
	0.0299	0.0243	0.0283	0.0304	0.0559	0.0324	0.0219	0.0238	0.0346	0.0305	0.0291	0.0319
HI	20.9	21.8	19.5	24.2	23.9	26.0	16.4	24.4	9.8	22.5	29.3	29.6
	8.64	9.11	8.78	6.59	7.23	7.03	6.96	6.53	7.62	5.19	6.24	6.80
Dsent	1.050	0.920	2.210	1.010	1.930	2.070	0.980	0.860	1.210	0.910	2.090	2.230
Dsyl	0.210	0.184	0.398	0.202	0.386	0.414	0.196	0.172	0.242	0.182	0.376	0.231
	0.080	0.053	0.159	0.054	0.294	0.342	0.089	0.062	0.151	0.089	0.210	0.075
Dnss	0.320	0.240	0.650	0.270	0.830	1.000	0.310	0.250	0.380	0.300	0.520	0.330

Table 3: for each speaker (S3 and S6), and for each social affect expression (SAE), values estimated from the MFA analysis on nuclear stress syllables. For perceptual data: mean z-scores on the breathy/non breathy (Breath), falsetto/modal (Falset), high/low (High), loud/soft (Loud), tense/not tense (Tense), and twang/no twang (Twang) scales. For the production parameters observed over the three third of each nuclear stress vowel (the number at the end of the labels indicates the position of the measure), median value of: fundamental frequency (F0, ST), intensity (Int, dB), cepstrum peak prominence (CPP, dB), normalized amplitude quotient (NAQ), peak slope (PS), and Hammarberg index (HI, dB).

Spk	S3	S3	S3	S3	S3	S3	S6	S6	S6	S6	S6	S6
SAE	Auth	Decl	Irri	Sinc	Unce	Woeg	Auth	Decl	Irri	Sinc	Unce	Woeg
Breath	-0.365	-0.505	-0.466	0.031	-0.267	0.760	-0.599	-0.150	-0.879	0.815	0.282	1.325
Falset	0.176	0.002	0.371	-0.237	0.333	-0.894	0.946	-0.058	1.558	-0.684	-0.415	-1.162
High	-0.232	-0.378	0.657	-0.648	-0.267	0.290	-0.367	-0.190	-0.349	0.615	0.563	0.271
Loud	0.058	0.112	0.632	-0.522	0.360	-1.037	1.001	-0.075	1.266	-0.521	-0.747	-0.528
Tense	-0.299	-0.368	1.266	-0.813	0.201	-0.460	0.671	-0.023	0.896	-0.082	-0.111	-0.903
Twang	-0.003	0.169	0.588	-0.155	0.125	-0.413	0.823	0.142	1.048	-0.418	-0.915	-0.991
F01	91.6	89.2	94.9	86.6	92.6	88.8	94.4	91.7	94.1	91.2	91.6	81.1
F02	90.0	88.3	97.5	85.7	91.9	87.9	93.7	91.6	95.0	91.9	90.4	82.9
F03	88.7	88.2	100.2	85.8	91.2	87.6	93.2	91.3	97.5	92.0	91.4	83.4
Int1	67.0	62.7	65.1	64.0	65.9	63.0	68.8	68.9	71.8	69.6	67.1	68.2
Int2	68.0	65.0	70.0	60.1	63.3	60.5	64.1	65.7	70.0	65.7	62.3	62.3
Int3	68.7	64.4	80.3	58.6	63.3	57.9	65.6	62.7	70.2	62.0	59.9	62.2
CPP1	23.9	17.7	21.4	19.4	20.9	16.1	20.3	22.2	24.6	17.0	21.0	10.1
CPP2	21.4	23.1	20.1	23.4	22.6	17.0	25.2	24.7	24.5	17.2	16.8	11.5
CPP3	23.6	22.8	18.7	21.4	23.1	15.0	25.2	20.2	20.7	17.0	18.0	10.4
NAQ1	0.078	0.029	0.204	0.079	0.074	0.012	0.111	0.103	0.070	0.047	0.098	0.067
NAQ2	0.106	0.081	0.209	0.096	0.060	0.037	0.134	0.104	0.079	0.142	0.137	0.076
NAQ3	0.064	0.079	0.208	0.084	0.051	0.041	0.125	0.121	0.106	0.157	0.104	0.126
PS1	-0.5216	-0.5262	-0.4945	-0.5209	-0.5187	-0.5109	-0.5231	-0.5290	-0.4617	-0.4897	-0.5279	-0.4630
PS2	-0.5231	-0.5327	-0.5028	-0.4913	-0.5084	-0.4960	-0.4738	-0.5245	-0.4547	-0.5059	-0.5194	-0.4384
PS3	-0.5454	-0.5231	-0.5514	-0.4774	-0.5181	-0.4915	-0.4898	-0.5116	-0.4681	-0.4964	-0.5136	-0.4117
HI1	14.5	14.1	14.4	21.0	20.4	29.0	14.0	22.6	10.0	27.6	25.4	31.8

HI2	24.6	27.3	17.8	22.0	19.6	29.7	9.1	21.4	2.2	21.4	23.9	27.9
HI3	33.1	34.9	33.5	27.7	33.2	37.0	9.7	21.1	2.9	20.2	27.1	30.9

3 Results

3.1 Acoustic measurements

3.1.1 Tonal profiles

ToBI is a system used to annotate dialect or language-specific prosodic information, specifically, Tones (To) and Boundary Indices (BI). The notation marks F0 events (indicated with H(igh) and L(ow) markers to denote high and low F0, respectively), stress location (indicated with a * as in H*, L*, etc., to mark pitch accented syllables with high or low F0, respectively), and phrasal information (indicated with an – as in H- or L- to mark the end of an intermediate phrase that ends with high or low F0, respectively, or % as in H% or L% to mark the end of an intonational phrase that ends in high or low F0, respectively (e.g., Pierrehumbert 1980; Beckman and Ayers 1997). The ToBI annotation varies across dialects and languages and has, in general, been found to be a useful method for annotating intonational prosodic information. We use it here to describe the prosodic profiles of the six social affective expressions of the two female Caucasian speakers of Mainstream American English in our case study (see, e.g., Jun 2022, for an up-to-date account of ToBI transcription systems, its strengths, and challenges).

The intonation profiles for the six affective expressions produced by the two speakers are shown in Figure 1. To unclutter the plot, absolute F0 values are not shown. The relative height of F0 is marked with H or L, downstepped F0 height is marked with !, and syllable stress is marked with *, with the nuclear stress of the utterance falling on the last stressed syllable in the utterance. The utterance-final phrase is marked with %. Phrase breaks after the phrase “Mary was” (not shown in the figure) are such that Authority, Irritation, WOEG, and Uncertainty end with a high boundary tone (H-); the other two, Declaration and Sincerity, end with a low boundary tone (L-). The F0 max and mean values and F0 ranges were different for the two speakers (values are given here in Hz), with S3 showing a mean of 193, a median of 196, and a range of 112-260, and S6, a mean of 213, a median of 203, and a range of 144-306 (ranges are calculated from the fifth to the ninety-fifth percentiles for each speaker to avoid outliers).

Pitch accents on the nuclear stress syllable (the /æ/ in dancing) for the two speakers were H*+L for Authority, L* for Declaration, Sincerity, WOEG, L*+H for Irritation, and !H* for Uncertainty. The symbol !H* is used here to indicate a stressed syllable produced on a mid-level F0, lower than the preceding high F0, but not on a low F0. Note that for S3’s Irritation, she produced an additional phrase break after the auxiliary verb, was, with a stressed L*+H on was.

Pitch accents on the pre-nuclear stress region (Mary) were L*+H for all utterances, except for Declaration with H* for S3. Concerning the phrase accents and boundary tones, both speakers used a falling intonation (L-L%) for Authority, Declaration, Irritation, and Uncertainty. S3 used a falling intonation for Sincerity and WOEG while S6 used a rising intonation (H-H%) in both cases.

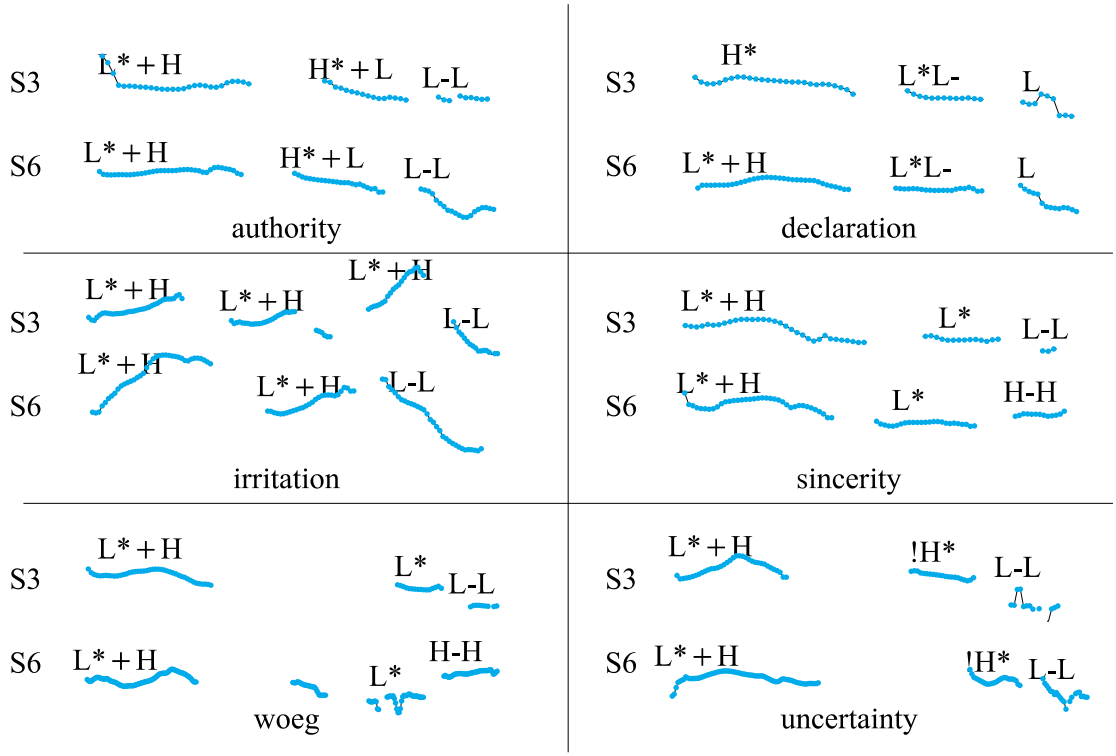

Figure 1: F0 contours of the social affective expressions, displayed in pairs, with S3 on top, and S6 on the bottom of each pair. Left pairs are as follows from top to bottom: Authority, Irritation, Walking on Eggs; Right pairs are as follows from top to bottom: Declaration, Sincere, Uncertainty. Y-axis indicates F0 (Hz) and x-axis, time. To focus on patterns of pitch accents, the scales of F0 values are not shown, albeit it was kept constant across sentences for coherence (ranging from 100 to 400 Hz).

As for the similarity of the intonational profiles of the two speakers, first of all, we mention that since the analysis was done by only one phonetician, this is a limitation of the study (see, e.g., Syrdal and McGory 2000; Yoon et al. 2004). Secondly, and as mentioned above, both speakers are Caucasian women from the same dialect region, both have similar cultural backgrounds, are roughly of the same generation, and are expressing well-recognized culturally determined social expressions. It, therefore, seems plausible that they might use similar intonation patterns. Since this is a case study with two speakers, it does not imply that all speakers of this dialect use these intonation patterns. As far as we know, no previous study of intonational patterns of these specific social affective expressions has been done. And, as an exploratory study, we hope this will inspire future work to explore intonational patterns of various social affective expressions.

3.1.2 Prosodic profiles for complete sentences

The prosodic profiles are described in terms of duration, intensity and F0 characteristics of the entire utterances. The duration profiles are shown in Figure 2. For both speakers, Declaration, Authority, and Sincerity are shorter than the other three; the others are longer, mostly because of longer pauses. For Uncertainty and WOEG both speakers inserted long pauses before the final word, dancing. Notice that for S3, Irritation was longer than Uncertainty. For Irritation for S3, two pauses were inserted, one before dancing, but also before was. For S6 for WOEG, the word "was" was repeated twice, but no pause before dancing.

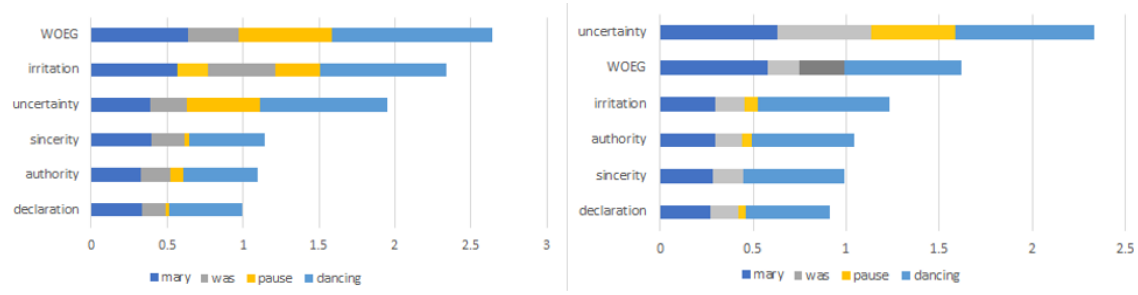

Figure 2: Duration profiles, in seconds, along x-axis. S3 (left) S6 (right)

F0 and intensity profiles are shown in Figure 3. S6 (right graph) generally speaks with a higher and louder voice than does S3; also note a more linear progression of mean intensity and F0 for S6, with Irritation the loudest and highest, and Uncertainty and WOEG, the softest and lowest. S3 also produces Irritation with the highest, loudest voice, and WOEG as the softest, but Sincerity is the lowest. For both speakers, Declaration, Sincerity, and Authority are in the middle range of intensity and F0.

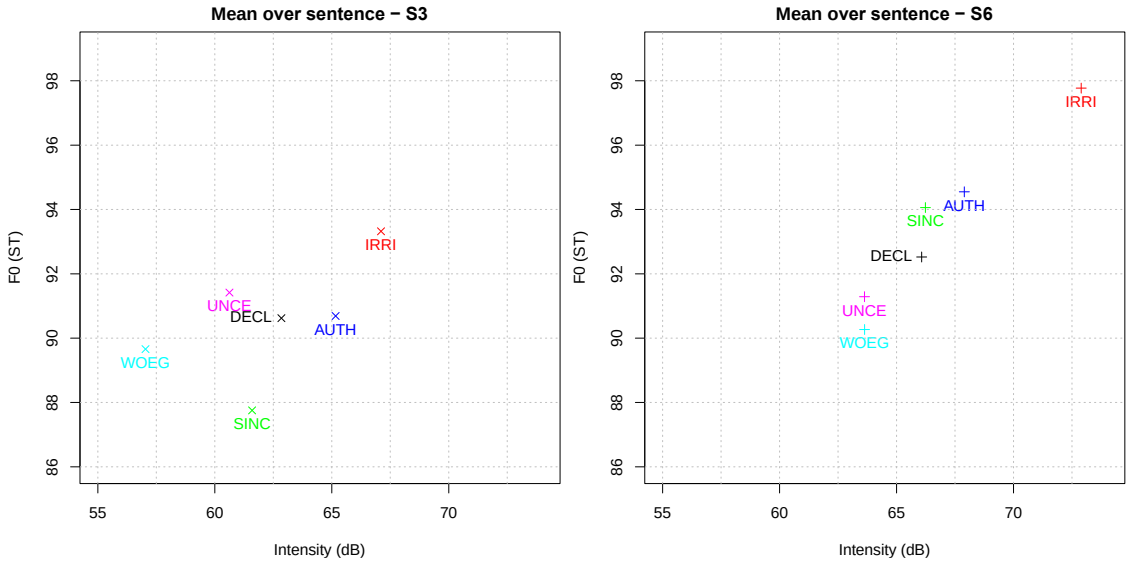

Figure 3: Mean F0 (ST) on the y-axis and mean intensity (dB) on the x-axis for all vowels in utterances. Speaker S3 is on the left panel, and S6 is on the right one.

Table 2 shows the mean and standard deviations over the vowels in the six affective expressions, as performed by the two speakers. (See the Methods section for a description of the acoustic values.)

3.1.3 Prosodic profiles for nuclear stress vowels

The prosodic profiles for the nuclear stress vowels are presented here. The values extracted from the stressed vowel for acoustic parameters and duration are detailed in Table 3, for each social affective stimuli of each speaker.

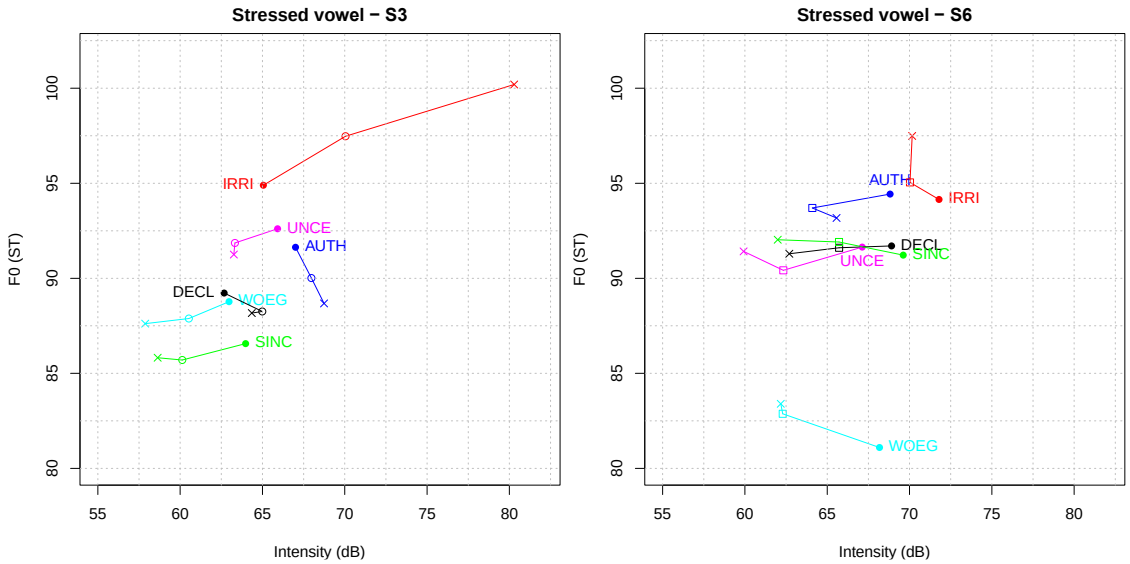

Figure 4: For each social affect, median intensity (dB, on the x-axis) and F0 (ST, on the y-axis) for the three thirds of the nuclear stress vowels (points respectively depicted by "•", "o", and "x" symbols for each third, linked by a line), for S3 (left) and S6 (right).

The intensity vs F0 profiles (Figure 4) for the nuclear stress vowels show some similarities and differences from those of the entire utterance (Figure 3). Note that the nuclear stress vowel for S3's Irritation is much higher and louder than for her utterance as a whole; it is also higher and louder than S6's nuclear stress vowel for Irritation: it displays a particularly large dynamic pattern. In general, the relation between intensity and F0 for S3's nuclear stress vowels is similar to that observed for the utterance as a whole, except that it is louder than for the utterance as a whole. As for S6's nuclear stress vowel, F0 is actually slightly lower than for the entire utterance, and interestingly, WOEG is considerably lower, about 7 semitones, than for the entire utterance. Otherwise for S6, in terms of F0 and intensity, her nuclear stress vowel and her entire utterance are very similar. Along a single nuclear stress vowel for most of the expressions, both speakers became softer at the end (the final third) of the vowel. But this is not the case for S3's Authority and especially not for Irritation (as mentioned above), expressions for which her intensity rises across the vowel. Note also for S3, F0 rises for Irritation but decreases for Authority. For S6, fewer dynamics are observed within the vowels (most of the changes are for decreasing intensity), and especially few along the F0 dimension, but for S6's Irritation expression, it has a rising F0 with a steady intensity level.

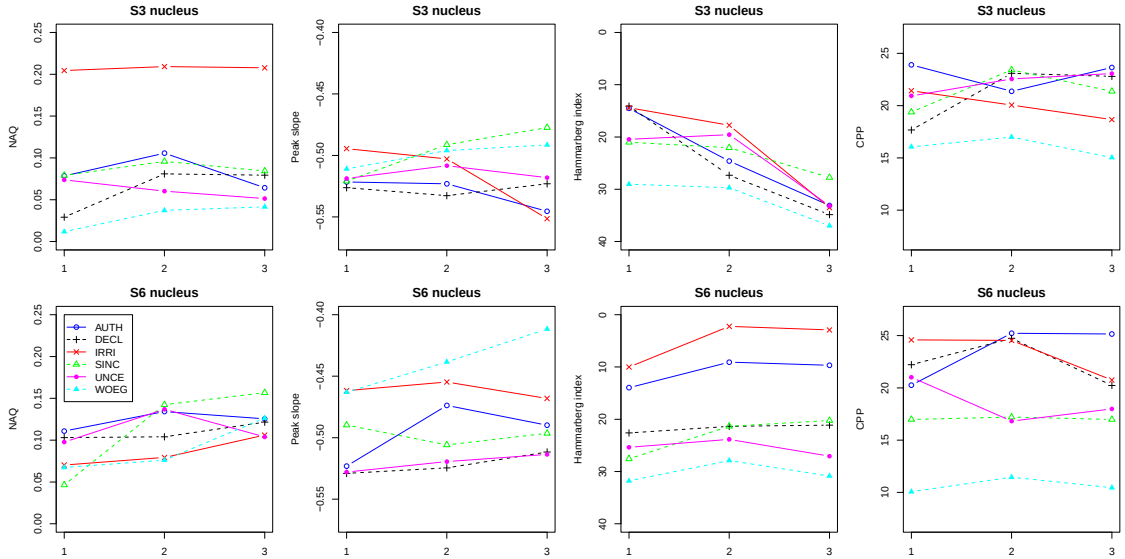

Figure 5: Median values, measured at each third of the nuclear stress syllable (x-axis 1, 2, and 3 points) of the NAQ, Peak slope, Hammarberg index, and CPP parameters (on separate plots by columns) for both speakers (S3: top row; S6: bottom row), for each social affect (individual lines). The columns are referred to in the text below as 5a, 5b, 5c and 5d, respectively.

As for voice quality profiles of the nuclear stress vowels, Figure 5 shows how the estimates of NAQ, Peak slope, Hammarberg index, and CPP vary for each of the expressions, for each of the speakers, and for each third of the vowels. These figures provide a graphic illustration of how variable voice quality is across speakers, across expressions, and even within in single vowel. Here we comment on a few of our observations.

Looking at NAQ in Fig. 5a, S3's Irritation has a very high value compared to her other expressions, and her WOEG has the lowest value. For S6, however, both Irritation and WOEG have the lowest NAQ values, and in general, S6's NAQ parameters show fewer changes. However, with regard to Peak Slope measurements in Fig. 5b, S6 shows more variation than S3; especially her Irritation and WOEG expressions received higher values (closer to zero): shallower Peak slopes are indicative of "breathier" voices (Kane and Gobl 2011) , which may include high flow voices (see the discussion about breathiness in Laver, 1980, p. 132ff).

Next, we look at Figures 5c (Hammarberg Index) and 5d (CPP). The Hammarberg Index (5c) provides information about the steepness of the spectral slope, where large values (displayed at the bottom of the y-axis) indicate steep drop off of energy in the upper frequencies, and small values (top of the y-axis) indicate no steep drop off of energy in the upper frequencies. The CPP (5d) estimates harmonic strength, with large values indicating strong harmonics.

Comparing figures 5c and 5d, notice that it is only the expressions with medium or low Hammarberg indices that have low or lowered CPP values. For both S3 and S6, WOEG clearly stands apart as having a steep spectral tilt (high Hammarberg index) and weak harmonics (low CPP values). For S3, the other five attitudes cluster in the range of having a shallower spectral tilt (more energy in the upper frequencies) and fairly strong harmonics, with Authority having slightly stronger harmonics. Note for S3, the intra-vowel variation of the spectral slope is systematic – with a spectral slope going steeper along the vowel. This systematic change may be due to the fact that S3 produces a diphtongation of the /æ̃/ vowel, with an upward shift of the second formant to be more [I]-like (as seen in spectrograms, as well as auditorily perceptible).

S6, on the other hand, showed a steady [æ̃]-like vowel quality. It is interesting that the Hammarberg index is sensitive to vowel articulation, suggesting the possibility of vowel-intrinsic spectral tilt, in addition to vowels having intrinsic F0 (e.g., Shadle 1985), and intrinsic duration/ intensity (Lehiste 1970).

For S6, Sincerity, Uncertainty and Declaration have relatively low energy in the upper frequencies, as evidenced by low Hammarberg indices (5c), and fairly strong harmonic structure as evidenced by high CPP values (5d), with Declaration having the stronger harmonicity. Irritation and Authority both have large spectral energy in the upper frequencies (5c) as well as strong harmonics (5d). As for intrasyllabic behavior, S6 tends to have different intrasyllabic behavior than S3, with decreasing spectral slopes for Authority and Irritation, and relatively few changes along the vowels for the others.

Comparing the voice quality strategies of the two speakers, S6 shows a more linear relationship between spectral tilt and harmonic strength than S3, i.e., as energy increases in the upper frequencies, so does harmonic strength. The voice quality characteristics for S3 seem to be not that much differentiated by these measurements (i.e., spectral tilt and harmonic strength): the nuclear stress vowels for all of her expressions, except WOEG, have comparable values of energy in the upper frequencies and of harmonic strength. Her WOEG showed a steep drop of energy, similar to S6's WOEG, but with higher harmonicity than S6.

The intra-vocalic changes observed for S3, which do not follow what is observed for Irritation in her case, may be linked to articulatory differences between the two speakers (at a phonemic level): typically, a potential nasal coda may systematically veil the higher energy at the end of the vowel. This is not observed on other parameters, that shall be less sensitive to articulatory variations.

One interpretation of these differences in voice quality estimates is that S3 tends to use a more modal voice (more energy in the upper frequencies with stronger harmonics) than S6, which uses a more falsetto-like voice. This interpretation is also borne out by the auditory impressions of the professional speaker who produced the exemplar sounds for the auditory perception test. As will be discussed later, the speakers used of different voice registers, i.e., modal vs falsetto, also affects the social code they use in their social affective expressions.

3.2 Perception tests: VAD

Perceptual evaluation of the V, A, and D dimensions on a scale of 1 to 5 were standardized (centered and scaled) for each listener and each scale to remove individual changes in using these scales. The variation in these z-scores was evaluated with one linear model for each scale with, as independent variables, the 6 attitudes and the two speakers plus their interaction (note the two speakers are not modeled here as a random variable, as this dataset is reduced: we do not try to extrapolate their production characteristics to a more general population but look at how these specific expressions were perceived). The analysis of variance output for each scale is presented in Table 4; post-hoc comparisons were run to compare the levels of each attitude using a Bonferroni correction.

Table 4: ANOVA tables of the linear models on the V, A, D scales, with factors attitude (Att) and speaker (Spk) and their interaction; the F value, its degrees of freedom, and associated p-value and η2 are reported. (from Erickson et al. 2022)

Valence	df	F	p	η2
Att	5	23.697	0.000	0.407
Spk	1	0.001	0.971	0.000
A:S	5	3.899	0.002	0.067
Residuals	153	–	–	0.526
Arousal	df	F	p	η2
Att	5	22.543	0.000	0.401
Spk	1	5.106	0.025	0.018
A:S	5	1.807	0.115	0.032
Residuals	154	–	–	0.547
Dominance	Df	F	p	η2
Att	5	32.026	0.000	0.470
Spk	1	6.350	0.013	0.019
A:S	5	4.422	0.001	0.065
Residuals	152	–	–	0.447

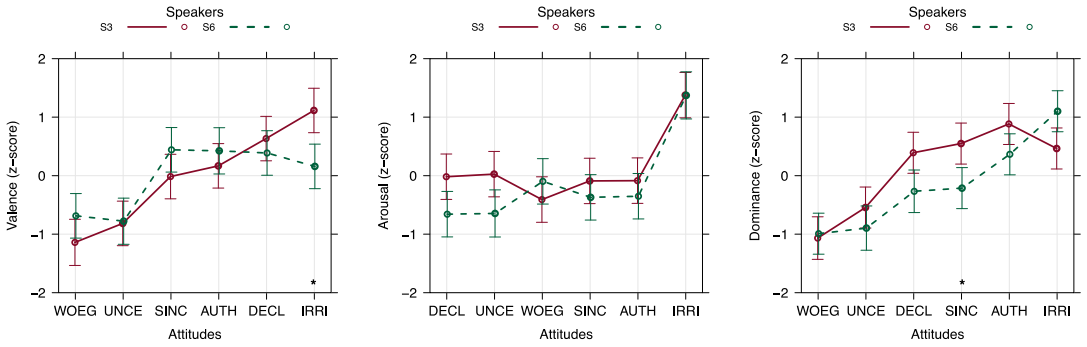

Figure 6: Three plots from left to right for negative-positive (Valence), calm-excited (Arousal), and accommodating-assertive (Dominance) ratings: mean and confidence intervals of for the 6 attitudes (ranked in increasing mean levels of each scale) for speakers S3 and S6. The stars (*) just above the horizontal axis indicate significant post-hoc pairwise comparisons. (from Erickson et al. 2022a)

Figure 6 shows the VAD ratings by listeners for the six affective expressions and the two speakers. In terms of Valence, WOEG and Uncertainty by both speakers were rated as significantly more negative than all the other ones, which are comparable in valence level – but with a significant difference between S3 and S6 valence for Irritation (S3's production receiving higher ratings). As for Arousal, most of the affective expressions were rated as calm (ratings below the mean) but, for both speakers, productions of Irritation were rated as very excited (significantly higher than all the others); the linear model also showed higher arousal perceived for S3 than S6; the interaction was not significant. As for Dominance ratings, the highest ratings of "assertive" are for both speakers' Authority and Irritation. In addition, we see Sincerity and Declaration rated as dominant for S3 but less for S6, and WOEG and Uncertainty significantly lower in Dominance ratings for both speakers.

We note an SAE is not fixed in terms of valence, dominance, or arousal. "Surprise, for example, can result from unexpected positive/pleasant news or from negative indignation; and, likewise, irritation can be expressed submissively (low dominance) or accusatory (high dominance), depending on who is assumed to be the culprit" (p.c. by one anonymous reviewer). In our design, we tried to control for this by specifying the hierarchical relationship between the

dialogue partners, e.g., social relationship, age, and location of dialogue. To be sure, however, the dialogue scenarios could have been more detailed. For instance, the speaker's personality, e.g., extrovert, neurotic, etc., is also known to affect the voice characteristics (see Erickson et al. 2018).

3.3 Perception tests: Auditory assessment of voice qualities

Answers for each of the six voice quality scales were standardized by listener (so to remove potential individual variation in using each scale). These z-score values were then submitted, for each scale, to linear modeling having as independent variables the six attitudes and the two speakers, plus their interaction. The output of these analyses of variance is presented in Table 5. Subsequent planned post-hoc comparisons of the between-speaker differences for each attitude were run with a Bonferroni correction; results are presented in Figure 7, with each significant pair differences being marked by a star (*).

Table 5: ANOVA tables of the linear models on the voice quality scales (Breathy, Falsetto, High, Loud, Tense, Twang), with factors attitude (Att) and speaker (Spk) and their interaction; the F value, its degrees of freedom, and associated *p*-value and η^2 are reported.

Breathy	df	F	*P*	η^2	Loud	df	F	*P*	η^2
Att	5	27.792	0.000	0.379	Att	5	31.824	0.000	0.382
Spk	1	7.257	0.008	0.020	Spk	1	1.982	0.161	0.005
A:S	5	3.851	0.002	0.053	A:S	5	10.239	0.000	0.123
Residuals	201	–	–	0.549	Residuals	204	–	–	0.490
Falsetto	df	F	*P*	η^2	Tense	df	F	*P*	η^2
Att	5	44.037	0.000	0.453	Att	5	25.262	0.000	0.348
Spk	1	0.733	0.393	0.002	Spk	1	2.422	0.121	0.007
A:S	5	12.443	0.000	0.128	A:S	5	6.204	0.000	0.086
Residuals	203	–	–	0.417	Residuals	203	–	–	0.560
High	df	F	*P*	η^2	Twang	df	F	*P*	η^2
Att	5	2.747	0.02	0.054	Att	5	19.596	0.000	0.289
Spk	1	2.303	0.131	0.009	Spk	1	0.987	0.322	0.003
A:S	5	7.252	0.000	0.142	A:S	5	7.129	0.000	0.105
Residuals	203	–	–	0.795	Residuals	204	–	–	0.602

The output of the analyses of variance (Figure 7) shows a globally coherent strategy of both speakers in their use of most voice quality scales: the majority of between-speaker-pairs was not significantly different. In particular, they are always similar for their expression of Declaration – or the reference neutral speaking way, which generally is found close to the middle of each scale (i.e., here close to 0 z-score). They also have similar voice quality judgments for WOEG. Observed differences in the voice qualities of their nuclear vowels were for Sincerity for breathiness, pitch, and tension judgments – S3 being lower than S6 in each case. For Authority, S3 is also lower than S6 for the falsetto, loudness, tension, and twang scales. In the case of Irritation, S3 was higher than S6, but less falsetto and less loud. For Uncertainty, S3 shows a lower pitch and a voice judged as louder, more twang and more falsetto.

Table 6 gives a qualitative description of the auditory perceptions of voice qualities for each expression; finer differences between the two speakers can be seen by examining the mean values displayed in Figure 7. In the table, auditory ratings above 0 are considered as having a certain quality, below 0 as having the opposite quality, and values at or close to 0 are described as "not that quality", i.e., "not breathy." More specifically, the breathy scale is qualitatively

described as breathy vs not breathy; the falsetto scale is described as falsetto vs modal; the high scale is described as high voice vs low voice; the loudness scale, as loud vs soft; the tense scale, as tense vs. lax; the twang scale, as twang vs no twang.

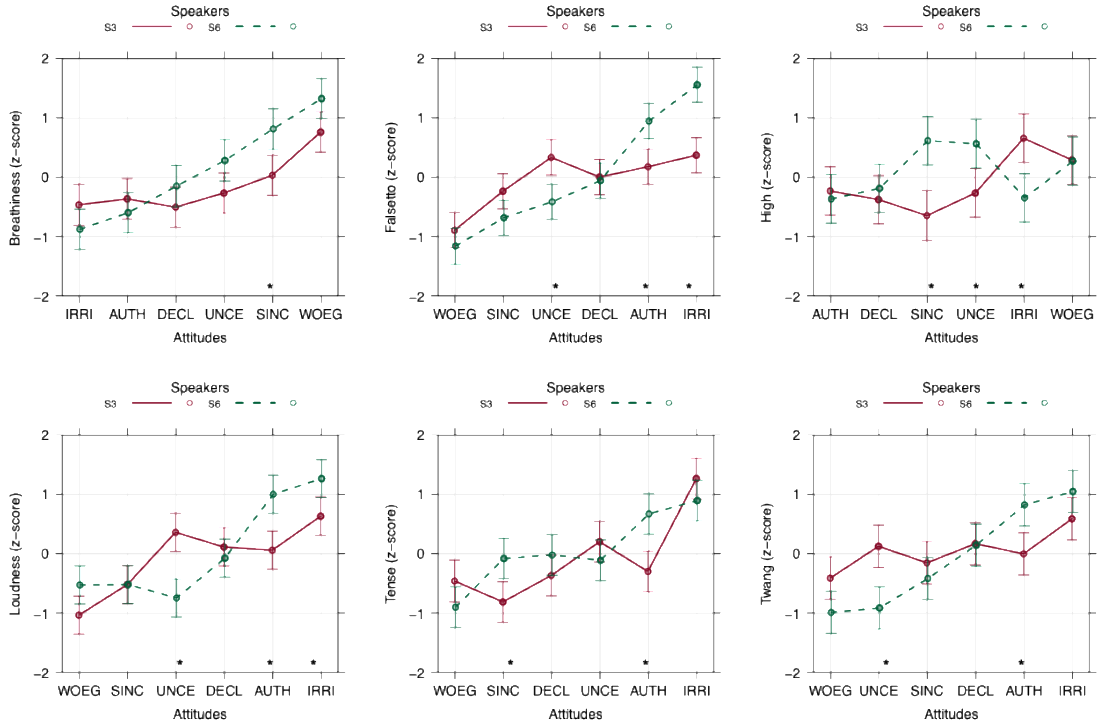

Figure 7: Six plots for the perceptual VQ scales; from left to right, top row: Breathy, Falsetto, High; bottom row: Loud, Tense, Twang. Plots present the mean and confidence intervals predicted by the linear models, with the 6 attitudes along the horizontal axis (ranked in increasing mean levels of each scale) and speakers S3 and S6 on separate lines. The stars (*) just above the horizontal axis indicate significant post-hoc pairwise comparisons.

Table 6. Voice quality summary for each of the expressions, based on the MFA in Table 3.

Authority	Declaration	Irritation	Sincerity	Uncertainty	WOEG
Not breathy or high voice; for S6, high tenseness, falsetto, loudness, twang.	Not breathy, or high voice; slightly lax; not falsetto, loud or twang. Speakers very similar.	Tense, loud and twang; for S3, high voice; for S6, lots of falsetto.	Soft, not tense, twang or falsetto; for S3, higher, more lax; for S6, breathy.	Not tense or twang; for S3, slightly low, loud & falsetto; no breathiness; for S6, high, less falsetto & breathy.	Very breathy, soft, modal, not especially high, lax, no twang. Speakers very similar.

3.4 Relations between Prosodic characteristic and VAD scales

The mean score assigned by listeners to each Valence, Arousal, and Dominance scale (expressed in z-scores) was organized in a table, with rows as the twelve utterances performed by the two speakers, and as columns, the perceptual results. The mean and standard deviation of acoustic measures on the utterances' vowels were added in further columns: F0, intensity, Hammarberg index, CPP, NAQ, and peak slope. This two-part table, summarizing perceptual and acoustic aspects of the complete utterance, was then submitted to a Multiple Factorial Analysis (MFA: see

details in the Methods part and in Husson et al. 2017). The MFA allows the comparison of several datasets, keeping the characteristics of each one.

Only the first three axes of the MFA were kept, that explain about 75% of the total variance. The coordinates, contributions, and squared cosines of the columns and of the rows are presented in Table 7. The distribution of the utterances along the first two main axes are represented on Figure 8, distinguished for the clusters they were attributed to, so to propose a visual support for the spread obtained with such parameters (please note the third dimension is not taken into consideration with such a representation).

Table 7: Coordinates (D), contributions (Ct), and squared cosines (cs, multiplied by 100 for convenience) of each row of the MFA on utterance perceptual and acoustic data, for the three main dimensions (1, 2, 3)

	D1	D2	D3	Ct1	Ct2	Ct3	cs1	cs2	cs3
A	0.25	0.58	0.07	2.7	25.5	0.8	14	79	1
V	0.60	0.07	0.03	15.9	0.3	0.2	81	1	0
D	0.61	0.18	0.22	16.8	2.6	9.1	73	7	9
F0 m	0.40	0.48	-0.59	1.7	4.2	16.0	16	23	35
F0 sd	-0.29	0.69	-0.11	0.9	8.7	0.5	9	47	1
SPL m	0.86	0.23	-0.24	8.0	1.0	2.6	74	5	6
SPL sd	-0.28	0.56	-0.01	0.8	5.9	0.0	8	32	0
CPP m	0.80	0.00	0.46	6.9	0.0	9.8	64	0	21
CPP sd	-0.58	0.13	0.43	3.6	0.3	8.4	33	2	18
NAQ m	0.60	0.57	-0.23	3.9	6.1	2.6	36	33	6
NAQ sd	0.13	0.73	-0.46	0.2	9.7	9.8	2	53	21
PS m	-0.38	0.55	-0.36	1.5	5.6	6.0	14	30	13
PS sd	-0.58	0.50	0.00	3.7	4.7	0.0	34	25	0
HI m	-0.71	-0.49	-0.04	5.5	4.4	0.1	51	24	0
HI sd	0.32	0.37	0.74	1.1	2.5	25.1	10	13	54
Dur. total	-0.73	0.47	0.04	6.7	4.7	0.1	53	22	0
Dur. syll. m	-0.68	0.52	0.27	5.9	5.9	4.1	46	27	7
Dur. syll. sd	-0.75	0.41	0.16	7.2	3.6	1.4	56	17	3
Dur. stress	-0.73	0.45	0.25	6.9	4.5	3.5	54	21	6
AUTH S3	1.00	-0.30	1.00	3.5	0.7	16.5	32	4	35
DECL S3	1.30	-0.80	0.70	6.6	3.8	8.4	46	16	14
IRRI S3	0.90	2.20	0.50	3.3	31.3	4.1	11	61	3
SINC S3	0.30	-0.90	0.50	0.4	4.8	3.5	5	31	9
UNCE S3	-2.40	1.20	0.30	21.9	9.3	1.4	65	16	1
WOEG S3	-2.60	-0.10	0.40	25.8	0.1	3.0	81	0	2
AUTH S6	1.50	-0.80	0.00	7.9	3.9	0.0	52	15	0
DECL S6	0.70	-1.60	-0.20	1.6	15.6	0.9	12	70	2
IRRI S6	1.70	2.00	-0.50	10.3	26.5	4.5	32	48	3
SINC S6	0.70	-0.30	-1.70	1.8	0.5	43.6	12	2	65
UNCE S6	-1.60	-0.70	-0.10	9.5	3.4	0.1	59	13	0
WOEG S6	-1.40	0.00	-0.90	7.4	0.0	14.2	37	0	17

The first axis of the MFA is positively correlated to the Valence and Dominance scales. It is also positively related to the mean of intensity, CPP and NAQ; negative correlations with the three duration measurements were found, as well as with the mean spectral slope (Hammarberg index) and standard deviation in peak slope and CPP. This axis has high loadings on the WOEG and Uncertainty expressions (negative side). The second axis is correlated to Arousal judgements, and to high standard deviations of NAQ and F0; it is related to the expression of Irritation. The third axis opposes the performances of both speakers, based on the parameters of Hammarberg

index (positive correlation) and F0 mean (negative): S3, on the positive side of the axis, tends to have a louder and lower voice than S6.

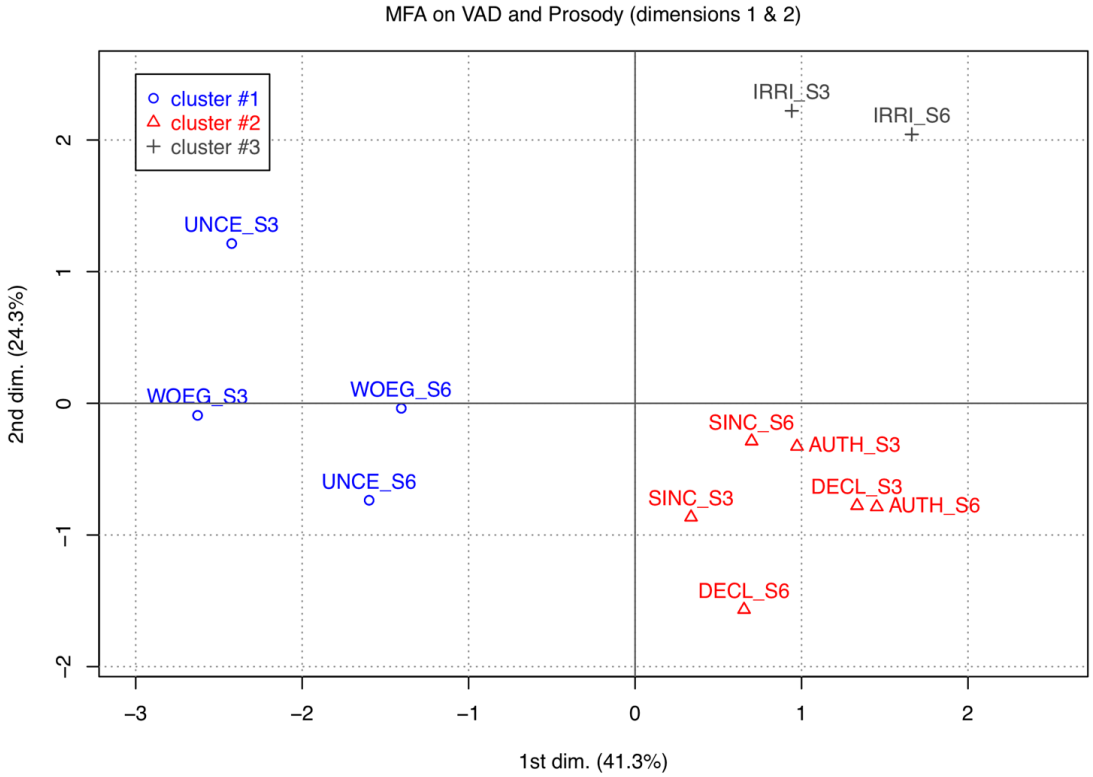

Figure 8: Position of each utterance along the first two main axes of the MFA on acoustic and VAD measurements; the shape and colors of the point indicates the cluster it was attributed to.

From the distribution of the twelve utterances on this three-dimensional space, a hierarchical agglomeration algorithm was applied that leads to a three-cluster solution according to a criterion of inertia gain. These clusters are described hereafter.

Cluster #1: based on the Uncertainty and WOEG expressions, it is characterized by duration parameters (longer utterances, with irregular rhythm mixing short/ long syllables and hesitations/pauses), steep spectral slopes (Hammarberg index), and low mean intensity and weak noisy harmonics (CPP). These expressions were both perceived as having negative Valence and Dominance.

Cluster #2: based on utterances expressing Authority, Declaration, and Sincerity by both speakers, they were shorter with comparatively regular rhythm (no large deviations in syllable duration and relatively no pauses/hesitations) and relatively flat (non-dynamic) F0 patterns of intonation (small standard deviation of F0).

Cluster #3: regroups the two expressions of Irritation; it was characterized by high Arousal judgments, non-steep (sustained) spectral slopes, high intensity, and high mean NAQ with large variations of NAQ. That is, Irritation, which was judged as aroused, has high energy in the upper frequencies, high intensity, and the glottal closing behaviors are those associated with a tense voice, but with high irregularity in glottal closing behavior.

To summarize the findings for this case study, a soft / noisy voice, with weak harmonics and irregular rhythm with pauses and hesitations, as in the expressions of Uncertainty and WOEG, was perceived by listeners as accommodating (not Dominant) and not positive (negative Valence).

Loud, tense voices with energy in the upper frequencies, as in the expression of Irritation, are perceived as Aroused. Expressions of Authority, Declaration, and Sincerity tended to have a comparatively regular rhythm and relatively flat intonation.

Note that the term "tense" is used in this paper to refer to a number of different acoustic and auditory parameters: high NAQ and high Peak slope values were designed to indicate a tense voice, but the former refers to tense voice as a function of glottal closing behavior, the latter to tense voice as a function of spectral configuration. In addition, the term tense voice is used as an auditory percept. Future work is needed to examine in more detail how these terms relate to actual voice production strategies.

3.5 Relations between VQ measurements and perception

The mean score assigned by listeners to each voice quality scale (expressed in z-scores) was organized in a table, with rows as the twelve utterances performed by the two speakers, and as columns, the six perceptual scales. Other columns with acoustic measures corresponding to the vowel with nuclear stress were added, estimating the median value of the following parameters on each third of the vowel: F0, intensity, Hammarberg index, CPP, NAQ and peak slope. These two parts of the table, which summarize respectively perceptual and acoustic aspects of the nuclear stress vowel, were then submitted to a Multiple Factorial Analysis. The first three main axes of the MFA were kept, as they explain more than 85% of the total variance. The coordinate, contribution, and squared cosines of the columns and of the rows are presented in table 8. The spread of the twelve utterances along the first two main axes of the MFA is also presented in Figure 9, with each utterance being attributed one of the five clusters.

Table 8: Coordinates (D), contributions (Ct), and squared cosines (cs, multiplied by 100 for convenience) of each row of the MFA on nucleus perceptual and acoustic data for the three main dimensions (1, 2, 3)

	D1	D2	D3	Ct1	Ct2	Ct3	cs1	cs2	cs3
Breathy	-0.58	0.25	0.01	9.0	5.5	0.0	81	15	0
Falsetto	0.70	-0.09	-0.20	13.1	0.7	4.3	90	1	7
High	-0.12	0.30	0.25	0.4	8.0	7.1	7	47	34
Loud	0.63	0.05	-0.18	10.7	0.2	3.8	84	1	7
Tense	0.56	0.20	0.11	8.6	3.7	1.4	79	10	3
Twang	0.56	-0.07	-0.12	8.4	0.4	1.5	85	1	4
F0 1	0.86	-0.07	0.30	4.7	0.1	2.4	74	1	9
F0 2	0.88	0.21	0.30	4.9	1.0	2.4	77	4	9
F0 3	0.85	0.35	0.28	4.6	2.6	2.0	72	12	8
Int 1	0.37	0.52	-0.42	0.9	5.8	4.7	13	27	18
Int 2	0.76	0.32	0.16	3.7	2.1	0.7	58	10	3
Int 3	0.73	0.37	0.31	3.4	2.9	2.4	54	14	9
CPP 1	0.80	-0.33	0.15	4.1	2.3	0.6	64	11	2
CPP 2	0.75	-0.56	-0.16	3.6	6.7	0.6	56	31	2
CPP 3	0.69	-0.66	0.00	3.0	9.3	0.0	47	43	0
NAQ 1	0.55	0.38	0.46	1.9	3.1	5.6	30	15	21
NAQ 2	0.43	0.46	0.62	1.2	4.5	9.9	18	21	38
NAQ 3	0.32	0.77	0.33	0.6	12.5	2.9	10	59	11
PS 1	0.00	0.79	-0.44	0.0	13.2	5.1	0	62	19
PS 2	-0.04	0.59	-0.72	0.0	7.5	13.5	0	35	51
PS 3	-0.45	0.40	-0.77	1.3	3.4	15.6	21	16	59
Ham 1	-0.89	0.27	0.07	5.1	1.6	0.1	80	7	0
Ham 2	-0.84	-0.23	0.42	4.5	1.2	4.7	70	5	18

Ham 3	-0.56	-0.28	0.57	2.0	1.6	8.6	32	8	33
AUTH S3	0.35	-0.73	0.35	0.5	7.8	2.2	9	41	9
DECL S3	-0.10	-1.07	0.14	0.0	16.7	0.3	1	67	1
IRRI S3	1.76	1.09	1.39	13.6	17.4	34.6	47	18	29
SINC S3	-0.84	-0.76	-0.38	3.1	8.4	2.6	36	29	7
UNCE S3	0.35	-0.69	-0.08	0.5	6.9	0.1	13	52	1
WOEG S3	-1.81	-0.32	0.05	14.4	1.5	0.1	82	3	0
AUTH S6	1.63	-0.05	-0.56	11.7	0.0	5.6	76	0	9
DECL S6	0.27	-0.27	0.19	0.3	1.1	0.6	11	11	5
IRRI S6	2.33	0.55	-1.33	23.9	4.4	31.9	70	4	23
SINC S6	-0.71	0.87	0.31	2.2	11.2	1.7	25	39	5
UNCE S6	-0.75	0.09	0.71	2.5	0.1	9.1	33	0	30
WOEG S6	-2.49	1.29	-0.79	27.2	24.3	11.2	70	19	7

The first axis of the MFA is correlated positively to the Falsetto, Twang, Loud, and Tense scales and Negatively to the Breathy scale. The perception of High voice is correlated to the second and third axes.

The first axis is positively linked to high F0 and CPP values, and high intensity values (on the two last measurement points of the vowel), and negatively to steep spectral slopes (large negative Hammarberg index values on the first two measurement points). The second axis shows a positive correlation with Peak Slope (points 1 and 2) and NAQ (point 3) values, and a negative correlation to the third CPP measure – while the third axis has a positive correlation to the second NAQ measurement point (mid-vowel), and negative one to peak slope (points 2 and 3). The second axis is mostly related to "modal" voices (Declaration and Uncertainty for both speakers), while the third opposes S3 and S6 performances of Irritation.

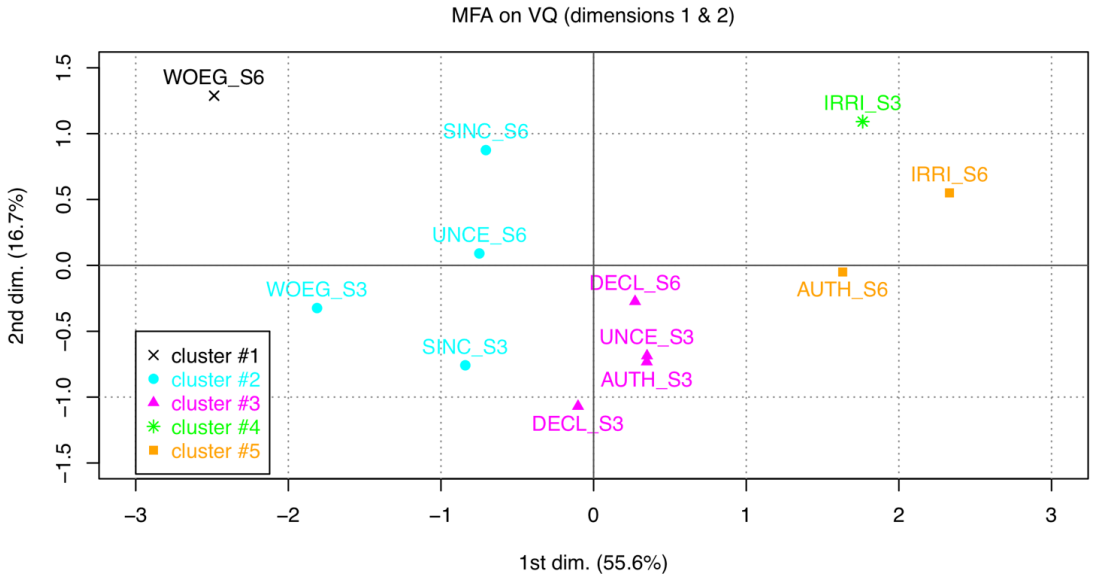

Figure 9: Position of each utterance along the first two main axes of the MFA on acoustic and voice quality measurements; the shape and colors of the point indicate the cluster it was attributed to.

From the distribution of the twelve utterances on this three-dimensional space, a hierarchical agglomeration algorithm was applied that leads to a five-cluster solution according to a criterion of inertia gain. These clusters are described hereafter.

Cluster #1: based on S6 WOEG. Perceived as Breathy, it has small peak slope values, low F0, and low CPP values; that is, WOEG is produced with a lax voice, low F0, and weak, noisy harmonics, which is coherent with a breathy perception.

Cluster #2: based on Sincerity, S3 WOEG, and S6 Uncertainty; the expressions were perceived as non-Loud (soft) and have low intensity values (3rd point).

Cluster #3: based on Declaration, and S3 Uncertainty and Authority; it has a steeper peak slope than the mean observed value – i.e., less energy in the upper frequencies.

Cluster #4: based on S3 Irritation; perceived as tense, it is characterized by high NAQ throughout the vowel, and high final F0 and intensity; that is, S3 Irritation, perceived as tense, is produced with glottal closing behavior associated with tense voice throughout the whole vowel, high F0 and high intensity.

Cluster #5: based on S6 Irritation and Authority; perceived as Falsetto, Loud, Twang, it has shallow, non-steep spectral slopes (small Hammarberg index values); that is, S6 Irritation and Authority are produced with relatively high energy in the upper frequencies

To summarize the findings for this case study, WOEG, perceived as breathy, has the acoustic characteristics of low F0, low peak slope values, and low CPP values. That is to say, breathy voices do not have much energy in the upper frequencies, and the harmonics are weak compared to noise. The expressions that that are perceived as soft, with low intensity values are Sincerity and S3's WOEG and S6's Uncertainty. The expressions with less energy in the upper frequencies are Declaration, and S3's Uncertainty and Authority. Irritation, produced by S3, was perceived as tense, acoustically characterized by high intensity, high final F0, and high NAQ throughout the vowel. Thus, it seems that the auditory impression of voices with high NAQ measurements is indeed tenseness; moreover, we see that the acoustic measure of tenseness due to glottal closing behavior (NAQ) corresponds with listeners' auditory perceptions of tenseness. However, S6's Irritation, as well as Authority, were perceived as loud, falsetto, and twang, and are acoustically characterized as having relatively high energy in the upper frequencies, but the MFA analysis does not refer to either tenseness or NAQ in describing S6's Irritation. It is interesting that perceived Falsetto was produced with high energy in the upper frequencies, but yet, these two utterances were also perceived as being loud and twangy. Increased loudness is characteristic of twang, due to increased energy in the 2.5-3kHz area as a result of narrowing the pharyngeal area, similar to the production of the singer formant (e.g. Erickson et al. 2020b). Also, there are reasons to believe that twang is usually produced in falsetto (thin folds) voice, rather than modal (thick folds) voice since too much pressure, when coupled with thick folds, may invite constriction (p.c. Kerrie Obert). Notice that the Irritation expressions by both speakers were heard as loud. Thus, a possible interpretation is that S3, who tends to use a more modal voice (with thicker vocal folds), uses tenseness (changes glottal closure configuration) to increase loudness to express Irritation, while S6, who tends to use a more falsetto voice (with thinner folds), changes vocal tract configuration to amplify the loudness when expressing Irritation.

4 Discussion

This fine-detail analysis of the voice quality of two speakers uttering a set of six social affective expressions brings a set of information on (1) the perception of general dimensions linked to speakers' expressivity and their interpretation along the VAD dimensions, and (2) the susceptibility of acoustic cues to rapid changes and their adaptation to each speaker's individual physiological characteristics and voice habits, due to a variety of factors, such as their speaking styles, personalities, interpersonal experiences, etc. Thus, voice quality acoustic measurements can reveal important variations across speakers. Even if a global evaluation of these two speakers

showed they were comparably efficient in expressing these prosodic attitudes, they did so by selecting parts of the voice quality cues in an idiosyncratic way.

The parts of the acoustic measurements that are comparable in the two speakers strategies are primarily linked to the opposition between strong and weak voices, an opposition that fits the Arousal aspects of vocal emotional expression (Banse and Scherer 1996; Goudbeek and Scherer 2010) as a major dimension of voice change; it also goes along with the notion of vocal effort (Liénard and Di Benedetto 1999; Traunmüller and Eriksson 2000; Liénard 2019) and the adaptation of one's voice loudness to a given situation (this includes individual voice settings or voice registers). Along this dimension, low or weak voices have low energy, a steeper spectral slope, and will tend to have higher additive noise linked to breathy phonation and lower harmonic energy; on the contrary, strong voices have higher overall energy, shallow, flatter spectral slopes, and the difference in energy levels in the harmonic and non-harmonic parts of the spectrum is higher (i.e., high CPP values, the voice is not breathy).

Works by Titze and colleagues have long shown that changes in subglottal pressure lead to variation in the voice's fundamental frequency (Titze 1989): as increasing the subglottal pressure is a main way to increase the vocal intensity, and raising one's vocal effort (especially in a large amount, as can be found at each end of the "weak-strong" voice quality axis observed here) will increase the voice F0. It may be said for the reverse: in order to increase F0 to a great extent, a speaker will have to use a louder voice, i.e. higher intensity levels (Titze and Sundberg 1992; Lamesch et al. 2012). But the perceptual effects of a speaker trying to raise F0, and hence the voice's pitch (as a perceptual quantity), are not the same as those of raising its intensity or effort – if in both cases, increases of F0 and intensity are observed. This is illustrated in the two speakers by their varying strategies for the spread of their mean intensity and F0 values across attitudes. Both speakers seem to have different voice settings, with S3 showing lower mean intensity and F0 values, with F0 changes relatively independent of the intensity level, while S6 has higher intensity and F0 ranges, and shows a mostly linear change of F0 with intensity.

For both speakers, Irritation has the highest F0 and intensity, Authority having the second highest intensity, but an intermediate F0 level. For speaker S6, the increase in intensity and F0 follows a quasi-linear relation: for Irritation she increases her vocal effort (in terms of mean intensity over the complete sentence) of 6.8 dB above the mean reference level observed for Declaration, reaching a high F0 level (5.3 st above the mean reference level) – but even with such an F0 increase, S6's nuclear stress vowel was rated as having a low pitch (her lowest pitch in fact, comparable to the pitch of Authority). Results are similar for S6's Authority, perceived as loud, twangy, and falsetto, but not as high pitched. Speaker S3's voice quality ratings of these two expressions differ, with her Authority being rated as relatively loud but not high pitched (as seen for S6), while the expression of Irritation is even louder but was also perceived has high pitched (unlike for S6). In terms of VAD dimensions, Irritation which is consistently loud across speakers was given high Arousal levels for both speakers, but S3's performance was rated as more positive than S6's – while for their Authority performances, both are relatively not Aroused (negative z-score), with a neutral Valence, but are relatively Dominant: S3's Authority is her more Dominant expression, S6's Authority being her second most Dominant after Irritation. Such results, with loudness or vocal effort correlated to judgements of Arousal and Dominance, with yet the intensity-independent rise of F0 being judged as positive (for S3's Irritation), confirms the predictions of Gussenhoven's Effort Code. This code predicts high Effort for important speech acts, such as in imposition and dominant expressions (Gussenhoven 2004) – here Authority and S6's Irritation.

Observe that the strategy used by S3 for Irritation resorts primarily to a rise of F0 not driven by intensity (with a high peak on the nuclear stress), and is perceived with a positive Valence. In

the study by Goudbeek & Scherer (2010), they did not find the combination of high arousal with high F0 as having positive valence; instead, they reported for high levels of arousal, "positive emotion portrayals [had] less Intensity variation and a steeper negative slope than negative ones" (Goudbeek and Scherer 2010, p. 1334). The comparatively steeper spectral slope of aroused positive affect compared to aroused negative ones may signal a higher perceived pitch for these positive expressions – a fact that would match S3's Irritation vocal strategy. Some comments about differences between our corpus and that analyzed by Goudbeek and Scherer: their corpus was the large Geneva Multimodal Emotional Portrayals corpus, with ten actors portraying 18 different emotional states; ours was small, based on two female speakers, six types of social affective expressions unevenly spread along the Valence Arousal Dominance dimensions, not on emotional portrayals by actors. As such, our corpus did not encompass the variety of voice quality possibilities that a larger database does, such as e.g., hoarse, rough, creaky voices, etc. Moreover, the VAD descriptions by Goudbeek and Scherer were based on assumed VAD qualities of emotions, not on empirical grounds of listeners' ratings, nor on non-acted social affective expressions, typical of daily ordinary conversation situations. Finally, the acoustic measurements in Goudbeek and Scherer were extracted from the entire utterance, not just the vowels, as was done in our study.

Along these lines, our study illustrates that the voice quality characteristics of the nuclear stress vowel can differ from the prosodic estimates on all the vowels in the utterance. This is seen especially for the expressions of Irritation by S3, who had a relatively non-high F0 for the Irritation utterance as a whole but a high F0 for the nuclear stress vowel, as discussed above.

At the low end of the intensity values are, for both speakers, the walking-on-eggs and Uncertainty attitudes. For S6, F0 values again follow these intensity changes, with both attitudes being her expressions with low intensity. For S3, WOEG has her lowest energy and a relatively low F0, while Uncertainty has a sustained F0 (at a comparable level with S3's Declaration); for this speaker, the lowest F0 is found for Sincerity, an expression with an intensity level comparable to Declaration. This low end of energy illustrates again the differences in vocal strategy between the two speakers, with S3 showing F0 variations independent of intensity, while the covariation of both parameters is strong for S6. In perceptual terms, and for both speakers, WOEG and Uncertainty are rated at the lowest end of the Dominant and Valence scales: low effort (intensity) seems to drive these perceptual judgments. WOEG is also remarkably comparable across speakers in terms of the vocal quality of the stressed vowel – being judged for both as breathy and having low levels of falsetto, loudness, and twang; similar judgments were found for S6's Uncertainty, while the S3 one was not breathy but relatively loud and twangy. An interesting point is the link with the perception of highness in WOEG: it has low F0 values for both speakers, but it was rated as having a relatively high pitch (positive z-score). The link between submissive behavior and high pitch is typical of the prediction of Ohala's Frequency Code (Ohala 1994). In this case, the link was not done on the basis of the fundamental frequency measurement (that shall be corrected for the very low effort), but on the perceived pitch of the two voices.

The Sincerity expression was also performed with varying strategies by both speakers. While they both have intensity levels close to their Declaration, they are opposed in F0 levels: S3 has her lowest voice (in terms of F0 and perceived pitch), while S6 Sincerity received the highest pitch values (for F0 near her mean values), and her highest Valence. The link between pitch (as a perceived judgment) and Valence is again confirmed (S3 highest pitch and Valence being found for Irritation). One thought here: for S3, her strategy for Sincerity may have been to convey Sincerity about the "truth" that "Mary was dancing", and hence a more assertive voice; for S6, her strategy for Sincerity may have been more similar to Sincerity politeness, as defined by Shochi

et al. (2023) as expressing an "honest, respectful feeling toward the hearers"; hence, a more accommodating voice.

To summarize our findings about Valence, Arousal, and Dominance in comparison with previous studies—the expression with the most positive valence was S3's Irritation, characterized acoustically as having a relatively high mean F0 over the complete utterance, and the highest F0 peak on the nuclear stress; it also received high NAQ measurements, but modest intensity levels compared to S6's Irritation. S6's most positive expression is Sincerity, which was rated as high pitched for her voice, with a lower arousal level than S3's Irritation. These two positive and high pitched expressions differ in terms of Arousal and Dominance – thus the importance of Arousal in the interpretation of the other dimensions, already shown by Goudbeek and Scherer (2010). The Arousal dimension seems to be mostly linked to the vocal effort, and this appears to be fairly robust even on a reduced dataset. Finally, the Dominance dimension is related to perceived low pitch – a prediction of the Frequency Code, even if that does not systematically correlate with measured fundamental frequency.

The acoustic study of acted Mandarin emotions (e.g., anger, disgust and happiness) by Liu and Xu (2014) is mentioned here. They interpreted their findings of high F0 for both anger and happiness as possible lack of support for the Frequency Code. Our SAE study does not include happiness or anger; however, our finding was that one speaker raised F0, not lowered F0, to express Authority and Irritation, while the other speaker lowered F0. Both speakers increased intensity. We use these examples to discuss how both the Frequency Code and the Effort Code are compatible: A speaker with a modal voice may choose to use the Frequency Code, i.e., lower the voice to express dominance, while a speaker with a falsetto voice, may increase the range of F0 to express dominance, thus using the Effort Code.

4.1 Description of common behavior in terms of ToBI labels.

Social affective expressions are used to convey specific pragmatic meanings, e.g., "I am irritated." or "I am being polite." etc., and the acoustic (and visual cues) used to convey these meanings are somewhat stereotypical according to the cultural norms (see e.g. Rilliard et al. 2013, 2017). As such, we wonder if perhaps the pragmatic framework proposed by Pierrehumbert and Hirschberg (1990) might be relevant to account for the pitch accents found for the six affective expressions by these two American English speakers. According to them, "a speaker (S) chooses a particular tune to convey a particular relationship between an utterance, currently perceived beliefs of a hearer or hearers (H), and anticipated contributions of subsequent utterances." (Pierrehumbert and Hirschberg 1990, p. 271). In other words, the choice of pitch accent depends on (a) the specific utterance, (b) the currently perceived beliefs of the listener/s, and, (c) the anticipated contribution of subsequent utterances. They also state that stressed L and H tones convey different pragmatic meanings in terms of what the speaker intends the hearer to know. Simply put, a stressed L tone conveys that the speaker, for various reasons, does not feel the accented item should be added to the hearer's belief. The authors write "when the starred tone is L (L*, L*+H, and H+L*), S does not convey that the instantiation of the open expression by the accented item should be added to H's mutual beliefs. For one of a variety of reasons–it may already be there, S may not be certain of its appropriateness. S may not wish or be able to predicate the open expression of the accented item--S does not intend to contribute this instantiation to H's mutual beliefs." (Pierrehumbert and Hirschberg 1990, p. 301). On the other hand, a stressed H tone conveys to the hearer that this information is to be added to the hearer's beliefs, i.e., "When the starred tone is H (H*, L+H*, H*+L), S does intend to instantiate the open expression in H's mutual belief space" (Pierrehumbert and Hirschberg 1990, p. 301).

In our data, both S3 and S6 produce the prenuclear contour on Mary with an L*+H pitch accent. Given that in the dialogue situation, the speakers were asked "What was Mary doing?", it was clear to the speaker that the hearer (interviewer) already knew that Mary was the subject, and hence they may have used a starred L pitch accent to convey that the hearer already knew about Mary. S3, however, used an H* pitch accent on Mary, perhaps making sure the hearer knew that it was Mary who was dancing.

As for the nuclear contour on dancing, both speakers used the L* pitch accent for expressions of Declaration, Sincerity, and WOEG—perhaps to convey a lack of strong feeling for adding this information to the hearer. In our data, both speakers employ !H* to indicate a stressed syllable that is neither low nor high and is used to express Uncertainty. (See also Thorson and Burdin, 2022, for more discussion of the !*H tone being a tone with a smaller F0 drop.)

Interestingly, both speakers mark Irritation with an L*+H pitch accent on the nuclear stress syllable. The L*+H pitch accent is reported by Ward and Hirschberg (1985) to be used frequently to express Uncertainty. Perhaps in our data, the starred L was used because the interviewer asked the speaker three times, "What was Mary doing?" and the speaker was uncertain that the listener had heard her the first time; hence the L*+H pitch accent.

As for starred H tones, according to Pierrehumbert and Hirschberg (1990), H*+L may frequently be used to convey pedagogical type information. In our data for Authority, both speakers use the H*+L pitch accent on the nuclear stress syllable as if to convey "this is new information, please pay attention." S3 also uses this for Declaration for the nuclear contour on, while S6 uses the L* pitch accent. Interestingly, listeners heard S3's Declaration (with an H* on Mary) as significantly more Dominant than that of S6 (with an L*+H on Mary).

To summarize, the two speakers in our case study produced the following nuclear stress pitch accents on the following affective expressions: H*+L for Authority; L*+H for Irritation; L* for Declaration, Sincerity, WOEG; and !H* for Uncertainty.

However, we wish to strongly say that the role pitch accents and overall tunes play in terms of any pragmatic meaning is something we offer here as something to think about. It is an idea proposed by Pierrehumbert and Hirschberg (1990) that has not been tested. Whether these pitch accent patterns are found for the six affective expressions for all speakers of Mainstream America English is an interesting question to be explored. Specifically, we wonder if there is an equivalence between pitch accents/intonational tunes and pragmatic functions for stereotypical culturally-specific social affective expressions of the type we have collected for our second language teaching purposes? Or, is it that the variability in pitch accents/intonational tunes negates any possible pragmatic meaning to specific pitch accents? To the best of our knowledge, intonational analyses have not been systematically done for social affective expressions; it may well be that a re-visit of the Pierrehumbert and Hirschberg approach as it applies to social affective expressions, expressions that are somewhat stereotypical for a certain dialect/culture, might be appropriate as well as lead to some interesting insights about how we express ourselves in culturally-regulated social interactions. That is, in certain well-defined pragmatic situations, is there an equivalence between pitch accents/overall tunes and pragmatic functions?

Finally, we acknowledge that the ToBI annotation was done by a single rater (the first author), which is a limitation of this study. Future work of annotating intonational profiles of social affective expressions will require more than one annotator, as well as a larger data base to work with.

4.2 Discussion of phonetic profiles of nuclear stress vowel in six affective expressions by two speakers

Table 9 presents a description of the nuclear stress vowel in the six affective expressions for the two speakers examined in this study. The third column shows the pitch accents for the nuclear stress vowels. As discussed above, the two speakers seem to agree on which pitch accents to use for each of the affective expressions. If this finding is borne out by future studies, this is an interesting finding from our study.

The fourth column describes in relative terms the mean acoustic values for each of the nuclear stress vowels, based on the estimates in Table 1. For instance, F0 values are described relatively as high, moderately high, not high, moderately low, and similarly for the other values. These assessments were made by sorting the data from high to low values, and then dividing them into the five descriptive categories. For future explorations of phonetic profiles of social affective expressions, we plan to work with a larger set of data in order to allow inferential statistical analyses.

The fifth column describes the auditory perceptions of voice qualities of the nuclear stress vowel. Table 4, presented earlier in the paper, gave an overall summary of voice quality perceptions of the nuclear vowel. Here (Table 9) we present the data again, but separately by speaker. The bold items indicate the differences in voice quality production by the two speakers for a given attitude.

One motivation for presenting acoustic and auditory summaries in the same table is to highlight that it is not trivial to ascertain which acoustic measurements relate to which auditory perceptions. The MFA analyses discussed above show some relationships between acoustic measurements and auditory perceptions, but a larger corpus is needed in order to assess the acoustic measurements that best account for auditory judgments. The final column is the VAD ratings of the entire utterance. A rating above 0 (Figure 6) is interpreted as being assertive, excited, or as having a positive valence; below 0, as accommodating, calm, or negative valence. The assumption is that the nuclear stress acoustic estimates, together with the auditory perceptions, influence the VAD judgments.

Nevertheless, this fine detail examination of the phonetic profiles of a small set of social affective expressions has revealed some interesting similarities and differences. For instance, even though both speakers use the same pitch accent for the nuclear stress vowel of a given type of expression, listeners do not give the same 3-D emotional ratings VAD ratings. Both speakers used social codes for making social expression, but one speaker tends to use the Frequency Code, the other the Effort Code. Our working assumption is that the speaker who uses a more modal voice tends to use the Frequency Code, while the one with a more falsetto voice, uses the Effort Code. This assumption needs to be tested with more speakers.

As for acoustic measurements for measuring voice quality, it seems that NAQ may indeed be a good estimate for tense voice due to glottal closing behavior, CPP, a good estimation for strong non-noisy harmonics, Peak Slope, a good estimate of spectral-related tense voice, and Hammarberg Index, for distribution of spectral energy, i.e., strong or weak energy in the upper frequencies. However, it is also fairly clear that these measures do not completely account for the auditory judgments of voice quality. Currently, we are working on a more in-depth study to better clarify the acoustic and auditory parameters of voice quality and, specifically, types of voice qualities used in everyday social conversations.

Table 9: Phonetic profiles of nuclear stress syllable, and VAD ratings for utterances; ratings from the acoustic and auditory analysis were split into a scale of 1 to 5 (for low, medium low, medium, medium high, high) that is used to describe the values of these parameters in the Acoustics and Auditory columns.

Spk	Expression	Pitch accent	Acoustics	Auditory	VAD ratings of utterance
S3	Authority	H*+L	F0: 3; intensity: 3-5; CPP: 5; NAQ: 2; PS: 5; H-index: 4	High: 2; loud: 1; falsetto: 1; tense: 3; twang: 1; breathy: 1	Assertive, Aroused, Positive
S6	Authority	H*+L	F0: 4; intensity: 4; CPP: 4; NAQ: 3; PS: 5; H-index: 4	High: 2; loud: 4; falsetto: 4; tense: 4; twang: 4; breathy: 1	Assertive, Positive
S3	Declaration	L*	F0: 2; intensity: 2; CPP: 4; NAQ: 2; PS: 5; H-index: 4	High: 2; loud: 1; falsetto: 1; tense: 2; twang: 1; breathy: 1	Assertive, Positive
S6	Declaration	L*	F0: 3; intensity: 4; CPP: 4; NAQ: 3; PS: 5; H-index: 3	High: 2; loud: 1; falsetto: 1; tense: 2; twang: 1; breathy: 1	Positive
S3	Irritation	L*+H	F0: 5; intensity: 5; CPP: 2; NAQ: 5; PS: 2; H-index: 3	High: 4; loud: 4; falsetto: 2; tense: 4; twang: 4; breathy: 1	Assertive, Aroused, Positive
S6	Irritation	L*+H	F0: 3-5; intensity: 5; CPP: 5; NAQ: 2; PS: 2; H-index: 1	High: 2; loud: 4; falsetto: 4; tense: 4; twang: 4; breathy: 1	Assertive, Aroused
S3	Sincerity	L*	lF0: 1; intensity: 1; CPP: 4; NAQ: 2; PS: 3; H-index: 3	High: 1, loud: 1; falsetto: 1; tense: 1; twang: 1; breathy: 1	Assertive
S6	Sincerity	L*	F0: 3; intensity: 3; CPP: 2; NAQ: 2; PS: 2; H-index: 2	High: 4; loud: 1; falsetto: 1; tense: 1; twang: 1; breathy: 4	Positive
S3	Uncertainty	!H*	F0: 3; intensity: 3; CPP: 4; NAQ: 2; PS: 3; H-index: 3	High: 2; loud: 2; falsetto: 2; tense: 1; twang: 1; breathy: 1	Negative, Accomodating
S6	Uncertainty	!H*	F0: 3; intensity: 2; CPP: 2; NAQ: 3; PS: 3; H-index: 4	High: 4; loud: 1; falsetto: 1; tense: 1; twang: 1; breathy: 4	Negative, Accomodating
S3	Woeg	L*	F0: 2; intensity: 2; CPP: 2; NAQ: 1; PS: 3; H-index: 5	High: 3; loud: 1; falsetto: 1; tense: 1; twang: 1; breathy: 5	Negative, Accomodating
S6	Woeg	L*	F0: 1; intensity: 3; CPP: 1; NAQ: 3; PS: 1; H-index: 5	High: 3; loud: 1; falsetto: 1; tense: 1; twang: 1; breathy: 5	Negative, Accomodating

5 Conclusion

Our main aim was to draw some connections between acoustics, perception, and production of attitudinal expressions. Based on MFA analysis of the prosodic aspects of the six affective expressions, we find that a soft / noisy voice, with weak harmonics and irregular rhythm with pauses and hesitations, as in the expressions of Uncertainty and WOEG, is perceived by listeners as accommodating (not Dominant) and not positive (negative Valence). Loud, tense voices with energy in the upper frequencies, as in the expression of Irritation, are perceived as Aroused.

Expressions of Authority, Declaration, and Sincerity tend to have a comparatively regular rhythm and relatively flat intonation.

This article focuses on the importance of some of the acoustic, auditory, and phonological details in a speaker's voice as they engage in various types of socially prescribed interactions; and how this linguistic information is then interpreted by listeners in terms of 3-dimensional emotional perceptions of Valence, Arousal, and Dominance. Each individual trained themself through years of social experience in expressive strategies, given their physiological characteristics and socially developed skills. The choice of S3 and S6 to position their voices differently is an acquired habit used in the kinds of social interactions elicited during the recording process, as well as used in other social interactions. Voice habits, however, can also be moderated in keeping with various aspects of the perceived social environment. For instance, S6, while teaching, reports using a "lower" voice (modal or thick fold vibration mode of the vocal folds – see Erickson et al. 2020a), an example of character persona (Drager 2015; Sadanobu 2015), for the same individual in different social situations. Given their chosen characters during the recording process, each speaker selected available strategies to tune their voice in order to express the various targeted social affective expressions.

Our study shows that several expressive strategies are possible, and they do not impede listeners in interpreting vocal cues: the expressions were identified and described adequately in several evaluation experiments (Rilliard et al. 2013, 2017). A potential explanation for this capacity to interpret complex and varied acoustic cues under similar expressive labels may lie in their symbolism derived from emotional experience and biological codes. A difficulty in accessing such an interpretation based on acoustic cues is the crudeness of these measures, considered individually, that capture a physical dimension (e.g., F0) that does not necessarily match its perceptual counterpart (in that case, pitch): we have seen how one speaker produces F0 variations that are mostly inversely related to her perceived pitch. It must be noted that most studies of "prosody" use F0 as the correlate of pitch, without this relation raising problems, and certainly for good reasons: most of these studies deal with linguistic functions and are based on conversational speech, with much-reduced variations of vocal effort. The theoretical models of such F0 variations, as the Fujisaki model (Fujisaki 1983; Fujisaki and Hirose 1984), show the importance of different factors for the production of F0 changes – among which what is modeled as "accent command" is articulatory produced by an augmentation of the vocal folds' tension due to momentarily change the vocal fold configuration (Fujisaki 1988). Such F0 changes are rapid and mostly independent of subglottal pressure, which is another means to raise F0, typically used here by S6 for her expressive strategy, but of much less use in conversational language (Collier 1975; see also the motivation of Gussenhoven 2004's "production code"). F0 is certainly the primary cue to perceived pitch, when speakers are asked to change just that – their pitch (Bishop and Keating 2012). In the case of expressive speech, at least the vocal effort exerts important modifications on the configuration of the source and on the vocal tract, so that most of the predictions made for conversational speech do not necessarily hold. For example, important changes in formants linked to vocal tract opening, fronted articulation, forced smile have been reported for emotional expressions (Rilliard et al. 2018). The twangy voice observed here, mostly for S6, is another strategy to enhance one's voice power. The question of the discriminability of the voice quality changes linked to twang, or other energy-enhancing strategies remains to be explored (see Gerratt and Kreiman, 2001, for such a work on non-modal phonations).

In the case of expressive speech, it thus seems that combinations of several acoustic parameters (or model-derived estimates such as NAQ) are important to attain a better account of perceptual correlates like pitch and loudness. Goudbeek and Scherer (2010) have shown that the Arousal dimension changes many aspects of vocal performance (see also Liénard 2019

specifically on the influence of vocal effort on the spectral characteristics of the speech output). Experimental and theoretical descriptions of voice source behavior (Titze 1989; Titze and Sundberg 1992) have shown correlations between intensity and F0 parameters, but how they are linked to perception remains to be fully understood.

In this paper, we have reported connections between perceptual variation in affective dimensions, along with acoustic/prosodic phonatory voice-related variation in these affective dimensions. There are some very new things in our data: (a) we examine affective expressions, not emotions, and (b) based on the perception of these attitudes, we explore their acoustic and auditory correlates with a strong focus on voice quality parameters. We also report in passing some speaker-specific differences as a secondary observation. The overarching pattern is that we see different acoustic and auditory characteristics for different well-perceived affective expressions. The findings from this pilot study await to be verified by studying larger datasets. We hope that some of the avenues explored here are sufficiently interesting to encourage future research in the area of acoustic and perceptual profiles of social affective expressions across various dialects and languages.

6 Data Availability Statement

The audio sound files for the six pairs of affective expressions for two speakers, along with the six pairs of voice quality exemplar sounds examined in this study, can be found in 10.5281/zenodo.6406846.svg , https://zenodo.org/badge/DOI/10.5281/zenodo.6406846.svg.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Author Contributions

DE, AR, ET, JM, and TS were substantially involved in various aspects of the design, the data collection, the analysis and interpretation, and the drafting of the current work. DE took the leading role in organizing the paper, along with AR, who also played a key role in the acoustic and statistical analysis, as well as the interpretation of the data; ET in the VAD data collection and interpretation, and JM together with TS in the design of the project. in making sure the paper got published. Finally, all authors worked together in criticizing and revising the work, and approving the version for publication.

9 Funding

This work has been partially funded by the French National Research Agency (project Gender Equality Monitor - ANR-19-CE38-0012).

10 Acknowledgments

The first part of this research, the recording and preliminary analyses, was done thanks to the ANR PADE grant from the French government (2012-2014) awarded to the second author. We are thankful to the two female American English speakers who recorded the social affective

expressions, the university students in northern California, who rated the Valence, Arousal and Dominance dimensions of the affective expressions, and the listeners who rated the auditory qualities of the nuclear stress vowel in the utterances. Most of all, we wish to thank Helen Rowson, Vocal Coach, trained in the Estill Vocal Training Method, for her recordings of the voice quality exemplar sounds.

11 Ethical statements

Ethics statements
Studies involving animal subjects
Generated Statement: No animal studies are presented in this manuscript.
Studies involving human subjects
Generated Statement: The participants provided their written informed consent to participate in this study.
Inclusion of identifiable human data]
Generated Statement: No potentially identifiable human images or data is presented in this study.

REFERENCES

1. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010 Jul;2(4):433–59.

2. d'Alessandro C. Voice source parameters and prosodic analysis. In: Sudhoff S, Lenertova D, Meyer R, Pappert S, Augurzky P, Mleinek I, et al., editors. Methods in empirical prosody research. Berlin: Walter de Gruyter; 2006. p. 63–87.

3. d'Alessandro C, Darsinos V, Yegnanarayana B. Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. IEEE Trans Speech Audio Process. 1998; 6(1):12–23.

4. Alku P, Bäckström T, Vilkman E. Normalized amplitude quotient for parametrization of the glottal flow. J Acoust Soc Am. 2002 Aug;112(2):701–10.

5. Anikin A. A Moan of Pleasure Should Be Breathy: The Effect of Voice Quality on the Meaning of Human Nonverbal Vocalizations. Phonetica. 2020 Sep 1;77(5):327–49.

6. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Pers Soc Psychol. 1996;70:614–36.

7. Bänziger T, Scherer KR. The role of intonation in emotional expressions. Speech Commun. 2005 Jul;46(3–4):252–67.

8. Barrett LF. Are Emotions Natural Kinds? Perspect Psychol Sci. 2006 Mar;1(1):28–58.

9. Beckman ME, Ayers G. Guidelines for ToBI labelling. OSU Res Found. 1997;3(30):255–309.

10. Bishop J, Keating P. Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. J Acoust Soc Am. 2012;132(2):1100–12.

11. Boersma P, Weenink D. Praat: doing phonetics by computer [Computer program]. Version 6.2.08 [Internet]. 2022. Available from: http://www.praat.org/

12. Caballero JA, Vergis N, Jiang X, Pell MD. The sound of im/politeness. Speech Commun. 2018 Sep;102:39–53.

13. Camargo Z, Madureira S. Voice quality analysis from a phonetic perspective: Voice profile analysis scheme (VPAS) profile for Brazilian Portuguese. In: Proc 4th International Conference of Speech Prosody, Campinas, Brazil. 2008. p. 57–60.

14. Camargo Z, Madureira S, dos Reis N, Rilliard A. The phonetic approach of voice qualities: challenges in corresponding perceptual to acoustic descriptions. In: Lahoz-Bengoechea JM, Pérez Ramón R, editors.

Subsidia Tools and resources for speech sciences [Internet]. 2019. Available from: https://riuma.uma.es/xmlui/handle/10630/18177

15. Collier R. Physiological correlates of intonation patterns. J Acoust Soc Am. 1975 Jul;58(1):249–55.

16. Couper-Kuhlen E. An introduction to English prosody. London: Arnold; 1986.

17. Culpeper J, Bousfield D, Wichmann A. Impoliteness revisited: with special reference to dynamic and prosodic aspects. J Pragmat. 2003 Oct;35(10–11):1545–79.

18. Damasio AR. Emotion in the perspective of an integrated nervous system. Brain Res Rev. 1998;26(2–3):83–6.

19. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP: A collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [Internet]. Florence, Italy: IEEE; 2014 [cited 2021 Mar 10]. p. 960–4. Available from: http://ieeexplore.ieee.org/document/6853739/

20. Drager KK. Linguistic variation, identity construction and cognition. Berlin: Language Science Press; 2015. (Studies in laboratory phonology).

21. Elfenbein HA, Ambady N. Is there an in-group advantage in emotion recognition? Psychol Bull. 2002a;128(2):243–9.

22. Elfenbein HA, Ambady N. On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychol Bull. 2002b;128(2):203–35.

23. Elfenbein HA, Ambady N. Universals and Cultural Differences in Recognizing Emotions. Curr Dir Psychol Sci. 2003 Oct;12(5):159–64.

24. Erickson D. Expressive speech: Production, perception and application to speech synthesis. Acoust Sci Technol. 2005;26(4):317–25.

25. Erickson D, Hayashi S, Hose Y, Suzuki M, Ueno Y, Maekawa K. Perception of American English sarcasm by Japanese listeners. In: Acoustical Society of Japan Spring Meeting. 2002. p. 333–4.

26. Erickson D, Kawahara S, Rilliard A, Hayashi R, Sadanobu T, Li Y, et al. Cross cultural differences in arousal and valence perceptions of voice quality. In: Speech Prosody 2020 [Internet]. ISCA; 2020a [cited 2021 Sep 28]. p. 720–4. Available from: https://www.isca-speech.org/archive/speechprosody_2020/erickson20b_speechprosody.html

27. Erickson D, Niebuhr O. Articulation of prosody and rhythm: Some possible applications to language teaching. In: Proceedings of the 13th International Conference of Nordic Prosody [Internet]. Sciendo; 2023 [cited 2024 Dec 16]. p. 1–45. Available from: https://www.sciendo.com/chapter/9788366675728/10.2478/9788366675728-001

28. Erickson D, Rilliard A, Li Y, Menezes C, Kawahara S, Sadanobu T, et al. Cross Cultural perception of Valence and Arousal. In: Skarnitzl R, Volín J, editors. Proceedings of the 20th International Congress of Phonetic Sciences. Prague, Czech Republic: Guarant International; 2023. p. 1776–80.

29. Erickson D, Rilliard A, de Moraes J, Shochi T. Personality Judgments Based on Speaker's Social Affective Expressions. In: Fang Q, Dang J, Perrier P, Wei J, Wang L, Yan N, editors. Studies on Speech Production [Internet]. Cham: Springer International Publishing; 2018 [cited 2021 Sep 27]. p. 3–13. (Lecture Notes in Computer Science; vol. 10733). Available from: http://link.springer.com/10.1007/978-3-030-00126-1_1

30. Erickson D, Rilliard A, Thurgood E, de Moraes JA, Shochi T. A Valence-Arousal-Dominance Study of American English Social Affective Expressions. In 2022a [cited 2022 Aug 9]. p. 595–9. Available from: https://www.isca-speech.org/archive/speechprosody_2022/erickson22_speechprosody.html

31. Erickson D, Yoshida K, Menezes C, Fujino A, Mochida T, Shibuya Y. Exploratory Study of Some Acoustic and Articulatory Characteristics of Sad Speech. Phonetica. 2006 Mar 1;63(1):1–25.

32. Erickson D, Yun J, Gao J, Obert K. Interaction between phonation mode and pharyngeal narrowing: A pilot EGG study. In: Tiede M, Whalen DH, Gracco V, editors. Proceedings of the 12th International Seminar on Speech Production. New Haven, CT, USA: Haskins Press; 2020b. p. 190–3.

33. Erickson D, Yun J, Obert K, Reeve M, Rowson H, Møller K. Voice quality: Interactions among F0, vowel quality, phonation mode and phayngeal narrowing. In: Nordic Prosody 13. 2022b.

34. Esling JH, Moisik SR, Benner A, Crevier-Buchman L. Voice Quality: The Laryngeal Articulator Model [Internet]. 1st ed. Cambridge University Press; 2019 [cited 2022 Jan 8]. Available from: https://www.cambridge.org/core/product/identifier/9781108696555/type/book

35. Fitch HL, Halwes T, Erickson DM, Liberman AM. Perceptual equivalence of two acoustic cues for stop-consonant manner. Percept Psychophys. 1980 Jul;27(4):343–50.

36. Fónagy I, Bérard E. «Il est huit heures»: contribution à l'analyse sémantique de la vive voix. Phonetica. 1972;26(3):157–92.

37. Fontaine JRJ, Scherer KR, Roesch EB, Ellsworth PC. The World of Emotions is not Two-Dimensional. Psychol Sci. 2007 Dec;18(12):1050–7.

38. Fujimura O. The C/D Model and Prosodic Control of Articulatory Behavior. Phonetica. 2000 Dec 1;57(2–4):128–38.

39. Fujisaki H. Dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage P, editor. The production of speech. New York, NY: Springer; 1983. p. 39–55.

40. Fujisaki H. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: Fujimura O, editor. Vocal fold physiology: voice production, mechanisms and functions. New York, NY: Raven; 1988. p. 347–55.

41. Fujisaki H, Hirose K. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J Acoust Soc Jpn E. 1984;5(4):233–42.

42. Geng P, Gu W, Johnson K, Erickson D. Acoustic-Prosodic and Articulatory Characteristics of the Mandarin Speech Conveying Dominance or Submissiveness. In: Speech Prosody 2020 [Internet]. ISCA; 2020 [cited 2022 Mar 11]. p. 424–8. Available from: https://www.isca-speech.org/archive/speechprosody_2020/geng20_speechprosody.html

43. Gerratt BR, Kreiman J. Toward a taxonomy of nonmodal phonation. J Phon. 2001 Oct;29(4):365–81.

44. González-Fuente S, Escandell-Vidal V, Prieto P. Gestural codas pave the way to the understanding of verbal irony. J Pragmat. 2015;90:26–47.

45. Goudbeek M, Scherer K. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. J Acoust Soc Am. 2010;128(3):1322–36.

46. Gussenhoven C. The Phonology of Tone and Intonation [Internet]. 1st ed. Cambridge University Press; 2004 [cited 2020 Aug 1]. Available from: https://www.cambridge.org/core/product/identifier/9780511616983/type/book

47. Hammarberg B, Fritzell B, Gaufin J, Sundberg J, Wedin L. Perceptual and Acoustic Correlates of Abnormal Voice Qualities. Acta Otolaryngol (Stockh). 1980 Jan;90(1–6):441–51.

48. Hanson HM. Glottal characteristics of female speakers: Acoustic correlates. J Acoust Soc Am. 1997;101(1):466–81.

49. Hareli S, Kafetsios K, Hess U. A cross-cultural study on emotion expression and the learning of social norms. Front Psychol [Internet]. 2015 Oct 2 [cited 2022 Aug 9];6. Available from: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.01501/abstract

50. Henrich N, Bezard P, Expert R, Garnier M, Guerin C, Pillot C, et al. Towards a common terminology to describe voice quality in western lyrical singing: Contribution of a multidisciplinary research group. J Interdiscip Music Stud. 2008;2(1 & 2):71–93.

51. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic Correlates of Breathy Vocal Quality. J Speech Lang Hear Res. 1994 Aug;37(4):769–78.

52. Hillenbrand J, Houde RA. Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. J Speech Lang Hear Res. 1996 Apr;39(2):311–21.

53. Husson F, Lê S, Pagès J. Exploratory multivariate analysis by example using R. Second edition. Boca Raton: CRC Press; 2017.

54. Idemaru K, Winter B, Brown L. Cross-cultural multimodal politeness: The phonetics of Japanese deferential speech in comparison to Korean. Intercult Pragmat. 2019 Nov 1;16(5):517–55.

55. Jackson PJB, Shadle CH. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. IEEE Trans Speech Audio Process. 2001 Oct;9(7):713–26.

56. Johnstone T, Scherer KR. Vocal communication of emotion. In: Lewis M, Haviland J, editors. Handbook of emotions. New York: Guilford; 2000. p. 220–35.

57. Jun SA. The ToBI Transcription System: Conventions, Strengths, and Challenges. In: Barnes J, Shattuck-Hufnagel S, editors. Prosodic Theory and Practice [Internet]. The MIT Press; 2022 [cited 2022 Oct 11]. p. 151–81. Available from: https://direct.mit.edu/books/oa-edited-volume/5259/Prosodic-Theory-and-Practice

58. Jürgens R, Hammerschmidt K, Fischer J. Authentic and Play-Acted Vocal Emotion Expressions Reveal Acoustic Differences. Front Psychol [Internet]. 2011 [cited 2024 Dec 16];2. Available from: http://journal.frontiersin.org/article/10.3389/fpsyg.2011.00180/abstract

59. Juslin PN, Laukka P. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion. 2001;1(4):381–412.

60. Kane J, Gobl C. Identifying regions of non-modal phonation using features of the wavelet transform. In: Interspeech 2011 [Internet]. ISCA; 2011 [cited 2022 Mar 22]. p. 177–80. Available from: https://www.isca-speech.org/archive/interspeech_2011/kane11_interspeech.html

61. Lamesch S, Doval B, Castellengo M. Toward a more informative voice range profile: The role of laryngeal vibratory mechanisms on vowels dynamic range. J Voice. 2012;26(5):672-e9.

62. Laukka P, Elfenbein HA. Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis. Emot Rev. 2021 Jan;13(1):3–11.

63. Laver J. The Phonetic Description of Voice Quality. Cambridge, UK: Cambridge University Press; 1980. (Cambridge Studies in Linguistics).

64. Lehiste I. Suprasegmentals. Cambridge, Mass: M.I.T. Press; 1970.

65. Liénard JS. Quantifying vocal effort from the shape of the one-third octave long-term-average spectrum of speech. J Acoust Soc Am. 2019 Oct;146(4):EL369–75.

66. Liénard JS, Di Benedetto MG. Effect of vocal effort on spectral properties of vowels. J Acoust Soc Am. 1999 Jul;106(1):411–22.

67. Liu X, Xu Y. Body size projection by voice quality in emotional speechEvidence from Mandarin Chinese. In: Speech Prosody 2014 [Internet]. ISCA; 2014 [cited 2024 Dec 16]. p. 974–7. Available from: https://www.isca-archive.org/speechprosody_2014/liu14c_speechprosody.html

68. Loveday L. Pitch, Politeness and Sexual Role: An Exploratory Investigation into the Pitch Correlates of English and Japanese Politeness Formulae. Lang Speech. 1981 Jan;24(1):71–89.

69. Maryn Y, Weenink D. Objective Dysphonia Measures in the Program Praat: Smoothed Cepstral Peak Prominence and Acoustic Voice Quality Index. J Voice. 2015 Jan;29(1):35–43.

70. Mauchand M, Pell MD. Emotivity in the Voice: Prosodic, Lexical, and Cultural Appraisal of Complaining Speech. Front Psychol. 2021 Jan 18;11:619222.

71. Mauchand M, Vergis N, Pell MD. Irony, Prosody, and Social Impressions of Affective Stance. Discourse Process. 2020 Feb 7;57(2):141–57.

72. Mello H, Raso T. Illocution, modality, attitude: different names for different categories. In: Mello H, Panunzi A, Raso T, editors. Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation [Internet]. 1st ed. Firenze: Firenze University Press; 2011 [cited 2022 Aug 9]. p. 1–18. (Strumenti per la didattica e la ricerca; vol. 120). Available from: https://books.fupress.com/isbn/9788866550846

73. Mixdorff H, Rilliard A, Lee T, Ma MKH, Hönemann A. Cross-cultural (A)symmetries in Audio-visual Attitude Perception. In: Interspeech 2018 [Internet]. ISCA; 2018 [cited 2022 Jan 9]. p. 426–30. Available from: https://www.isca-speech.org/archive/interspeech_2018/mixdorff18_interspeech.html

74. de Moraes JA. From a prosodic point of vew: remarks on attitudinal meaning. In: Mello H, Panunzi A, Raso T, editors. Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation [Internet]. 1st ed. Firenze: Firenze University Press; 2011 [cited 2022 Aug 9]. p. 19–37.

(Strumenti per la didattica e la ricerca; vol. 120). Available from: https://books.fupress.com/isbn/9788866550846

75. de Moraes JA, Rilliard A, Erickson D, Shochi T. Acoustic analysis of a corpus of Brazilian Portuguese attitudes. In: Proceedings of the VIIth GSCP International Conference?: Speech and Corpora. Firenze University Press; 2012. p. 162–6.

76. Mozziconacci S. Speech variability and emotion?: production and perception [Internet]. Technische Universiteit Eindhoven; 1998 [cited 2022 Mar 11]. Available from: https://research.tue.nl/en/publications/speech-variability-and-emotion--production-and-perception(d8025d2b-2371-4cd5-83ad-0e62867f9fc7).html

77. Nadeu M, Prieto P. Pitch range, gestural information, and perceived politeness in Catalan. J Pragmat. 2011 Feb;43(3):841–54.

78. Niebuhr O. "A little more ironic" – Voice quality and segmental reduction differences between sarcastic and neutral utterances. In: 7th International Conference on Speech Prosody, Dublin, Ireland. 2014. p. 608–12.

79. Niebuhr O, Reetz H, Barnes J, Yu ACL. Fundamental Aspects in the Perception of f0. In: Gussenhoven C, Chen A, editors. The Oxford Handbook of Language Prosody [Internet]. Oxford University Press; 2020 [cited 2024 Dec 16]. p. 28–42. Available from: https://academic.oup.com/edited-volume/34870/chapter/298314264

80. Niebuhr O, Tegtmeier S, Schweisfurth T. Female Speakers Benefit More Than Male Speakers From Prosodic Charisma Training—A Before-After Analysis of 12-Weeks and 4-h Courses. Front Commun. 2019 Apr 3;4:12.

81. Ohala JJ. Cross-Language Use of Pitch: An Ethological View. Phonetica. 1983 Jan 1;40(1):1–18.

82. Ohala JJ. An Ethological Perspective on Common Cross-Language Utilization of F? of Voice. Phonetica. 1984 Jan 1;41(1):1–16.

83. Ohala JJ. The frequency code underlies the sound-symbolic use of voice pitch. In: Hinton L, Nichols J, Ohala JJ, editors. Sound Symbolism [Internet]. 1st ed. Cambridge University Press; 1994 [cited 2021 Nov 24]. p. 325–47. Available from: https://www.cambridge.org/core/product/identifier/CBO9780511751806A036/type/book_part

84. Osgood CE, May WH, Miron MS. Cross-cultural universals of affective meaning. University of Illinois Press; 1975.

85. Pagès J, Husson F. Inter-laboratory comparison of sensory profiles. Food Qual Prefer. 2001 Jul;12(5–7):297–309.

86. Perta K, Bae Y, Obert K. A pilot investigation of twang quality using magnetic resonance imaging. Logoped Phoniatr Vocol. 2021 Apr 3;46(2):77–85.

87. Pierrehumbert J, Hirschberg J. The Meaning of Intonational Contours in the Interpretation of Discourse. In: Cohen PR, Morgan J, editors. Intentions in communication. Cambridge, Mass.: MIT Press; 1990. p. 271–311. (System development foundation benchmark series).

88. Pierrehumbert JB. The phonology and phonetics of English intonation [PhD Thesis]. Massachusetts Institute of Technology; 1980.

89. Rilliard A, d'Alessandro C, Evrard M. Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis. J Acoust Soc Am. 2018 Jan;143(1):109–22.

90. Rilliard A, Erickson D, de Moraes JA, Shochi T. Cross-cultural perception of some Japanese politeness and impoliteness expressions. In: Baider F, Cislaru G, editors. Linguistic Approaches to Emotions in Context [Internet]. Amsterdam: John Benjamins Publishing Company; 2014 [cited 2021 Nov 24]. p. 251–76. (Pragmatics & Beyond New Series; vol. 241). Available from: https://benjamins.com/catalog/pbns.241.15ril

91. Rilliard A, Erickson D, de Moraes JA, Shochi T. Perception of expressive prosodic speech acts performed in USA english by L1 and L2 speakers. J Speech Sci. 2017 Nov 1;6(1):27–45.

92. Rilliard A, Erickson D, Shochi T, de Moraes JA. Social face to face communication — American English attitudinal prosody. In: Interspeech 2013 [Internet]. ISCA; 2013 [cited 2021 Dec 30]. p. 1648–52. Available from: https://www.isca-speech.org/archive/interspeech_2013/rilliard13_interspeech.html

93. Rilliard A, de Moraes JA. Social affective variations in Brazilian Portuguese: a perceptual and acoustic analysis. Rev Estud Ling. 2017 Jun 13;25(3):1043–74.

94. Rilliard A, Shochi T, Erickson D, de Moraes JA. Developmental perception of polite & impolite non-verbal behaviours in Japanese. In: Mello H, Pettorino M, Raso T, editors. Proceedings of the VIIth GSCP International Conference?: Speech and Corpora. Firenze University Press; 2012. p. 167–71.

95. Rilliard A, Shochi T, Martin JC, Erickson D, Aubergé V. Multimodal Indices to Japanese and French Prosodically Expressed Social Affects. Lang Speech. 2009 Jun;52(2–3):223–43.

96. Rossi M. Interactions of intensity glides and frequency glissandos. Lang Speech. 1978;21(4):384–96.

97. Russell JA. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull. 1994;115(1):102–41.

98. Sadanobu T. "Characters" in Japanese Communication and Language: An Overview. Acta Linguist Asiat. 2015 Dec 29;5(2):9–28.

99. San Segundo E, Foulkes P, French P, Harrison P, Hughes V, Kavanagh C. The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. J Int Phon Assoc. 2019 Dec;49(3):353–80.

100. Scherer K. Vocal communication of emotion: A review of research paradigms. Speech Commun. 2003 Apr;40(1–2):227–56.

101. Scherer KR. Emotions are emergent processes: they require a dynamic computational architecture. Philos Trans R Soc B Biol Sci. 2009a Dec 12;364(1535):3459–74.

102. Scherer KR. The dynamic architecture of emotion: Evidence for the component process model. Cogn Emot. 2009b Nov;23(7):1307–51.

103. Scherer KR, Banse R, Wallbott HG, Goldbeck T. Vocal cues in emotion encoding and decoding. Motiv Emot. 1991 Jun;15(2):123–48.

104. Scherer KR, Ladd DR, Silverman KEA. Vocal cues to speaker affect: Testing two models. J Acoust Soc Am. 1984 Nov;76(5):1346–56.

105. Schmidt J, Janse E, Scharenborg O. Perception of Emotion in Conversational Speech by Younger and Older Listeners. Front Psychol [Internet]. 2016 May 31 [cited 2022 Mar 30];7. Available from: http://journal.frontiersin.org/Article/10.3389/fpsyg.2016.00781/abstract

106. Schröder M, Heylen D, Poggi I. Perception of non-verbal emotional listener feedback. In: Proc Speech Prosody 2006. 2006. p. paper 072.

107. Shadle CH. Intrinsic fundamental frequency of vowels in sentence context. J Acoust Soc Am. 1985 Nov;78(5):1562–7.

108. Shochi T, Guerry M, Rilliard A, Erickson D, Rouas JL. The combined Perception of Socio-affective Prosody: Cultural Differences in Pattern Matching. J Phon Soc Jpn. 2020 décembre;24:84–96.

109. Shochi T, Rilliard A, Aubergé V, Erickson D. Intercultural perception of English, French and Japanese social affective prosody. In: Hancil S, editor. The role of prosody in affective speech. Bern, Switzerland: Peter Lang AG; 2009. p. 31–60. (Linguistic Insights).

110. Shochi T, Rilliard A, Erickson D. Chapter 8. Perceptual changes between adults and children for multimodal im/politeness in Japanese. In: Jucker AH, Hübscher I, Brown L, editors. Pragmatics & Beyond New Series [Internet]. Amsterdam: John Benjamins Publishing Company; 2023 [cited 2024 Apr 12]. p. 213–49. Available from: https://benjamins.com/catalog/pbns.333.08sho

111. Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, et al. TOBI: a Standard for Labeling English Prosody. In: Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP'92), Banff, Alberta, Canada. 1992. p. 867–70.

112. Steinhauer K, McDonald Klimek M, Estill J. The Estill voice model: theory & translation. Pittsburgh, Pennsylvania: Estill Voice International; 2017.

113. Stevens KN. Acoustic phonetics. Vol. 30. MIT press; 2000.

114. Stoet G. PsyToolkit: A software package for programming psychological experiments using Linux. Behav Res Methods. 2010 Nov;42(4):1096–104.

115. Stoet G. PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. Teach Psychol. 2017 Jan;44(1):24–31.

116. Syrdal AK, McGory J. Inter-transcriber reliability of toBI prosodic labeling. In: 6th International Conference on Spoken Language Processing (ICSLP 2000) [Internet]. ISCA; 2000 [cited 2022 Oct 3]. p. vols. 3, 235-238–0. Available from: https://www.isca-speech.org/archive/icslp_2000/syrdal00_icslp.html

117. Székely É, Mendelson J, Gustafson J. Synthesising Uncertainty: The Interplay of Vocal Effort and Hesitation Disfluencies. In: Interspeech 2017 [Internet]. ISCA; 2017 [cited 2023 Mar 16]. p. 804–8. Available from: https://www.isca-speech.org/archive/interspeech_2017/szekely17_interspeech.html

118. Thorson J, Burdin RS. The interpretation and phonetic implementation of?!H* in American English. In 2022 [cited 2022 Aug 17]. p. 749–53. Available from: https://www.isca-speech.org/archive/speechprosody_2022/thorson22b_speechprosody.html

119. Titze IR. On the relation between subglottal pressure and fundamental frequency in phonation. J Acoust Soc Am. 1989 Feb;85(2):901–6.

120. Titze IR, Sundberg J. Vocal intensity in speakers and singers. J Acoust Soc Am. 1992 May;91(5):2936–46.

121. Traunmüller H, Eriksson A. Acoustic effects of variation in vocal effort by men, women, and children. J Acoust Soc Am. 2000 Jun;107(6):3438–51.

122. Uldall E. Attitudinal meanings conveyed by intonation contours. Lang Speech. 1960;3(4):223–34.

123. Ward G, Hirschberg J. Implicating Uncertainty: The Pragmatics of Fall-Rise Intonation. Language. 1985 Dec;61(4):747.

124. Weiss B, Trouvain J, Barkat-Defradas M, Ohala JJ, editors. Voice attractiveness: studies on sexy, likable, and charismatic speakers. Singapore: Springer; 2021.

125. Wichmann A. The attitudinal effects of prosody, and how they relate to emotion. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. 2000. p. 143–8.

126. Wierzbicka A. Defining emotion concepts. Cogn Sci. 1992;16(4):539–81.

127. Williams CE, Stevens KN. Emotions and Speech: Some Acoustical Correlates. J Acoust Soc Am. 1972 Oct 1;52(4B):1238–50.

128. Xue Y, Hamada Y, Akagi M. Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. Speech Commun. 2018 Sep;102:54–67.

129. Yanagida M. Discriminating ironies from praising-Acoustic parameters vs. prosodic parameters. In: Proceedings of the Symposium Prosody and Speech Processing. University of Tokyo; 2002. p. 143–6.

130. Yoon TJ, Chavarria S, Cole J, Hasegawa-Johnson M. Intertranscriber reliability of prosodic labeling on telephone conversation using toBI. In: Interspeech 2004 [Internet]. ISCA; 2004 [cited 2022 Oct 3]. p. 2729–32. Available from: https://www.isca-speech.org/archive/interspeech_2004/yoon04b_interspeech.html