Title: Cross-language differences in how voice quality and f_0 contours map to affect[1]

Authors: Irena Yanushevskaya[2], Christer Gobl, Ailbhe Ní Chasaide

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Dublin, Ireland

Phone: +353 1 8961348

[The manuscript uploaded to the JASA submission system: October 12, 2018]

Running title: Voice quality and cross-language affect perception

[1] Portions of this work were presented in 'Universal and language-specific perception of affect from voice', Proceedings of the XVIIth International Congress of Phonetic Sciences, Hong Kong, August, 2011.

[2] Author to whom correspondence should be addressed. Electronic mail: yanushei@tcd.ie

ABSTRACT

The relationship between prosody and perceived affect involves multiple variables. This paper explores the interplay of three: voice quality, f_0 contour, and the hearer's language background. Perception tests were conducted with speakers of Irish English, Russian, Spanish and Japanese using three types of synthetic stimuli: (1) stimuli varied in voice quality, (2) stimuli of uniform (modal) voice quality incorporating affect-related f_0 contours, and (3) stimuli combining specific non-modal voice qualities with the affect-related f_0 contours of (2). The participants rated the stimuli for the presence/strength of affective colouring on six bipolar scales, e.g., *happy-sad*. The results suggest that stimuli incorporating non-modal voice qualities, with or without f_0 variation, are generally more effective in affect cueing than stimuli varying only in f_0. Along with similarities in the affective responses across these languages, many points of divergence were found, both in terms of the range and strength of affective responses overall and in terms of specific stimulus-to-affect associations. The f_0 contour may play a more important role, and tense voice a lesser role in affect signalling in Japanese and Spanish than in Irish English and Russian. The greatest cross-language differences emerged for the affects *intimate, formal*, *stressed* and *relaxed.*

I. INTRODUCTION

This paper explores how the voice communicates affect and considers whether the language background of the listeners impacts on the affective colouring associated with variations in the voice. It looks specifically at how the dimensions of voice quality and f_0 map individually to affect and also at how certain combinations of these dimensions might work. Perception tests were carried out on native speakers of four languages, Irish English, Russian, Spanish and Japanese.

It is generally appreciated that voice quality plays a fundamental role in spoken communication, carrying an affective strand of information that conveys not only the speaker's emotion, mood and state, but also her/his attitude towards the interlocutor or the current situation. This aspect of speech communication has nonetheless remained elusive, being relatively neglected within the linguistic and speech sciences, and researched mostly within psychology, e.g. Scherer (2003). As man-machine communication is becoming ever more speech enabled, this aspect of spoken communication has moved centre stage – speech dialogue systems are increasingly needed that are sensitive to the affective dimension of the speaker's voice and generate speech with a voice quality that is accordingly modulated in appropriate ways (Burkhardt and Stegmann, 2009). Increasingly, the detection of the speaker's affective state is also important to applications monitoring affective and mental wellbeing (Cummins *et al.*, 2015).

Despite the increased research focus on this topic in recent decades, summarised in several reviews, e.g., Murray and Arnott (1993), Scherer (2003), Gobl and Ní Chasaide (2003b), Juslin and Laukka (2003), Juslin and Scherer (2005), Laukka (2008), Cowie

(2009), Kreiman and Sitdis (2011), Scherer (2013), reliable information on the specific voice source correlates of affect is relatively scarce. This area of research is beset with numerous conceptual and methodological difficulties. To begin with, defining affects/emotions is a complex issue (Cowie and Cornelius, 2003). A simple term like 'angry' can be interpreted to mean rather different emotions ('hot' and 'cold' anger are commonly differentiated in the literature, e.g., Juslin and Laukka (2001), but other states belonging to the same *family* of this emotion can be identified). Obtaining reliable speech samples with specific affects is difficult, particularly if one needs recordings that are amenable to voice source analysis. Likewise, the terms that are commonly used to describe voice quality are often used in vague and inconsistent ways in the literature – see discussion of this issue in Gobl and Ní Chasaide (2003b), although the use of Laver's (1980) classification framework enables a more rigorous treatment. Probably the greatest roadblock in the field concerns the difficulty in obtaining reliable, accurate production measurements of voice source features, particularly for non-modal voice qualities and in connected speech (Gobl and Ní Chasaide, 2010) – which is precisely the type of data that is needed. These issues are elaborated on in some detail in Section II below.

The picture is further complicated by the fact that the voice is not uniquely a carrier of affect: modulation of the voice source is an integral part of the linguistic (non-affective) prosody about which rather little is known – but see for example Heldner (2003), Iseli *et al.* (2006), Ní Chasaide *et al.* (2013), as linguistic-phonetic research on prosody is largely focused on f_0 modulation. Furthermore, the segmental context induces substantial microprosodic effects on the source (Ní Chasaide and Gobl, 1993) – and this can also be an important consideration when interpreting voice source data.

Cross-language studies (see Section II) typically entail the presentation of an utterance produced with affect in one language to listeners of other languages, who are asked to identify the affect they hear. This provides many insights but does not elucidate which aspects of the voice are cueing listener's perceptions. The present experiment follows a different approach, used in earlier studies by the authors to explore voice-to-affect mapping for speakers of Irish English, e.g., Gobl *et al.* (2002), Gobl and Ní Chasaide (2003b), Ryan *et al.* (2003). It entails the presentation of stimuli synthesised with different voice qualities to listeners of different language backgrounds and elicitation of the affective colouring they impart. The stimuli were constructed to provide exemplars of specific voice qualities chosen to provide a representative sample of qualities most frequently mentioned in the literature. The synthesised utterances incorporate the kinds of voice modulations that reflect the linguistic prosody as well as segment-dependent perturbations that have been alluded to above.

As in those earlier experiments, Laver's (1980) theoretical framework provides a basis for the synthesis of the voice quality stimuli used here: Tense Voice, Whispery Voice, Breathy Voice and Lax-creaky Voice. Laver's framework links auditory vocal qualities to their underlying physiological mechanisms (laryngeal tension settings) as well as to their main acoustic features. Past analytic studies by the authors of voice qualities produced within Laver's system have guided stimulus construction (Gobl, 1989; Gobl and Ní Chasaide, 1992; Ní Chasaide and Gobl, 1995). It is important to note that a specific voice quality is not a set point (like a cardinal vowel) but may vary on a continuum, and strength of the associated affect appears to vary accordingly (Ryan *et al.*, 2003). The voice quality used here are not extreme exemplars: the point was to

explore how moderate shifts, not necessarily extreme ones, in voice quality might influence perception of affect.

Differences in the f_0 contour (level, range and dynamics) have also been widely reported production correlates of affect, even though there are questions as to how perceptually important they may be (see Section II.C). To consider the role of f_0 in the present study, two further sets of stimuli were included: a series where the f_0 contours were varied and the voice quality kept constant and a series where these f_0 contours were combined with different voice qualities from the first series. The f_0 contours used here were based on production data, elicited for affective utterances of Dutch (Mozziconacci, 1995). The contours varied in terms of their level, range and dynamics. Unlike the voice quality series, it did include rather extreme contours, with large deviations from the neutral.

The two main research questions addressed by the present study are:
(1) how do voice quality and f_0 dimensions of the voice communicate affect?
(2) how does the language background of the listeners influence the mapping of voice to affect?

The underlying principle of this experiment was to present the listeners with a 'palette' of vocal stimuli and have them 'paint the picture' as to their affective colouring. Consequently, the focus was not on testing specific hypotheses. However, there are certain expectations regarding results, arising from our understanding of the physiological basis for the different voice qualities (following Laver's framework), and based also on the findings of earlier studies.

On the one hand, one would expect that tense voice (with high laryngeal tension settings) would be associated with high activation, and breathy (or lax) voice with low activation affects. This expectation is rather like the 'effort code' proposed to account for aspects of linguistic intonation by Gussenhoven (2004) which suggests that language prosody 'rules' are built on underlying production correlates. This might emerge therefore as an expected common feature, a potential candidate for universality. This expected trend did occur for tense voice in Gobl and Ní Chasaide (2003b), but in that and subsequent studies the lax-creaky setting was found to be rather more effective than breathy voice in conjuring low activation states, and so, both of these qualities are included here. As whispery voice entails more laryngeal tension than breathy voice, it was expected that this quality would be relatively less strongly associated with low activation. Language specific effects might also be expected (see Section II). For example, whispery voice has been mentioned as being associated with fear in English (Boula de Mareüil *et al.*, 2002), but with different qualities in other languages; breathy voice is traditionally associated with intimacy in English (Laver, 1980) but with formality/politeness in Japanese (Ito, 2004; Ishi *et al.*, 2008). Creaky voice tends also to be associated with different affective states in different languages.

The f_0 contours were based on affective production data, elicited in Dutch for a number of affects (sad, bored, etc.), and these affective labels are retained in this study. Given our 'broad palette' approach, we would refrain from strong predictions concerning how speakers of other languages might perceive these contours. As discussed in Section II, past studies have found such affective f_0 contours to be relatively ineffective in cuing the specific affects of the original production data. However, an expectation based on

the 'effort code' would be that the contours with the more extreme deviations from neutral would be associated with greater affective responses, e.g., the higher the f_0 contour the more strongly it would be associated with high activation states.

As regards the stimuli combining both voice quality and f_0 manipulations, the broad expectation would be that these dimensions would work synergistically, so that combined stimuli would be more effective in signalling affect than those stimuli involving simply voice quality or f_0 modifications on their own. Furthermore, there is a possibility that even though f_0 manipulations might not be effective in signalling the affect (of original production data) when in association with an appropriate voice quality setting such affects (sad, bored, etc.) would more clearly emerge. Although not generally attested in the earlier experiments on English, this could be the case for other languages.

Although the phonetic literature has typically assumed that specific voice qualities map directly to specific affects (e.g., creaky voice associates with boredom), our past experiments suggest that there is no one-to-one mapping of voice to affect. Rather, a single stimulus was found to be typically associated with several affects, not always related in any obvious way. As a corollary, a given affect might be conjured by more than one stimulus.

II. BACKGROUND AND LITERATURE REVIEW

Cross-language production and perception studies are discussed here, along with methodological difficulties that arise, particularly in elucidating the voice source correlates of affect. It is hoped that this broad review can help provide insight into the

complexities and difficulties pertaining to research in this field. It is also aims to provide a backdrop that explains the motivations for the present approach.

A. Expression of affect: universal or language specific?

As with the facial expression of emotion (Ekman *et al.*, 1987; Ekman, 1993), the spontaneous vocal expression of full-blown 'basic' (primarily negative) emotions is generally assumed to be universal (Scherer, 2000; Zinken *et al.*, 2008; Sauter *et al.*, 2010b; Scherer *et al.*, 2011; Laukka *et al.*, 2016) reflecting the influence of physiological 'push effects' (e.g., changes in rate of respiration and muscular tension) on the mechanism of voice production. However, in real interactions the affective expression of mood, attitude and interpersonal stance is not so extreme, involving much more subtle modulations of the voice than 'full blown' emotions and less directly linked to physiological 'push' effects.

Furthermore, the spontaneous expression of affect is constrained by 'pull effects' − external factors related to the socially accepted 'display rules', the listener's expectations, the speaker's self-presentation, etc. (Johnstone and Scherer, 2000; Juslin and Scherer, 2005). Cultural factors codified in language may thus intervene in determining how acceptable it may be to express or supress the display of affect (Mesquita and Walker, 2003; Oatley *et al.*, 2006). Complex cognitive states such as irony or sarcasm seem to involve intentional mismatching of voice and verbal content: these are particularly likely to be governed by culturally informed display rules (Zinken *et al.*, 2008; Sauter *et al.*, 2010b; Scherer *et al.*, 2011). The pragmatic context of the interaction is also important, regardless of culture. We can and routinely do control our

voice quality. As pointed out by Scherer (2013), the listener's attribution of affect and the speaker's underlying emotion do not need to coincide, and this is a feature of normal human interactions.

B. Production studies

1. Data elicitation

Obtaining reliable and valid affectively coloured speech data is not a straightforward task; e.g., discussion in Juslin *et al.* (2017). Naturally occurring spontaneous expression can be recorded during TV games or reality shows (Douglas-Cowie *et al.*, 2003) or in everyday social interactions (Campbell, 2002). However, the recording conditions and quality and the high variability of the spoken content tend to render these kinds of data unsuited to acoustic analysis that would isolate those voice source measures that can be correlated to affect. Speech samples obtained with mood induction techniques (Westermann *et al.*, 1996) allow control over the verbal and emotional content but produce relatively low in intensity and not clearly differentiated emotional expressions (Scherer, 2003).

The dominant approach remains the collection of acted expressions of affect (Bänziger and Scherer, 2007; Scherer, 2013; Laukka *et al.*, 2016). Typically, the utterances used are content-neutral, i.e. repetitions of a single word (e.g., a name) or specially constructed nonsense sentences (Scherer *et al.*, 2001). Such simulated data can produce intense prototypical expressions but allows experimental control over the lexical content of the utterances and affects expressed as well as cross-speaker comparison of results. A study on two emotions (*positive/happy* and *negative/sad*) by Scherer (2013) found no difference between acted and induced affective vocalisations. However, the

approach has been criticised (Bachorowski and Owren, 2003; Russell *et al.*, 2003; Juslin *et al.*, 2017) mostly on the grounds that portrayals may reflect exaggerated, culture-specific stereotypes rather than genuine real-life expressions.

2. Production correlates of affect

Not surprisingly therefore, empirical analytic studies to date have yielded relatively little information on the voice quality correlates of affect and have tended to focus rather on aspects that are more easily measured, particularly f_0 and intensity, as well as speech rate. (Speech rate emerges as important but is not covered in the present review.) Scherer and colleagues (Pakosz, 1983; Scherer *et al.*, 1984; Ladd *et al.*, 1986; Carlson *et al.*, 1992; Mozziconacci, 1995; Banse and Scherer, 1996; Mozziconacci and Hermes, 1999; Paeschke *et al.*, 1999; Mozziconacci, 2002; Paeschke, 2004; Bänziger and Scherer, 2005; Grandjean *et al.*, 2006) have provided extensive data on how the level, range and dynamics of f_0 and intensity are correlated with affect in speech production. Typically, global statements about these acoustic parameters are reported, e.g., greater level and range of f_0 for high activation affects (Banse and Scherer, 1996). However, when the perceptual relevance of production findings on f_0 differences was tested in a number of the above studies, results were disappointing, leading to a conclusion that the key to the differentiation of emotion lies in the voice quality (Scherer, 1986).

More recent studies (Patel *et al.*, 2011; Sundberg *et al.*, 2011) have emphasised the need to move towards more direct measurement of the voice source. Scherer (2013) has pointed out that the acoustic measures hitherto most commonly used in emotion research and inherited from phonetics, such as f_0 and intensity, are not in fact optimal indicators of the vocal changes that occur with affect. Unique acoustic patterns for

different emotions do exist, but the most appropriate acoustic measures to describe them have not yet been identified, suggesting the need for more complex acoustic measures, based on direct source measurements (Patel *et al.*, 2011; Bänziger *et al.*, 2015).

Direct measurements of the voice source are not easy to obtain, however, given the difficulty in obtaining the glottal flow signal itself. A non-invasive technique for estimating the glottal flow involves inverse filtering of the speech pressure waveform, whereby the effect of the vocal tract transfer function is cancelled (e.g., Gobl and Ní Chasaide, 2010; Alku, 2011). This type of analysis generally requires stringent recording conditions (Gobl, 2003; Alku, 2011), and as automatic methods tend to yield considerable errors, manual editing is ideally required – a technically complex and time-consuming task, which greatly limits the quantity of data that can be analysed (Gobl and Ní Chasaide, 2010). Indirect inferences about the source can also be made based on measurements carried out on the speech signal (Hanson, 1997; Hanson and Chuang, 1999; Keller, 2005; Alku, 2011). However, caution is needed in interpreting such data as they are influenced by the vocal tract filter; see discussion in Gobl and Ní Chasaide (2010).

The most commonly used indirect measures of voice quality in emotional speech research mainly describe spectral slope or spectral balance, i.e. the relative amount of acoustic energy above and below a certain cutoff frequency, e.g., the alpha ratio (Sundberg and Nordenberg, 2006) or the Hammarberg index (Hammarberg *et al.*, 1980), and also formant bandwidths (Juslin and Scherer, 2005; Laukka *et al.*, 2005; Goudbeek and Scherer, 2010; Laukka *et al.*, 2011). High activation or negative valence

states have been reported to have a less steep spectral slope (likely to be indicative of tenser phonation) while low activation or positive valence states show the opposite trend (Pittam *et al.*, 1990; Laukka *et al.*, 2005; Goudbeek and Scherer, 2010; Guzman *et al.*, 2013). Given the potential influence of the vocal tract filter setting, such measures may only be partially indicative of what is happening at the level of the voice source.

Given the technical complexity of voice source analysis, the quantity and scope of emotional speech voice source production data available is very limited. Most studies have tended to focus on a single vowel (or sometimes on a single glottal pulse) extracted from utterances produced with different affects, by typically one or very few speakers. Although the studies provide important insights, such measures cannot capture the important dynamic aspects of voice modulation in affect expression.

Analysis of glottal waveforms, reported in Cummings and Clements (1995), Laukkanen *et al.* (1996), Murphy and Laukkanen (2009) revealed significant differences in the glottal pulse characteristics/shape in speech produced with different speaking styles and different affective colouring. Similar findings were reported in a more extensive study (Patel *et al.*, 2011; Sundberg *et al.*, 2011), where several voice source parameters were measured in the [a] vowel, as produced by 10 speakers for a range of affective states. In a different approach, the dynamic variation across entire utterances was analysed for a single male speaker portraying a range of affective states (Yanushevskaya *et al.*, 2007; Yanushevskaya *et al.*, 2009). Results show distinct patterns of source parameter settings for each affect in terms of the within-utterance averages, but there was considerable overlap in parameter trajectories in parts of these utterances.

One of the issues arising in comparing different studies concerns the number and choice of source parameters measured. The 'global' parameter NAQ (Alku and Vilkman, 1996, see also Fant, 1995; 1997 on the global waveshape parameter R_d) intended to capture the overall characteristics of the glottal waveshape has been found to correlate to affect (Airas and Alku, 2004; Laukkanen *et al.*, 2008; Waaramaa *et al.*, 2008), particularly to the activation dimension.

Other studies like Sundberg *et al.* (2011), Patel *et al.* (2011), Cummings and Clements (1995), Murphy and Laukkanen (2009), Yanushevskaya *et al.* (2007), have measured more numerous, but not always the same, source parameters. This raises the question which pertains to all work on the voice source, be it directed at emotional expression or not, of what the optimal set of parameters might be. When large numbers of parameters are measured, there are considerable redundancies, as there are inevitable interdependencies among them. The issue of how many parameters are needed to capture production differences on the one hand and which can on the other hand lead to perceptually distinct changes to voice quality is one that still requires elaboration. It has been suggested by Eyben *et al.* (2016) that studies on the vocal expression of affect should all include comprehensive coverage of 65 acoustic features (not all of which are voice source measures), to ensure comparability across studies. While this is undoubtedly a desirable goal, when it comes to measurements of the voice source, a scattergun approach may be of little value if the accuracy of the source measurements is not assured or indeed for any given acoustic measure whose perceptual relevance has not been demonstrated. An interesting model in this context is that of Kreiman *et al.*

(2014), which aims to integrate voice production and perception, and whose acoustic parameters specifically aim to capture variation in voice quality.

In many studies there is an assumption that variation in a voice source parameter (or set of parameters) can be mapped directly to affect, and this assumption may not be borne out – for several reasons. NAQ is considered to correlate with audible variation in the tense-lax dimension of voice quality (higher NAQ indicating breather voice) (Airas and Alku, 2007) and as such, has been used as a possible source correlate of affect (Waaramaa *et al.*, 2008; Patel *et al.*, 2011). However, as pointed out in Gobl and Ní Chasaide (2003a) a relatively high NAQ value is not invariably indicative of a relatively breathier voice quality and the correlation may only pertain to phonation in a given f_0 range. And, as mentioned above, in interpreting voice source measures, the non-affective (linguistic-prosodic and segment-related) sources of voice source variation need to be borne in mind, as these are not perceived by the listener as shifts in auditory voice quality. Ultimately, voice source variation is unlikely to express affect unless there is a perceptible shift in the auditory voice quality – hence the approach adopted in this study.

3. Cross-language production studies

Several cross-language production studies suggest that basic emotions (in acted portrayals) may not be encoded universally (at least based on the acoustic parameters measured). Anger is usually described as having high f_0 and high intensity, e.g., in German (Banse and Scherer, 1996) and English (Williams and Stevens, 1972); see also Juslin and Scherer (2005), Kappas *et al.* (1991), Juslin and Laukka (2003) for other languages. This was found also in Russian but not in Estonian, where pitch and

intensity were found to be lower in the expressions of anger than in the neutral state (Altrov, 2013). Significant cross-language differences in f_0 mean and range in the expression of anger have been reported in Pell *et al.* (2009a), Pell *et al.* (2009b). Laukka *et al.* (2016) reported significant effects of language in the analysis of affective speech for many types of acoustic cues (f_0, intensity, spectral balance, and temporal characteristics).

Auditory judgements of portrayals of the same emotion across-languages suggest differences in the voice qualities used. For example, fear was reported to be expressed with falsetto voice by German speakers (Burkhardt and Sendlmeier, 2000; Schröder, 2001), but with whispery voice by an English actor (Boula de Mareüil *et al.*, 2002) and with breathy voice in Italian (Drioli *et al.*, 2003). Similarly, the same voice quality may carry quite different connotations for speakers with different language backgrounds. Thus, creaky voice tends to be associated with boredom in English and German (Laver, 1980; Burkhardt and Sendlmeier, 2000). In Japanese, where the use of pressed or creaky voice ('rikimi') follows very complex rules of social interaction, creaky voice has been reported to convey a range of affective states, e.g., surprise, irritation, disgust, anger and admiration (Sadanobu, 2004). In Chinese, creaky voice has been found in the expression of anger (Yuan *et al.*, 2002), while in Italian a creaky voice quality was found in expressions of disgust (Drioli *et al.*, 2003).

While these observations are possibly indicative of language-specific vocal expression of emotion, one must be cautious in interpreting such apparent differences. There is often uncertainty as to whether a given impressionistic voice quality label refers to the same acoustic-auditory phenomenon in different studies. Can one be sure that the

breathy voice and the whispery voice (reported for expressions of fear for the Italian and English speakers respectively) are in fact different qualities? Similarly, is the creaky voice reported as associated with boredom in English and German in fact the same quality as the creaky voice of Japanese 'rikimi'? Furthermore, even in a single language a particular voice quality can be associated with a number of different affects and a particular affect (such as sadness) can be indicated by more than one voice quality. The lack of bi-uniqueness raises the possibility that such cross-language differences might sometimes be more apparent than real – particularly as the context and methodologies of different studies can differ considerably.

C. Perception studies

Although many production studies have described shifts in the level, range and dynamics of f_0 in affectively coloured speech (Scherer *et al.*, 1984; Ladd *et al.*, 1985; Banse and Scherer, 1996; Grandjean *et al.*, 2006), these features alone do not generally appear very effective in cueing affect (Mozziconacci, 1998; Bänziger and Scherer, 2005). Such lack of correspondence of the production and perception findings has led to the overall conclusion, mentioned above, that these are not the optimal acoustic features, and that more complex, voice source measures are needed to capture how listeners decode emotion (Scherer, 2013; Bänziger *et al.*, 2015). Not surprisingly, given the sparsity of production data on affect-related voice source features, rather little is known about their influence/salience in affective cueing.

As described in the Introduction, studies exploring the perceptual relevance of voice quality in the cueing of affect were carried out by the authors with Irish English

participants (Gobl and Ní Chasaide, 2000; Gobl *et al.*, 2002; Gobl and Ní Chasaide, 2003b). These are the precursors to the present cross-language investigation.

1. *Cross-language perception studies*

Unlike production studies, the preponderance of research on perception has been directed at cross-language aspects. The standard paradigm uses acted portrayals of a limited set of discrete emotion categories on content free utterances. Perception tests entail forced-choice judgement among the emotion categories being tested and report the accuracy with which the intended expressed emotions are identified (Bänziger *et al.*, 2015). Typically, individuals from many language/ethnic groups judge stimuli produced in one language, individuals from one language/ethnic group judge stimuli from many languages or a balanced design is used (Laukka *et al.*, 2016). Many such cross-language perception studies suggest broad similarities in affect perception from voice by listeners with different language background, e.g., McCluskey *et al.* (1975), McCluskey and Albas (1981), van Bezooijen *et al.* (1983), Scherer (2000), Graham *et al.* (2001), Pell and Scorup (2008), Koeda *et al.* (2013), Waaramaa and Leisiö (2013), Waaramaa (2014), see also review of perception studies in Laukka *et al.* (2016).

In general, it is reported that negative emotions tend to be recognised more accurately cross-linguistically than positive emotions (Sauter *et al.*, 2010a; Sauter *et al.*, 2010b) and basic emotions such as sadness and fear are recognised more accurately compared to more complex cognitive affective states, such as contempt or sarcasm, e.g., van Bezooijen *et al.* (1983), Cheang and Pell (2008), Sauter *et al.* (2010a).

It has been suggested that the same emotion may be expressed 'more clearly' and recognized with higher accuracy in certain languages; for example, in the studies by McCluskey *et al*. (McCluskey *et al.*, 1975; McCluskey and Albas, 1981) the Mexican subjects were found to be significantly more sensitive to expressions of happiness, sadness, love, and anger in speech than their Canadian counterparts. Emotions expressed by Mexicans were generally recognised with higher degree of accuracy by both Mexicans and Canadians than those expressed by Canadians.

As suggested in Kitayama and Ishii (2002), certain cultures may be more sensitive to vocal expression, while for others, the propositional content matters more than the tone-of-voice. Furthermore, due to in-group advantage, listeners tend to recognise emotions vocally expressed in their native language with more accuracy than those expressed by representatives of other cultures (Elfenbein and Ambady, 2002b; a; Bryant and Barrett, 2008).

Language relatedness may be a factor in the accuracy of cross-language emotion attribution. In a large-scale study (Scherer, 2000; Scherer *et al.*, 2001), listeners from differing language backgrounds judged simulated emotions produced by German actors in artificially constructed sentences. The greater the language dissimilarity from German, the less accurate the identification of emotion was. Similar findings are reported more recently by Waaramaa and Leisiö (2013) and Waaramaa (2014). However, no such advantage of the language/culture proximity was found in Pell *et al.* (2009a) and Altrov and Pajupuu (2015).

These studies do not clarify which aspects of the voice signal affect within a language group, or whether these differ across languages. Perception tests by Burkhardt *et al.* (2006) address such questions by presenting listeners with stimuli of MBROLA-synthesised, emotionally loaded sentences, where pitch range, duration and jitter have been manipulated. Listeners with different language backgrounds (French, German, Greek and Turkish) showed differences in sensitivity to the jitter and pitch range. A narrow pitch range was associated with a friendly affect by Turkish listeners, but not by the speakers from other countries. Turkish listeners also associated an increased amount of jitter with a threatening affect – an association not made by German and French listeners: for the French listeners the jitter-simulation evoked a frightened affect, whereas the converse was found for the Germans.

As detailed in the Introduction, the study presented here differs from the above approaches by presenting listeners from differing language backgrounds with stimuli containing auditorily distinct voice qualities in a semantically neutral utterance and eliciting how they are associated with affect by the different groups. In addition to stimuli differing in voice quality, stimuli with different f_0 contours (all with modal voice) were also included, as well as combined stimuli where voice quality and f_0 contours were varied. Of interest is to see how similar/different these language groups might be in associating the different dimensions of the voice to affect. Although not testing specific hypotheses, our expectations in terms of likely/possible results are described in Introduction.

III. MATERIAL AND METHOD

A. Synthesised stimuli

The stimuli used in this cross-language study are based on a recording of a short Swedish utterance ‘ja adjö’ [ˈjɑː aˈjøː] produced by a male speaker with a voice quality conforming to Laver’s (1980) description of modal voice. To create the modal voice stimulus, high quality copy synthesis of this natural utterance was carried out using the KLSYN88 formant synthesiser (Sensimetrics Corporation, Boston, MA, described in Klatt and Klatt, 1990) employing its version of the LF model for the glottal waveform generation. By modifying the modal voice stimulus, three series of distinct stimulus types were generated, referred to as the VQ stimuli, the F0 stimuli, and the VQ+F0 stimuli, as outlined in the following sections.

To obtain the source and filter settings used in the synthesis for the modal voice stimulus, the utterance was inverse filtered and the source signal modelled by matching the LF model (Fant *et al.*, 1985) to the output of the inverse filter. Both processes were carried out using the manual, interactive technique described in Gobl and Ní Chasaide (1999) and Gobl and Ní Chasaide (2010), which allows pulse-by-pulse optimisation of the parameter settings.

1. *VQ stimuli*

The non-modal voice qualities used in this study were *Whispery Voice, Breathy Voice, Lax-creaky Voice and Tense Voice*. As mentioned earlier, they were not extreme exemplars of these qualities. With some minor modifications, these stimuli are as described in detail in Gobl and Ní Chasaide (2003b) and they were generated from the modal voice stimulus by manipulating the following parameters of the KLSYN88

synthesiser: fundamental frequency (F0), amplitude of voicing (AV), spectral tilt (TL), open quotient (OQ), speed quotient (SQ), aspiration noise (AH), diplophonia (DI), and the bandwidth of the first formant (B1). The dynamic variation of parameters was determined by earlier analytic studies of specific voice qualities (Gobl, 1988; Gobl and Ní Chasaide, 1992; Ní Chasaide and Gobl, 1995) and was adjusted on the basis of auditory analysis of the voice qualities (Gobl and Ní Chasaide, 2003b). Informal evaluation of the VQ stimuli by phoneticians trained in Laver's (1980) framework was used to ensure that the intended voice quality was achieved. The parameter variation of the VQ stimuli is illustrated in Figure 1.

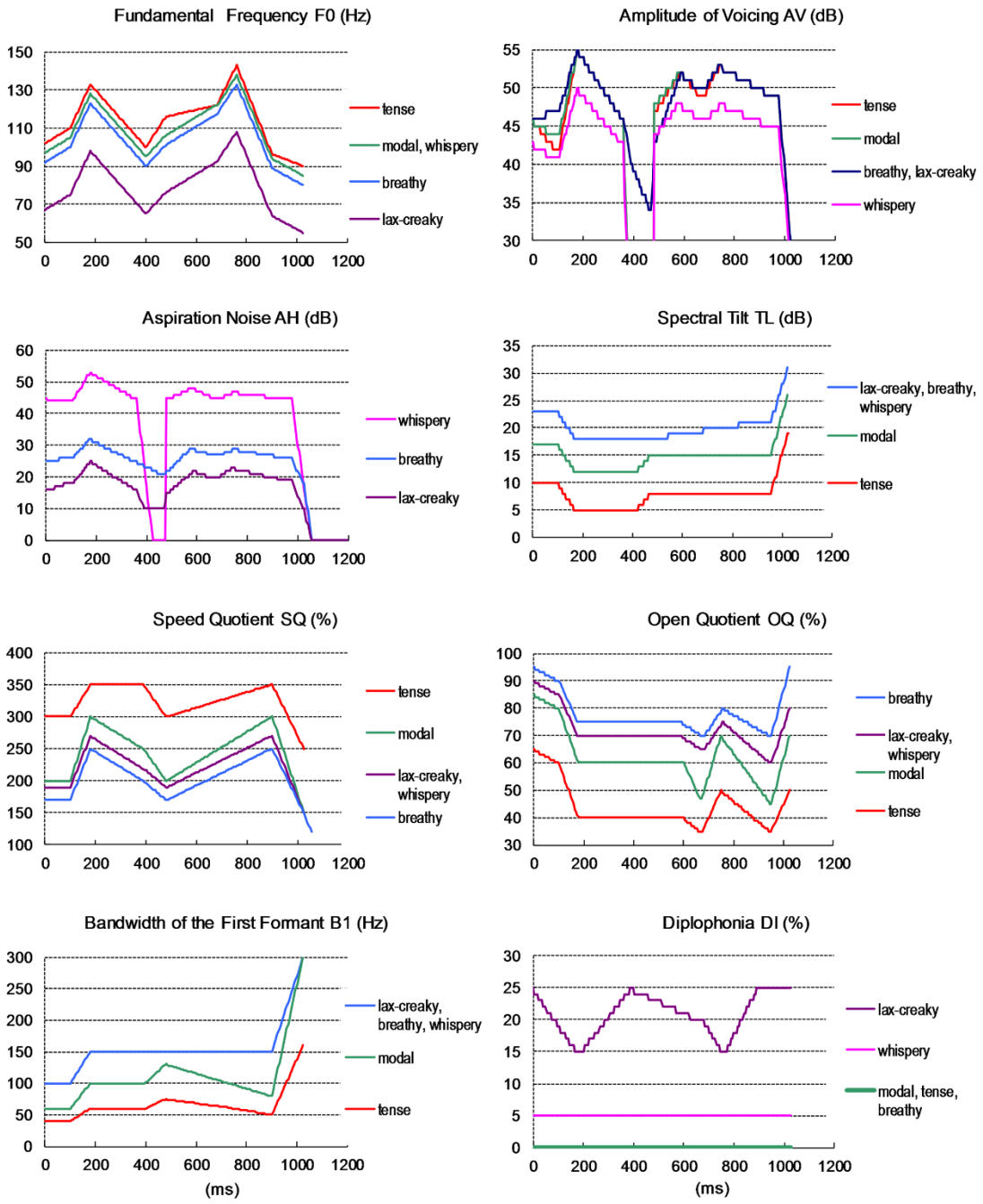


Figure 1 (color online). Parameter variation in the synthesised voice quality stimuli (after Gobl & Ní Chasaide, 2003).

The synthesised voice quality stimuli in Gobl and Ní Chasaide (2003b) also incorporated minor adjustments of f_0 values, deemed likely to be typical of these qualities. For example, f_0 is 5 Hz higher for the *Tense Voice* stimulus and 5 Hz lower

for the *Breathy Voice* stimulus compared to *Modal Voice*. In one case the f_0 difference was more substantial: for *Lax-creaky Voice* an f_0 value 30 Hz lower than that of the *Modal* (neutral) stimulus was used, as this quality most typically occurs with low f_0. While this feature was retained for the present experiment, it resulted in a certain anomaly necessitating consideration for the presentation of results. This is discussed further in Section III.A.3.

As can be seen in Figure 1, the trajectories for all source parameters show considerable utterance-internal dynamic variation. These reflect the variation due to the linguistic prosody of the utterance (Ní Chasaide and Gobl, 2004b) as well as perturbations which are the consequence of consonantal articulation (Ní Chasaide and Gobl, 1993) and these have been discussed in Section II. These variations were retained in the other VQ stimuli.

2. *F0 stimuli*

The creation of this series of stimuli entailed modifications to the *Modal* stimulus by manipulating the F0 parameter of the KLSYN88 synthesiser. Five further stimuli were produced with different f_0 contours by adapting the affect-related f_0 contours presented in Mozziconacci (1995). The latter were based on the analysis of a short semantically neutral utterance (containing two accented words) produced by a male Dutch speaker portraying indignation, joy, fear, boredom, sadness as well as a neutral state.

As the f_0 contour of the modal utterance used in this study was very similar to the neutral contour of Mozziconacci (1995), it served as a neutral reference f_0 contour. The five additional contours were obtained by proportional scaling of the f_0 values of our reference stimulus at nine anchor points, using scale factors derived from

Mozziconacci's contours. The contours are shown in Figure 2, and are labelled as follows: *F0 fear, F0 joy, F0 sadness, F0 boredom, F0 indignation,* and *F0 neutral.*

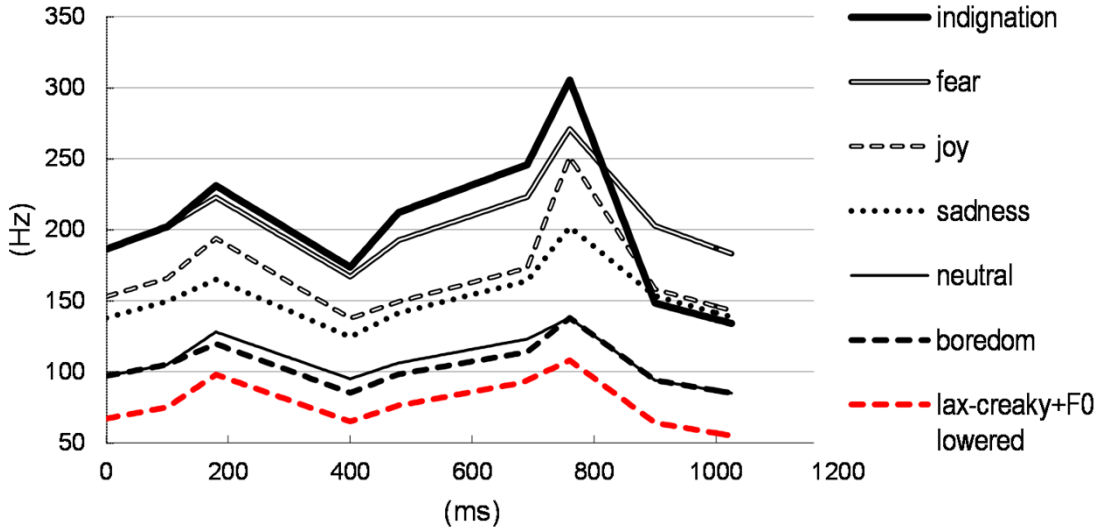


Figure 2 (color online). f_0 contours used in the F0 stimuli. The lowest contour shows the f_0 contour used with the lax-creaky voice quality (see text).

It should be noted that in selecting these contours, and in labelling them according to the affective states associated with them in Mozziconacci's (1995) study (i.e. *F0 fear, F0 joy*, and so on) no assumption is being made that these f_0 contours should necessarily be associated with those particular affective states by the speakers of the languages tested in this study. Rather, they can be regarded simply as a set of f_0 contours that reflect likely affect-related variations, and which encapsulate the kinds of shifts in f_0 level, range and dynamics that are often mentioned in production-based studies.

The f_0 contours have two pitch accents and the overall shape of the contours is characteristic of declaratives in all four languages selected for this study. These f_0 patterns differ not in terms of the basic contour shape, but in terms of f_0 level, range and

dynamics – the kinds of features generally reported in studies on f_0 and emotion (Bänziger and Scherer, 2005; Grandjean *et al.*, 2006). The contours for *boredom* and *neutral* are characterised by a low f_0 level and relatively small f_0 excursions. *Sadness* is represented by a higher f_0 level; *joy, fear* and *indignation* are all characterised by increasingly raised f_0 levels, and by an extended dynamic range, particularly in the second peak. *Indignation* has the very highest f_0 level and a rapidly falling f_0 in the final accent (see Figure 2).

3. *VQ+F0 stimuli*

The stimuli of the third type were generated by combining non-modal voice qualities with the affect-related f_0 contours used in the synthesis of the F0 stimuli. The combination of particular voice qualities with potentially affect-related f_0 contours was guided by the results of the earlier experiments (Gobl *et al.*, 2002; Gobl and Ní Chasaide, 2003b) as well as by the literature, e.g., Laver (1980), Scherer (1986; 2003), Juslin and Laukka (2001; 2003) – the latter is a meta-analysis of 104 studies of vocal expression of emotions, a summary/review in Schröder (2001). Anger, joy and fear tend to be associated with high f_0 mean, variability and range, and generally, sadness and boredom are associated with relatively lower f_0 mean, variability and range. According to Laver (1980), breathy voice has been associated with intimacy and creaky voice with boredom. Anger and joy are characterised by narrow B1 and 'steep' glottal waveform and an increase in high frequency spectral energy (suggesting tense phonation), while fear and sadness by broad B1 and 'rounded' glottal waveform (suggesting laxer, breathier phonation) (Scherer, 1986; Juslin and Laukka, 2003). Whispery voice was found effective in signalling fear in Gobl and Ní Chasaide (2003b); whispery voice combined with high dynamic f_0 was found in the expression of fear in a production study by Klasmeyer and Sendlmeier (2000).

Thus, *Whispery Voice* was combined with the *F0 fear* contour (referred to as the *Whispery+F0 fear* stimulus), *Breathy Voice* with *F0 sadness* (*Breathy+F0 sadness*), *Lax-creaky Voice* with *F0 boredom* (*Lax-creaky+F0 boredom*). *Tense Voice* was combined both with *F0 joy* (*Tense+F0 Joy*) and with *F0 indignation* (*Tense+F0 indignation*).

As constructed, a potential anomaly arose for the VQ and VQ+F0 stimuli which had a lax-creaky voice quality. The originally constructed *Lax-creaky* stimulus in the VQ series had an f_0 level which deviated considerably from neutral being 30 Hz lower, whereas the combined stimulus *Lax-creaky+F0 boredom* had an f_0 level very close to the neutral. On listening to these two stimuli, it was clear that the lax-creaky quality was fully audible in the stimulus even where f_0 was close to the neutral value. Insofar as the VQ+F0 stimuli are intended to test for the effect of large f_0 deviations from neutral coupled to particular voice qualities (as compared to voice quality alone), it would appear that these two stimuli are the 'wrong way around'. To remedy this anomalous situation, and to make the presentation of results more straightforward, in the following sections, these two stimuli are transposed, i.e. the *Lax-creaky+F0 boredom* stimulus is treated as belonging to the VQ series of stimuli (as all stimuli in this series entail only minor f_0 deviations from the neutral) and it will be referred to henceforth as *Lax-creaky*). The *Lax-creaky* stimulus with the large f_0 lowering is more appropriately treated as belonging to the VQ+F0 stimuli and is renamed as *Lax-creaky+F0 lowered*. This re-classification allows one to assess more directly the contribution of the f_0 lowering to the affective signalling of this voice quality, in a way that retains a certain consistency with the overall differentiation of the three stimulus types used here.

Table I provides a summary of the stimuli used in the perception experiments of this cross-language study. Based on the voice quality and f_0 contour combinations in the VQ+F0 series of stimuli, it is convenient to group the stimuli into five Stimulus Groups, each of which containing one stimulus from each of the three stimulus types. The five groups are named according to the voice quality and f_0 contour that characterise them. As mentioned in the discussion of the F0 stimuli, despite the use of labels such as 'sadness', there is no strong expectation that a stimulus thus labelled will yield the highest ratings for the particular emotion, e.g., sadness. In addition to the five groups, the stimulus with modal voice and the neutral f_0 contour provides a baseline to which responses for all other stimuli can be compared.

Table I. The synthesised stimuli used in the cross-language study. All F0 stimuli have modal voice. All VQ stimuli have neutral F0. Additionally, the stimulus *Modal+F0 neutral* is included for baseline comparison. For the stimuli with * see text.

Stimulus Group	**Stimulus Type** (type of manipulation)		
	VQ	F0	VQ+F0
WHISPERY‖FEAR	*Whispery*	*F0 fear*	*Whispery+F0 fear*
BREATHY‖SADNESS	*Breathy*	*F0 sadness*	*Breathy+F0 sadness*
LAX-CREAKY‖BOREDOM	*Lax-creaky**	*F0 boredom*	*Lax-creaky+F0 lowered**
TENSE‖JOY	*Tense*	*F0 joy*	*Tense+F0 joy*
TENSE‖INDIGNATION	*Tense*	*F0 indignation*	*Tense+F0 indignation*

Baseline Stimulus: ***Modal+F0 neutral***

B. Listening tests

1. *Procedure*

The perception experiments were conducted according to the procedure described in Gobl and Ní Chasaide (2003b), as a series of six subtests. In each subtest, the 15 stimuli were presented to the participants 10 times in random order and responses were obtained for a pair of opposite affective attributes (e.g., *happy-sad*). The stimuli were presented to the listeners in a quiet room through a high quality speaker with the volume set at the comfortable listening level. Each subtest took approximately 11 minutes. The subjects had a short training session before the tests and were given breaks between the subtests.

2. *Rating scales*

The participants were asked to judge the stimuli on six bipolar scales defined with contrastive adjectives at each end: *indignant-apologetic, interested-bored, formal-intimate, stressed-relaxed, happy-sad* and *fearless-scared.* A similar approach where stimuli were rated for affective content on bipolar scales was used, for example, in Uldall (1964) and Ladd *et al.* (1985). This kind of semantic differential scale is commonly used in the study of attitude (Heise, 1970; Osgood *et al.*, 1975; Russell and Carroll, 1999; Streiner and Norman, 2008), and allows one to measure directionality of reaction (e.g., bored vs. interested) as well as intensity (slight to extreme). The scale is usually interpreted as a seven point scale where the neutral attitude (or in our case, 'no affective colouring') is assigned the value of zero (Heise, 1970). The data obtained by

this kind of 'summative response scale' (Gamst *et al.*, 2008) can be analysed statistically using a general linear model, e.g., ANOVA.

The participants were instructed to judge each stimulus for the presence and strength of affect, and to mark their response on the answer sheet where the opposite affective labels were placed on each side with seven boxes in between. The choice of the centre box implied that no affective colouring was present in the utterance; checking the boxes to the left or right to the centre box indicated the presence and strength of a particular affect, the most extreme ratings being further from the centre box. The scale was later interpreted as ranging from -3 (low activation states) to +3 (high activation states), where 0 value corresponded to no affect perceived, ± 1 corresponded to mild affective colouring, ± 2 – to moderate and ± 3 – to strong affective colouring. Numeric values were not presented to the raters during the experiment.

The affective labels defining the opposite ends of each of the six scales have been chosen to cover a fairly broad range of emotions and milder affective states such as attitudes and interpersonal stances. The affective adjectives used in the scales are among those most frequently found in the lists of affect-related words (Juslin and Laukka, 2003). The choice of affective labels was in part guided by the synthesised voice qualities and by what is known about voice-to-affect mapping in the literature. The scales *stressed-relaxed, happy-sad, interested-bored, formal-intimate* were used in prior work and were adopted here in part to assure comparability of results.

3. *Translation of affective labels*

The stimuli were presented to speakers of four languages - Irish English, Russian, Spanish and Japanese. For the speakers of languages other than English the affective

labels on the answer sheets were given in their respective languages in translation. To render the translations as accurately as possible, we used team translation or the 'committee approach' that also involved back translation as recommended in Brislin (1980) and Streiner and Norman (2008). The translation of the labels from English was undertaken by at least two native speakers of the respective languages (university students and staff) who had a good command of English and who were also familiar with the nature of the research and the purpose of the rating scales. As there are usually several translation possibilities, the translators were asked to discuss them and come to consensus regarding the best possible choice. The affective labels used in the listening tests are shown in Table I in supplementary material[1].

4. *Participants*

The participants were speakers of Irish English, Russian, Spanish and Japanese. As it was desirable to exclude the influence of semantic content of the Swedish utterance used as the basis for the synthesised stimuli on the perception of affect, we made sure prior to the experiments that none of the participants spoke Swedish. It is virtually impossible in the modern world to completely exclude the influence of foreign languages and with this the influence of language-specific 'foreign' voice qualities, but such an influence was kept to a minimum in the present experiments as none of the participants had lived in a foreign country for any length of time or used foreign languages in an active way in their professional life (e.g., as an interpreter or a language teacher). The participants were college students recruited in Dublin (n = 20, 13 female, aged 18-35), St. Petersburg (n = 21, 10 female, aged 19-35), Barcelona (n = 20, 10 female, aged 18-22) and Tokyo (n = 21, 10 female, aged 18-22).

C. Statistical analyses

To compare the ratings of the three types of stimuli (the VQ stimuli relative to the VQ+F0 and the F0 stimuli) within each stimulus group taking the language factor into account, a separate repeated measures ANOVA was conducted for each subtest using the SPSS software package (IBM Corp., 2013).

A complex mixed design was applied. The two within-subjects factors were the Stimulus Group (5 levels: WHISPERY||FEAR, BREATHY||SADNESS, LAX-CREAKY||BOREDOM, TENSE||JOY, TENSE||INDIGNATION) and Stimulus Type (3 levels: VQ, F0, VQ+F0). The between-subjects factor was Language (4 levels: Irish English (E), Russian (R), Spanish (S), Japanese (J)). The dependent measure was mean affective rating (averaged across 10 randomisations for each speaker).

As *Tense Voice* was used twice, in the generation of two VQ+F0 stimuli in two Stimulus Groups, TENSE||JOY (*Tense+F0 joy*) and TENSE||INDIGNATION (*Tense+F0 indignation*), it was not possible to apply the 5 × 3 × 4 factorial design to the data. In each subtest, the analysis of variance was therefore done twice as a 4 × 3 × 4 design (4 Stimulus Groups × 3 Stimulus Types × 4 Languages), first for the stimuli in the Stimulus Groups WHISPERY||FEAR, BREATHY||SADNESS, LAX-CREAKY||BOREDOM and TENSE||JOY (Part I of the ANOVA) and second, for the stimuli in the Stimulus Groups WHISPERY||FEAR, BREATHY||SADNESS, LAX-CREAKY||BOREDOM and TENSE||INDIGNATION (Part II of the ANOVA). Modal voice was not included in the model.

To assess the ratings of VQ stimuli relative to *Modal* and the ratings of F0 and VQ+F0 stimuli relative to *Modal* (with neutral f_0), a separate repeated measures ANOVA test of

simple contrasts was conducted for each subtest in which pairwise comparisons of all the stimuli were carried out. The within-subject factor was Stimulus, the between-subject factor was Language. The dependent variable was the mean affective rating for each participant (averaged across 10 randomisations). To correct for multiple comparisons and to minimise the risk of random effects, a Bonferroni correction was applied in the post-hoc tests.

There are multiple sources of variability in voice stimuli ratings, such as listener factors (sensitivity, bias, error, fatigue, etc.), scale factors (scale resolution and the way the scale is defined), stimulus factors (the quality of voice samples or synthesised stimuli), as well as the interaction of these factors (Kreiman *et al.*, 2007). Consistency in assignment of similar ratings to the same stimulus will suggest stronger voice-to-affect association. Listeners' agreement/consistency in ratings was measured by Intraclass Correlation Coefficients (ICC), the two-way random model (Landis and Koch, 1977; Shrout and Fleiss, 1979; McGraw and Wong, 1996), which were calculated for each subtest for each language group of listeners. As it is of interest to establish whether the judgment of one rater is similar to that of the others, the single measures ICC (*r*) rather than the average measures ICC (*R*) will be considered here. The interpretation of ICC is based on Landis and Koch (1977): < 0 poor agreement, 0-0.20 slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, 0.81-1 almost perfect agreement.

IV. RESULTS

The statistical results are outlined and following a broad overview the detailed findings for the individual stimuli are summarised.

A. Statistical results

1. *Rater's agreement*

The participants' agreement, calculated as single measure (*r*) Intraclass Correlation Coefficients, overall was within the moderate to substantial range (moderate in 12 out of 24 cases, substantial in 6 out of 24 cases). Two groups of listeners, Japanese and Spanish, showed fair agreement in the *formal-intimate* and *fearless-scared* subtests. Spanish listeners also showed fair agreement in the *indignant-apologetic* subtest. Furthermore, the agreement of the Japanese listeners was poor in the *stressed-relaxed* subtest. Going from the greatest to the lowest degree of interrater agreement, the order is R > E > S > J (see supplementary material, Table II[2]).

2. *ANOVA results*

The results of the omnibus complex design ANOVA are given in supplementary material, Table III[3]. The results for the within-subject effects - part (a) of the ANOVA - showed significant effects of Stimulus Type and Stimulus Group in all subtests. Two-way Stimulus Group × Language and Stimulus Type × Language interactions were largely significant, except in *happy-sad* and *fearless-scared* subtests. A three-way Stimulus Group × Stimulus Type × Language interaction was found significant in all but two subtests, *indignant-apologetic* (Part I) and *happy-sad* (Part II). Given the interactions, it is not possible here to assess the relative contribution of individual factors. Thus, for a closer look at cross-language differences, the pairwise comparisons of the individual stimuli are also included.

In part (b) of the ANOVA, the between-subject effect of Language was found to be significant in the two subtests: *formal-intimate* and *stressed-relaxed*. It is for these affects, therefore, that we would expect to see the greatest differences among languages in terms of voice to affect association.

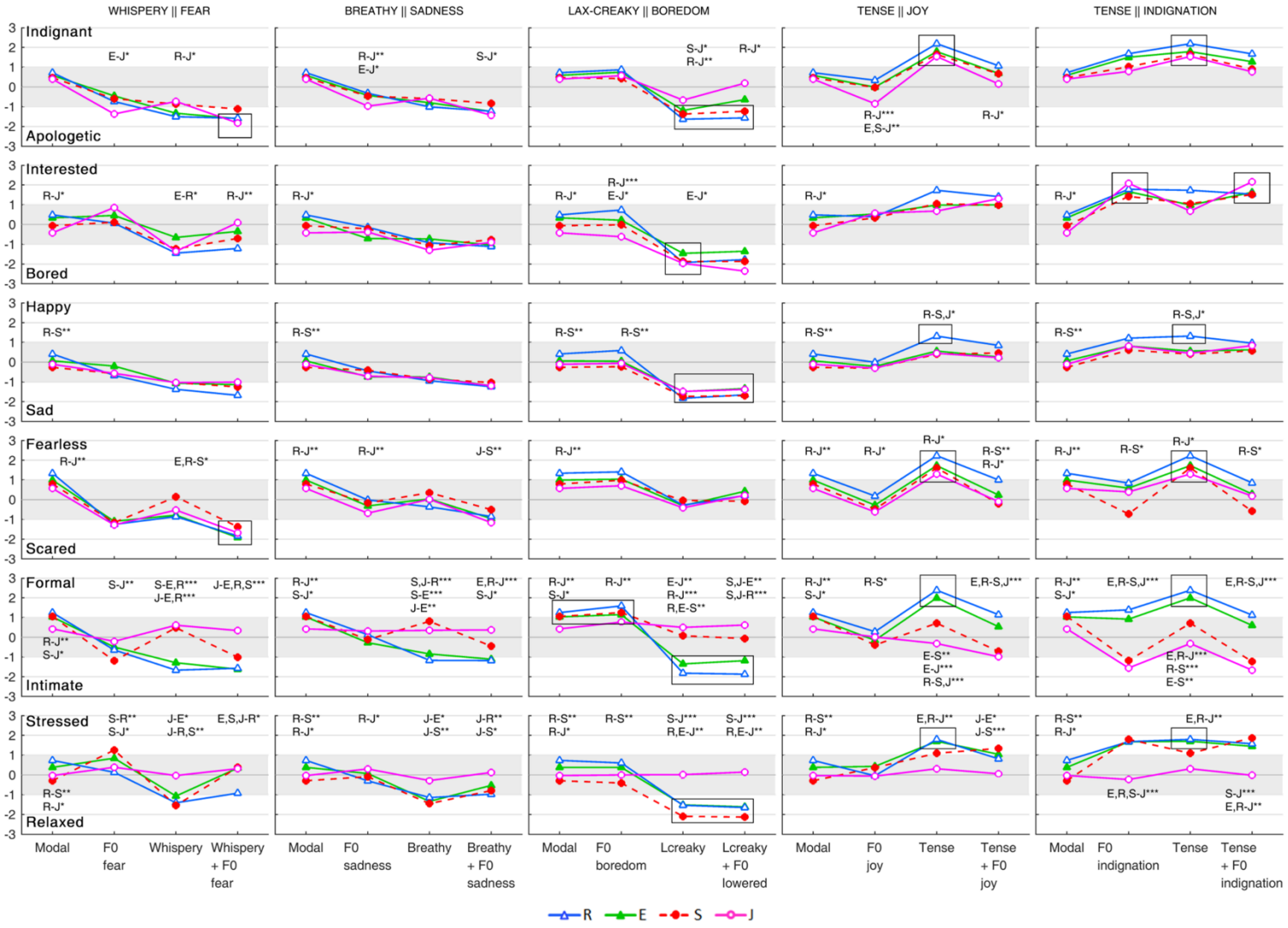
Details of statistically significant cross-language differences of the ratings for the individual stimuli obtained in the ANOVA test are shown in Figure 3 (explained in detail below). Pairwise comparisons of within-language differences in ratings for the different stimuli are not presented here due to space constraints; note that the differences in affective ratings for *Modal* and *Modal+F0 boredom* and for *Lax-creaky Voice* and *Lax-creaky+F0 lowered* were not found to be statistically significant in any language for any affective subtest.

B. Voice-to-affect mapping: broad overview

The five columns in Figure 3 show the ratings obtained organised in terms of the Stimulus Groups (see also Table I). Each row presents plots with the ratings for one of the six affective subtests (*indignant-apologetic, etc.*). In each plot, the rating for the stimulus *Modal Voice (+F0 Neutral)* provides a baseline to which ratings for other stimuli can be compared. For each affective subtest, the most highly rated stimulus is marked by a surrounding box. More than one box is used when two stimuli produce similar ratings or where there are differences among the languages in terms of their most highly rated stimulus.

In the individual plots positive values (0 to +3) correspond to high activation/power states (*indignant, interested, happy, fearless, formal, stressed*), while negative values (0

to −3) correspond to low activation/power states (*apologetic, bored, sad, scared, intimate, relaxed*). As in previous studies (e.g., Yanushevskaya *et al.*, 2013) discussion focuses on ratings that exceed ±1, and this region of weak affect is shaded grey. This ±1 threshold is admittedly arbitrary and indeed, statistically significant differences can occasionally be found between ratings that are low, but the intention here is to concentrate on the more salient voice-to-affect associations. Significant cross-language differences are indicated by asterisks (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).


WHISPERY || FEAR
BREATHY || SADNESS
LAX-CREAKY || BOREDOM
TENSE || JOY
TENSE || INDIGNATION
Indignant
Apologetic
Interested
Bored
Happy
Sad
Fearless
Scared
Formal
Intimate
Stressed
Relaxed
Modal
F0 fear
Whispery
Whispery + F0 fear
F0 sadness
Breathy
Breathy + F0 sadness
F0 boredom
Lcreaky
Lcreaky + F0 lowered
F0 joy
Tense
Tense + F0 joy
F0 indignation
Tense + F0 indignation
R
E
S
J

1

Figure 3 (color online). Affective ratings obtained in the six subtests: each row represents a listening test (e.g., *indignant-apologetic*), each column represents a particular stimulus group (e.g., WHISPERY||FEAR). Languages: Irish-English (E); Russian (R), Spanish (S), Japanese (J). Statistically significant cross-language differences: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The most highly rated stimulus for a particular affect is marked by a surrounding box.

A broad indication of the extent of cross-language differences/similarities can be gleaned from Figure 3. Two of the affective subtests stand out as having large statistically significant differences. In the *formal-intimate* subtest, languages patterned in terms of two groups, R and E in one, and S and J in the other. In the *stressed-relaxed* subtest, J diverges from the other three languages. In the remaining subtests results appear rather similar overall in terms of voice to affect association, although there are many cases (indicated with asterisks in Figure 3) where the same stimulus yields affective ratings of significantly different strength from listeners of different language groups.

Of the five Stimulus Groups, TENSE||INDIGNATION accounts for the highest ratings obtained for high activation/power states. Within this group, the VQ stimulus (*Tense Voice*) is particularly effective. The F0 stimulus (*F0 indignation*) is strongly associated with *interested* across languages but shows up prominent cross-language differences in its association with other affects – differences which persist in the combined VQ+F0 stimulus (*Tense Voice+F0 indignation*).

The LAX-CREAKY||BOREDOM group is the most effective for signalling low activation/power states. Within this group, the *Lax-creaky Voice* achieves the highest ratings overall, though as mentioned ratings are virtually identical to those obtained for the *Lax-creaky+F0 lowered* stimulus.

The group WHISPERY||FEAR also appears to be quite effective in signalling specific low activation/power states. The VQ+F0 stimulus *Whispery+F0 fear* emerges across the board as the most effective in conveying the affect *scared*.

The two stimulus groups, BREATHY||SADNESS and TENSE||JOY, are here relatively ineffective in signalling affect. (In the latter group *Tense Voice* achieves high ratings but note that this appears twice, being also represented in the TENSE||INDIGNATION group.) These two groups are largely omitted from the further discussion of results.

C. Voice-to-affect mapping across languages

Table II presents an overview of all stimuli which achieved ratings above the ± 1 threshold. Stimuli associated with high-activation states are grouped in the upper part of the table and low activation states in the lower. Bold font identifies those stimuli which yielded the highest rating for a given affect.

Table II. Stimuli that obtained the highest ratings (absolute values shown as subscripts) within each stimulus type for the affects tested. Stimuli yielding the highest ratings across stimulus types for at least one language are shown in bold type.

Affect	VQ	F0 (all with Modal Voice)	VQ+F0
indignant	**Tense *R* 2.18, *E* 1.78, *S* 1.64, *J* 1.53**	F0 indignation *R* 1.68, *E* 1.50, *S* 1.02	Tense+F0 indignation *R* 1.67, *E* 1.27
interested	Tense *R* 1.72, *S* 1.05	**F0 indignation** *J* 2.08, ***R* 1.78, *E* 1.66,** *S* 1.42	**Tense+F0 indignation** ***J* 2.15,** *E* 1.62, *R* 1.53, ***S* 1.51**
formal[a]	**Tense *R* 2.38, *E* 1.99** Modal *S* 1.05	**F0 boredom** *R* 1.58, ***S* 1.25,** *E* 1.14	Tense+F0 joy *R* 1.43
stressed[b]	**Tense *R* 1.79, *E* 1.70,** *S* 1.09	F0 indignation *S* 1.79, *R* 1.68, *E* 1.67	**Tense+F0 indignation *S* 1.87,** *R* 1.57, *E* 1.44
happy	**Tense *R* 1.31**	F0 indignation *R* 1.21	
fearless	**Tense *R* 2.22, *E* 1.72, *S* 1.61, *J* 1.30**	F0 boredom *R* 1.41, *E* 1.04	Tense+F0 joy *R* 1.00
apologetic	**Lax-creaky *R* 1.63, *S* 1.37** Whispery *E* 1.33	F0 fear *J* 1.37	Lax-creaky+F0 lowered *R* 1.57, *S* 1.23 **Whispery+F0 fear *J* 1.82, *E* 1.60**
bored	**Lax-creaky** *J* 1.96, ***R* 1.92, *S* 1.87, *E* 1.46**		**Lax-creaky+F0 lowered** ***J* 2.35, *S* 1.87,** *R* 1.77, *E* 1.35
intimate	Lax-creaky *R* 1.81 Whispery *E* 1.29	F0 fear *S* 1.19 F0 indignation *J* 1.55	**Lax-creaky+F0 lowered *R* 1.87** **Whispery+F0 fear *E* 1.62** **Tense+F0 indignation *J* 1.66, *S* 1.23**
relaxed[c]	Lax-creaky *S* 2.09, *R* 1.53, *E* 1.51		**Lax-creaky+F0 lowered *S* 2.13, *R* 1.64, *E* 1.61**
sad	**Lax-creaky *R* 1.82, *S* 1.74, *E* 1.49, *J* 1.49**		Lax-creaky+F0 lowered *S* 1.69, *R* 1.65, *J* 1.38, *E* 1.34
scared		F0 fear *J* 1.27, *R* 1.25, *S* 1.16, *E* 1.09	**Whispery+F0 fear *E* 1.91, *R* 1.83, *J* 1.67, *S* 1.37**

[a] None of the stimuli signalled *formal* to Japanese.
[b] None of the stimuli signalled *stressed* to Japanese.
[c] None of the stimuli signalled *relaxed* to Japanese.

Overall, Table II shows that the stimuli most effective in signalling affect are indeed those which entail voice quality differences, with or without concomitant f_0 shifts (i.e. the VQ and VQ+F0 series). As expected, *Tense Voice* is particularly associated with high activation states. *Lax-creaky Voice* is effective in signalling low activation affects, as is (unsurprisingly) *Lax-creaky+F0 lowered* and the *Whispery+F0 fear* stimuli. Although *Tense+F0 indignation* also achieves high affective ratings, rather unexpectedly, it gets associated with both positive and negative affects. The F0 stimuli were comparatively ineffective on the whole, but the *F0 indignation* stands out as an exception, being strongly associated with affect.

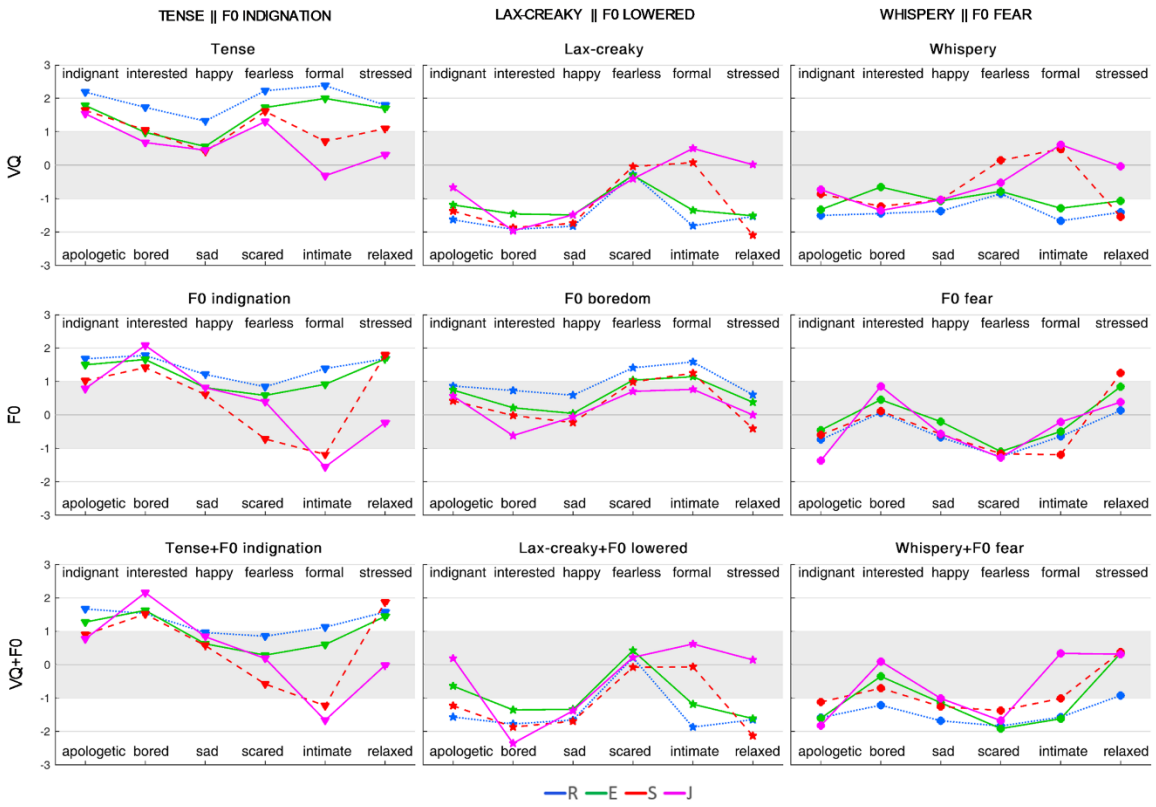


Figure 4 (color online). Ratings for the three stimulus types, from the three most effective stimulus groups in each language.

Figures 4 and 5 allow a closer look at how the individual stimuli map to affect. In Figure 4 results are shown for the subset of stimuli that feature in the three most effective Stimulus Groups of Figure 3, discussed above (TENSE||INDIGNATION, LAX-CREAKY||BOREDOM, WHISPERY||FEAR). Here, for each stimulus, ratings for each affective subtest are shown for the four languages. The spidergram summary plots in Figure 5 show, for the most highly rated stimuli, the network of associated affects in each language. Enlarged data points identify where the highest rating for a particular affect has been obtained (across all stimuli). The affective states have been loosely arranged so that high activation states are in the upper part of the spidergram, and low activation states in the lower. Positive affects are located to the right, and negative to the left.

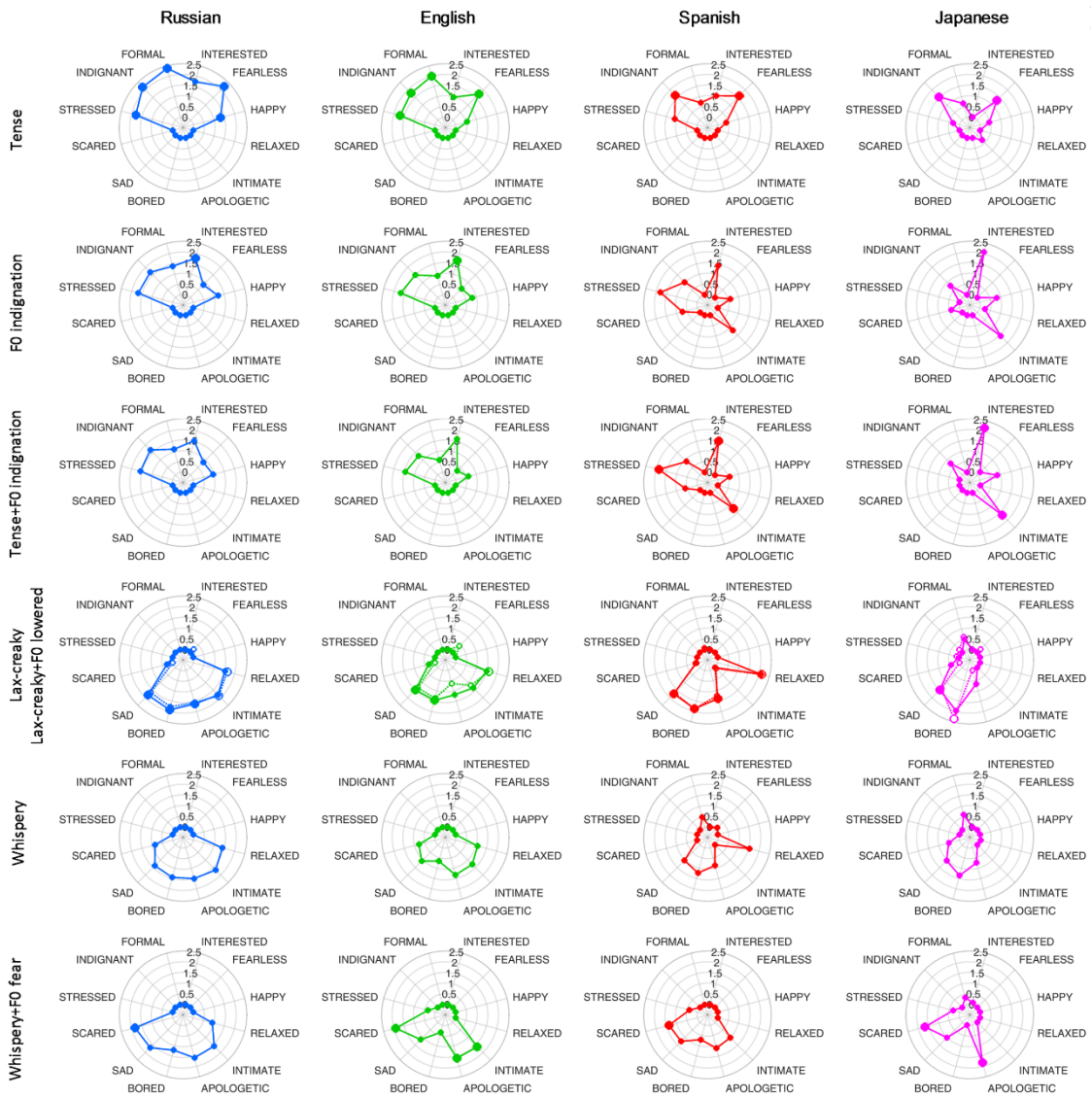


Figure 5 (color online). Summary plots of voice-to-affect associations for the most highly rated stimuli. The data for *Lax-creaky* (solid line) and *Lax-creaky+F0 lowered* (dotted line) stimuli are superimposed. Larger data points show cases when a stimulus yielded highest ratings compared to other stimuli for a particular affect.

1. The VQ stimuli

***Tense voice*:** *Tense Voice* did emerge as the stimulus most frequently associated with high activation states. One high activation states *happy* stands out as being only moderately signalled – and only for a single language, Russian. No other stimulus cues

happy for any language in this study. The elusive nature of vocal correlates of *happy* is reported in many studies.

Although results for *Tense Voice* conform to our initial expectations based on the 'effort code' and to our earlier findings for English, *Tense Voice* does not appear to carry the same affective load across these languages. In all four languages it elicits the highest ratings for *indignant* and *fearless.* Other than that, there are differences: as can be seen in Figures 4 and 5, it is strongly associated with a wide range of high activation states for Russian, somewhat less strongly in English, while for Japanese and Spanish subjects, the affect ratings are much lower, with many more 'gaps' in coverage, e.g., in the cueing of *stressed* and *formal.* (Note that for these two languages *formal* is not strongly signalled by any of the present stimuli, and for J, *stressed* in not rated at all). To sum up, although *Tense Voice* is associated with high activation states in all these languages, the range (number) of affects signalled and the strength of signalling varies considerably – in the order R > E > S > J.

***Lax-creaky Voice*:** (Given the similarity of their ratings, the values for *Lax-creaky Voice and Lax-creaky+F0 lowered* have been superimposed in Figure 5.) From the outset, the expectation was to find this quality strongly associated with low activation states. As with *Tense Voice,* the range/number of affects associated varied considerably across the four languages. Again, the most wide-ranging effects are attested for Russian and the least for Japanese.

This quality yields the highest ratings across the board for both *bored* and *sad*. There are also, however, large cross language differences. *Intimate* is rather well signalled by

this quality for Russian and English, but not at all in Japanese or Spanish. The affects *relaxed* and *apologetic* are also associated with this quality for Russian, Spanish, and (more weakly) English, but not at all for Japanese.

Whispery Voice: As can be seen in Figures 4 and 5, this quality does achieve reasonably high affective ratings for many low-activation states, and its affective profile (Figure 5) is somewhat similar to *Lax-creaky Voice* but more attenuated. The cross-language differences mirror those for *Lax-creaky Voice*, especially for the affects *intimate* and *relaxed*. The initial expectation that whispery voice might be associated with fear is not borne out: while R and E responses for *scared* show weak signalling of this affect (not indicated for the otherwise similar *Lax-creaky Voice*), ratings do not exceed ± 1 threshold.

Breathy Voice: Overall, this stimulus did not yield strong affective ratings (see Figure 3). Although this quality is frequently mentioned as associated with intimacy (e.g., Laver, 1980) and with sadness (e.g., Juslin and Laukka, 2003) but this did not manifest in a clear way. The suggested association with formality in Japanese (Campbell and Mokhtari, 2003) was not borne out here either.

2. The F0 stimuli

Overall, stimuli involving only f_0 manipulation were relatively ineffective in affect cueing (see Figure 3), and so these stimuli were not associated with the affects used in their labels. This was largely expected, given past studies of f_0 contours in affect perception (see Section II.C). In the Introduction, a possible expectation was discussed, based on the 'effort code', i.e. that the strength of activation would mirror the extent to

which a particular f_0 contour deviated from the neutral f_0. Extremely elevated f_0 contours might be expected to be associated with high activation states, and lowered f_0 contours with low activation states. (The very wide range of f_0 levels and the extreme nature of some of them can be seen Figure 2). The fact that most extreme f_0 contour in the series, *F0 indignation*, did yield strong affective responses could be seen as lending support to this proposition. However, the picture is rather more complex, as is discussed further below.

F0 indignation: This most extreme f_0 contour stands out as yielding high affective ratings. Across all four languages, this stimulus was the most highly rated for *interested*. Whereas one might have expected this contour to signal *indignant*, such an association only emerged for R and E (though with ratings that were lower than for *Tense Voice*). Overall, the ratings for R and E are rather similar, and show a broad association with high activation states (similar to *Tense Voice* but with much weaker rating strength).

In Japanese and Spanish this stimulus was rather unexpectedly associated with *intimate*, an association entirely absent in ratings for Russian and English. Thus, the expectation that extreme f_0 deviation from the neutral would be associated with high activation does not hold across these languages. Though it is true for R and E, the fact that both high and low activation states are signalled in both J and S suggests that the 'effort code' explanation is not generalizable. In addition, as can be gleaned in Figure 3, for the other F0 stimuli, there was little evidence of a correlation between the degree of f_0 deviation from the neutral, and degree of activation in responses.

3. The combined VQ+F0 stimuli

The combined stimuli were included to explore whether the combination of voice quality with differing f_0 contours might yield stronger affective responses than when voice quality alone differs (or f_0 alone). If, as many researchers have concluded, f_0 contours alone are ineffective in signalling affect because the voice quality dimension is missing, might not the combination of voice quality with f_0 contours provide a synergy to yield the most effective signalling? The present results did reveal synergies but rather less than might be expected. Affective ratings for three combined stimuli did achieve higher ratings than the VQ or F0 series alone: however, the gain was marginal in two of these three cases.

Whispery Voice+F0 fear: This proved to be the most clear-cut case of synergy, where the combination of voice quality and f_0 features was more effective than either dimension on its own. The *F0 fear* contour yielded relatively weak affective ratings overall but slightly above threshold values for *scared* (see Figure 4). Although *Whispery Voice* achieves stronger affect signalling on the whole, it is still much less effective than *Lax-creaky Voice,* for all affects other than *scared,* which is only weakly indicated. However, when *Whispery Voice* is combined with the *F0 fear* contour there is a distinct enhancement of affective signalling (see Figure 5). This combined stimulus yields the highest ratings for *scared* across all four languages; it yields the highest ratings for *apologetic* in E and J and the strongest signalling of *intimate* for E.

Tense Voice+F0 indignant: As *Tense Voice* is associated with high-activation states, ratings were expected to be higher for the combined stimulus. It was also thought likely that the affect *indignant* would emerge strongly.

These expectations were not borne out, and trends differed according to the language. In Spanish and Japanese, those affects which were associated with *F0 indignant* (*interested, intimate*, and in the case of Spanish, *stressed*) did yield higher ratings when in combination with *Tense Voice,* showing a synergy, even if the extent of the enhancement involved is not great. However, results for Russian and English run contrary to this in that the combined stimulus yields lower affect ratings than either of the VQ or F0 stimuli alone.

Responses for Spanish and Japanese are overall very like those obtained for the *F0 indignation* stimulus, and very different from those obtained for *Tense Voice.* This leads us to conclude that the extreme f_0 contour is the main determinant of the affects cued by the combined stimulus for these two languages.

Taking the results for this combined stimulus together with results the *F0 indignation*, we would conclude that the f_0 dimension works differently in the affect signalling of these two language groups, and that it plays a more important role in Japanese and Spanish than in Russian and English.

These results for the combined stimulus also prompt reflections on the role tense phonation plays in the affect signalling of these two language groups. The fact that a tense voice quality is compatible with, and (even slightly) enhances the impression of *intimate* in Spanish and Japanese underscores a basic difference in how this voice quality functions in the affect-system of the two language groups. While

unambiguously associated with high activation states in Russian and English, there is no such necessary linkage in Japanese or Spanish.

Lax-creaky+F0 lowered: although this combined stimulus yielded high affective ratings, these were almost identical to (and never significantly different from) those for *Lax-creaky Voice.* This strongly suggests that, although lax-creaky voice quality is often produced with low f_0, its affective signalling role appears to be due to the voice quality components other than the low pitch. Although f_0 lowering on its own (with modal phonation) was not included in the stimulus set here, we would tentatively conclude from the results from this combined stimulus that f_0 lowering is not greatly implicated in affect signalling.

V. DISCUSSION

This study set out to (1) explore how the different dimensions of the voice (voice quality and f_0) convey affect and provide insight into whether/how they combine in affect signalling, and (2) to explore the cross-language differences in how affect may be associated with voice.

A. How the dimensions of the voice signal affect

On the whole, results suggest that voice quality dominates the signalling of affect. As in earlier experiments by the authors, stimuli involving VQ – either alone or in combination with a specific f_0 contour – account for most of the highest affect ratings found, while most of F0 series of stimuli were relatively ineffective. However, the high affective ratings for *F0 indignation,* especially in J and S, and the clear synergy of f_0 contour and voice quality in the combined stimulus *Whispery+F0 fear* are indicators

that f_0 can be very important factor in conveying affect. The fact that the affective profiles of J and S are rather different from R and E for the stimuli *F0 indignation* and *Tense+F0 indignation* (Figure 5) provides an indication that (i) the relative contribution of either dimension is likely to be variable across languages, and that (ii) the ways they combine are consequently likely to differ considerably.

An 'effort code'? An expectation was expressed in the Introduction that the signalling of activation would mirror degree of underlying laryngeal effort (akin to the 'effort code' proposed to account for quasi-universal trends in intonation by Gussenhoven, 2004). This suggestion received support from earlier studies for English, where *Tense Voice* was clearly associated with high activation states, while *Lax-creaky Voice* was associated with low activation states. It was also suggested that the 'effort code' might equally be proposed for the F0 stimuli, where those contours deviating most extremely from the neutral contour would be expected to yield relatively stronger signalling of high activation states.

At first glance, this trend does emerge in the results for the voice quality materials. *Tense Voice* does signal high activation, while *Lax-creaky Voice* is associated with low activation. However, a close look at cross language differences demands a more nuanced account. *Tense Voice* is clearly associated with high activation states for R and E, but the effect is more limited for J and S. Furthermore, the fact that a tense voice quality, when coupled with *F0 indignation* can in the latter languages signal *intimate* (enhancing the affect ratings vis à vis the f_0 contour alone) means that we cannot expect a linkage to high activation to be universal.

In a similar vein, the results for the F0 stimuli show that the 'effort code' doesn't hold either in any simple way. Although the extreme f_0 contour *F0 indignation* does achieve high affective ratings, these can simultaneously involve both high (*interested*, *stressed*) and low activation (*intimate*) states. Also, the other very elevated f_0 contours of the F0 series were not more associated with high activation states than were those with lower f_0 contours.

Synergies? The initial proposal was that the combined stimuli would yield affect ratings beyond what the individual VQ or F0 stimuli achieved. This was not resoundingly demonstrated in this experiment, as a clear-cut enhancement was only found for one of the five combined stimuli. Nonetheless, the significant increase in affect signalling in this one case, *Whispery Voice+F0 fear*, provided a clear demonstration that synergies may be needed to signal certain states. As the pairings in this experiment represent a limited set, the present results do not allow of strong conclusions. This is an area that will need to be explored further.

One-to-one mappings? The present results, rather like our earlier experiments indicate that there is no one-to-one mapping of voice to affect: a stimulus such as *Tense Voice* or *Whispery Voice* maps to more than a single affect and may be associated with a cluster of affects. Nonetheless, Figure 5 shows that the mappings are more restricted in some languages than others, with more restricted in J and S than in R and E. For J and S, in the signalling of *intimate* the very extreme f_0 contour appears to be *the* determining factor, but even here, other affects like *interested* are also signalled.

B. Cross-language differences/similarities

Table III provides a simple listing of those stimuli to yield the highest ratings for each affect in each language, and a box surrounds cells where these corresponded across all four languages. The clear-cut cases where the mapping of stimulus-to-affect corresponds across all four languages are as follows:

Tense Voice → *indignant, fearless*

Lax-creaky Voice → *bored, sad*

Whispery+F0 fear → *scared, apologetic*

F0 indignation and *Tense voice+F0 indignation* → *interested*

The affective 'profiles' of Figure 5 provide a visual overview of cross-language convergence and divergence. Despite the overlap of values, no single stimulus here yields the same, or even similar, patterns across all four languages.

Table III. Summary of the stimulus to affect association for the languages tested (**R**=Russian; **E**=Irish-English; **S**=Spanish; **J**=Japanese). Only the stimuli yielding the highest rating are shown.

	Indign.	Interest.	Happy	Fearless	Formal	Stressed	Apologet	Bored	Sad	Scared	Intimate	Relaxed
Tense	**R E S J**		**R**	**R E S J**	**R E**	**R E**						
F0 indignation		**R E**										
Tense+F0 indignation		**S J**				**S**					**S J**	
Lax-creaky/ Lax-creaky+F0 lowered							**R S**	**R E S J**	**R E S J**		**R**	**R E S**
Modal/F0 boredom					**S**							
Whispery+F0 fear							**E J**			**R E S J**	**E**	
Gaps			**E S J**		**J**	**J**						**J**

Languages appear to differ in a number of ways:

Entirely different mappings of voice-to-affect is probably the most striking case. This is exemplified particularly in the signalling of *intimate.* The strong association with the *F0 indignation* stimulus, found for J and S finds no echo in the results for R and E. Similarly, the strong association with *Lax-creaky Voice* for R and E is not found for J and S.

The strength of affective responses can also be quite different across these languages. For example, *Tense Voice* yields much higher affect ratings in R and E than in S or J. In fact, high activation states were on the whole not well signalled for the J and S listeners. It is nonetheless entirely possible that J and S may communicate high activation affects by using voice and f_0 cues, but that the choice of voice qualities and f_0 contours of this study did not include the critical ones needed.

The range/number of affects associated with a particular stimulus also differs considerably in these languages. This can again be illustrated by responses for *Tense Voice,* which in R and E is associated with a wide range of high activation affects, but with a more restricted set in S and J. This also holds for responses to *Lax-creaky Voice*, which yields wide-ranging low-activation states for R and E, a more limited range for S, and an even more restricted range for J.

Gaps in coverage are a related phenomenon (see lowest row in Table III), which emerged particularly in responses for J, where neither *stressed* nor *relaxed* were signalled at all. This was in striking contrast to the other three languages, where these affects were effectively conveyed. The affect *formal* also failed to emerge for J, and

was only weakly signalled for S. Overall, J emerged as having the sparsest signalling of affect from this range of stimuli. It has been suggested (see discussion in Section II.C) that a given emotion may expressed more clearly and recognised with higher accuracy in some languages than in others. This could be the basis for some of the differences observed here, but other possibilities exist, such as that just mentioned, that the stimulus selection here might not be optimal for a particular language. Furthermore, one must consider that other, linguistic and cultural factors might also be at play, and we return to this issue in the Conclusions.

The relative importance of VQ and F0, and in ways in which these dimensions of the voice combine in affect signalling also appears to be different. As discussed above, the affective impact for J and S of the extreme pitch contour of the *F0 indignation* in the signalling of *intimate,* and the way it combines with *Tense Voice* in these languages to heighten the *intimate* affect lead us to propose f_0 likely plays a different, more important role in J and S than in R and E.

Broadly speaking, two groupings appear to emerge here: as suggested by many of the above observations and by the affective profiles of Figure 5, R and E show similar trends, and these differ in many respects from the trends observed for J and S. This grouping is of course not absolute: in the signalling of *relaxed* and *stressed* Spanish patterns closely with Russian and English rather than with Japanese.

Past studies (Scherer, 2000; Scherer, 2001) have proposed that the perception of affect tends to be more similar, the more closely related the languages (see Section II). If this were to pertain here, one should expect to find rather different groupings for these

languages. R, E and S are all Indo-European languages, even if they belong to different branches and are geographically fairly distant. Therefore, one might expect these three languages to form the main cluster, differing as a group from J. The grouping emerged in this data suggests that cross-language differences are not easily explained in terms of language relatedness.

Clearly, this study presents only a very limited contribution to the vast topic of cross-language differences in voice-to-affect signalling. Future extension of this work could include participants from more diverse languages. Of particular interest would be languages where voice quality is exploited for segmental/lexical contrast (e.g., vowels contrasting modal with tense, breathy, creaky voice etc.). Equally, tonal languages would be interesting to explore, where lexical contrasts involve f_0 and, sometimes also voice quality.

VI. CONCLUSIONS

Despite many points of convergence in affect attribution, there are considerable cross-language differences in the affective profile associated with every one of the most affect-carrying stimuli in this study (Figure 5). Results suggest that the two dimensions of the voice, voice quality and f_0, while they may act synergistically some of the time, are not necessarily coupled in easy-to-predict ways. Thus, for example, the initial expectation that *Tense Voice* would naturally combine with raised pitch to heighten the high-activation affect was not borne out in a clear-cut way: languages differed in (i) the extent to which this combination signalled high activation, and (ii) whether the combined stimulus enhanced affective ratings.

All in all, these perception results serve as a reminder that voice quality and f_0 are, in production terms, separately controllable. So, even though there are tendencies for voice quality and f_0 to covary, the potential to exploit them individually may be an important feature that allows for the richly nuanced expression of affect in speech. The cross-language differences observed provide indicators that these vocal parameters are being exploited in different ways depending on the language. This is highlighted by the finding here that f_0 appears to play a rather different and more important role in affect signalling in J and S than in R and E and that the way they combine to signal affect may differ considerably.

The gaps in affective coverage that emerge in these results, where specific affects were not signalled (or only weakly signalled) might be explained by certain factors. First, it should be borne in mind that the selection of voice qualities used in this experiment was not exhaustive, and that these were intentionally designed to be non-extreme exemplars of particular voice qualities. (The f_0 contours included more extreme exemplars.) It is possible that inclusion of extreme voice qualities might result in some of these gaps being filled, and/or in higher affective ratings in certain cases. Likewise, the inclusion of further voice qualities such as harsh voice and falsetto could be important for the signalling of certain affects in one or other language. Furthermore, the combinations of VQ with f_0 were necessarily limited in this study. The fact that the combined stimuli here yielded only a single very clear-cut instance of synergy (*Whispery+F0 fear*) may partly be due to this limitation. The number of participants from each language group was also admittedly small and a possibility remains that factors other than language/culture may have caused differences among groups. For all these reasons, the

present experiment must be seen as an initial exploration, providing pointers to where future work might be directed for a fuller teasing out of this complex question.

A major difficulty for both production and perception studies is that of ensuring that the same affect is being targeted by different subjects. This difficulty becomes more pronounced in a cross-language study: achieving conceptual/semantic equivalence of the scale anchor terms in translation is crucial for the interpretation of results, and emotion terms in one language do not always map perfectly onto terms in another language (Russell, 1991; Mesquita *et al.*, 1997; Sabini and Silver, 2005; Ogarkova *et al.*, 2009). Although considerable care was taken to mitigate this problem (Section III.B), in any such cross-language study, the possibility must be borne in mind that the affect labels in the different languages do not cover identical semantic fields.

The absence of a *formal* response in J to any of these stimuli was unexpected, as the use of breathy/whispery voice has been mentioned in the past as a specific marker of formality/politeness in this language (Campbell and Mokhtari, 2003; Ito, 2004). However, the fact that J is also known to codify formality in a system of honorifics – *keigo* (‘terms of respect’), e.g., Ofuka *et al.* (2000), Ito (2005), may result in voice quality being of relatively minor importance (or totally unimportant). In other words, if formality is already encoded in the lexicon and grammar, there may be little need to signal it through voice quality. Beyond such kinds of explanations, one must consider a further possibility to do with the broader culture: the rules of affect expression may differ in these languages and it may simply be less acceptable to express particular affects in one culture than another – at least in certain social situations.

The vocal expression of affect is widely viewed as lying beyond the linguistic system of the language. The present data suggest sometimes quite distinct language-related patterns of voice-to-affect association. Other than in the case of extreme emotions where vocal control gives way to involuntary effects, it seems that we can and do use our voice to encode psychological-social information, integral to the intended message. As argued elsewhere (Ní Chasaide and Gobl, 2004a; b), the temporal modulation of the voice source to express affect is an essential part of the prosody of a language, and is the hallmark not only of affective prosody, but also of linguistic prosody. One would expect to find for affective prosody (as with linguistic prosody) that both quasi-universal tendencies and language specific 'rules' (Gussenhoven, 2004, Chapter 5) apply. While children acquire these 'rules' in their L1, this is an aspect that becomes problematic in second language learning, being all the more problematic for being poorly understood and difficult to describe in an explicit way. The misinterpretations and misunderstandings that arise from the 'incorrect' vocal signalling of affect are all the more important as the listener is not aware of the 'error'. Whereas grammatical or segmental errors are likely to be identified as part of a foreign accent, the use of language-inappropriate voice prosody is not and tends to be simply interpreted in terms of the L1 code of the listener, impacting on the quality of communication.

Similar considerations pertain to speech technology. It seems clear that affect expression in speech synthesis needs to be language sensitive as one can readily imagine the misapprehensions that would be occasioned by the use of Japanese 'intimate' voice in an English speech synthesis system. There is an increasing demand for applications involving 'emotionally intelligent dialogue-partners' capable of providing affectively-appropriate speech output. In such applications and in future

technologies such as speech-to-speech translation systems language-sensitivity will be an important consideration.

Ultimately, a proper understanding of the prosody of languages will depend on being able to integrate their linguistic and affective dimensions within a single framework. For this, being able to capture the different prosodic functions of voice source modulation is key. A holistic understanding of prosody integrating the linguistic and affective prosody will be needed for many practical applications, not only the facilitation of the next generation of speech technologies, but also for language teaching and for an understanding of how voice disorders impact on the ability to communicate both linguistic and affective dimensions of the message.

As discussed in the Introduction, empirical research in this area is complex and technically challenging. Accurate voice source production data on affect are very difficult to obtain, especially for continuous speech where non-modal voice qualities are used – which is the kind of data that is relevant here. The elusive and difficult nature of the task makes empirical research more complex but does not diminish its importance. In the present study, as in all approaches, there are in-built limitations in scope, but results suggest that this approach nonetheless offers a useful instrument to explore the field and promises new insights into this aspect of spoken communication.

ACKNOWLEDGMENTS

This work was partly funded by the FP6 Network of Excellence HUMAINE and was further supported by An Roinn Cultúir, Oidhreachta agus Gaeltachta/Department of

1 Culture, Heritage and the Gaeltacht (*ABAIR* project). The authors would like to thank
2 the Editor and three anonymous reviewers for their helpful comments and suggestions.

[1] See supplementary material at [URL will be inserted by AIP] for the translations of the affective labels (Table I).
[2] See supplementary material at [URL will be inserted by AIP] for the data on interrater agreement (Table II).
[3] See supplementary material at [URL will be inserted by AIP] for the results of the ANOVA (Table III).

REFERENCES

Airas, M., and Alku, P. (**2004**). "Emotions in short vowel segments: effects of the glottal flow as reflected by the normalised amplitude quotient," in *Affective Dialogue Systems* (Springer-Verlag Berlin Heidelberg), pp. 13-24.

Alku, P. (**2011**). "Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications," Sādhanā: Academy Proceedings in Engineering Sciences (Indian Academy of Sciences) **36**, 623–650.

Altrov, R. (**2013**). "Aspects of cultural communication in recognising emotions," Trames **17(67/62)**, 159-174.

Altrov, R., and Pajupuu, H. (**2015**). "The influence of language and culture on the understanding of vocal emotions," Journal of Estonian and Finno-Ugric Linguistics **6**, 11-48.

Bachorowski, J.-A., and Owren, M. J. (**2003**). "Sounds of emotions," Ann. N. Y. Acad. Sci. **1000, Emotions Inside Out: 130 years after Darwin's The Expression of Emotions in Man and Animals**, 244-265.

Banse, R., and Scherer, K. R. (**1996**). "Acoustic profiles in vocal emotion expression," J. Pers. Soc. Psychol. **70**, 614-636.

Bänziger, T., Hosoya, G., and Scherer, K. R. (**2015**). "Path models of vocal emotion communication," PLoS One **10**, 1-29.

Bänziger, T., and Scherer, K. R. (**2005**). "The role of intonation in emotional expression," Speech Comm. **46**, 252-267.

Bänziger, T., and Scherer, K. R. (**2007**). "Using actor portrayals to systematically study multimodal emotion expression: the GEMEP corpus," in *Affective Computing*

and Intelligent Interaction (ACII 2007), edited by A. Paiva, and R. W. Picard (Springer-Verlag, Berlin), pp. 476-487.

Boula de Mareüil, P., Célérier, P., and Toen, J. (**2002**). "Generation of emotions by a morphing technique in English, French and Spanish," in *Speech Prosody 2002* (Aix-en-Provence, France).

Brislin, R. W. (**1980**). "Translation and content analysis of oral and written material," in *Handbook of Cross-Cultural Psychology*, edited by H. C. Triandis, and J. W. Berry (Allyn and Bacon, Inc, Boston), pp. 389-444.

Bryant, G. A., and Barrett, H. C. (**2008**). "Vocal emotion recognition across disparate cultures," Journal of Cognition and Culture **8**, 135-148.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., and Aubergé, V. (**2006**). "Emotional prosody - does culture make a difference?," in *Speech Prosody 2006* (Dresden, Germany).

Burkhardt, F., and Sendlmeier, W. F. (**2000**). "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ITRW on Speech and Emotion*, edited by R. Cowie, E. Douglas-Cowie, and M. Schröder (Newcastle, Northern Ireland), pp. 151-156.

Burkhardt, F., and Stegmann, J. (**2009**). "Emotional speech synthesis: applications, history and possible future," in *Elektronische Sprachsignalverarbeitung ESSV 2009* (TUDpress, Dresden, Germany).

Campbell, N. (**2002**). "Recording techniques for capturing natural everyday speech," in *Language Resources and Evaluation Conference ELREC'02* (Las Palmas, Spain), pp. 2029-2032.

Campbell, N., and Mokhtari, P. (**2003**). "Voice quality: the 4th prosodic dimension," in *XVth International Congress of Phonetic Sciences* (Barcelona, Spain), pp. 2417-2420.

Carlson, R., Granström, B., and Nord, L. (**1992**). "Experiments with emotive speech - acted utterances and synthesized replicas," in *2nd International Conference on Spoken Language Processing (ICSLP 92)* (Banff, Alberta, Canada), pp. 671-674.

Cheang, H. S., and Pell, M. D. (**2008**). "The sound of sarcasm," Speech Comm. **50**, 366-381.

Cowie, R. (**2009**). "Perceiving emotion: towards a realistic understanding of the task," Philosophical Transactions of the Royal Society B **364**, 3515-3525.

Cowie, R., and Cornelius, R. R. (**2003**). "Describing the emotional states that are expressed in speech," Speech Comm. **40**, 5-32.

Cummings, K. E., and Clements, M. A. (**1995**). "Analysis of the glottal excitation of emotionally styled and stressed speech," J. Acoust. Soc. Am. **98**, 88-98.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (**2015**). "A review of depression and suicide risk assessment using speech analysis," Speech Comm. **71**, 10-49.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (**2003**). "Emotional speech: towards a new generations of databases," Speech Comm. **40**, 33-60.

Drioli, C., Tisato, G., Cosi, P., and Tesser, F. (**2003**). "Emotions and voice quality: experiments with sinusoidal modelling," in *VOQUAL'03* (Switzerland), pp. 127-132.

Ekman, P. (**1993**). "Facial expression and emotion," Am. Psychol. **48**, 384-392.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K. R., Tomita, M., and Tzavaras, A. (**1987**). "Universals and cultural differences in the judgments of facial expressions of emotion," J. Pers. Soc. Psychol. **53**, 712-717.

Elfenbein, H. A., and Ambady, N. (**2002a**). "Is there an in-group advantage in emotion recognition?," Psychol. Bull. **128**, 243-249.

Elfenbein, H. A., and Ambady, N. (**2002b**). "On the universality and cultural specificity of emotion recognition: a meta-analysis," Psychol. Bull. **128**, 203-235.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (**2016**). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing **7**, 190-202.

Fant, G. (**1995**). "The LF-model revisited: transformations and frequency domain analysis," STL-QPSR **2-3**, 119-156.

Fant, G. (**1997**). "The voice source in connected speech," Speech Comm. **22**, 125-139.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). "A four-parameter model of glottal flow," STL-QPSR **4**, 1-13.

Gobl, C. (**1988**). "Voice source dynamics in connected speech," STL-QPSR **1**, 123-159.

Gobl, C. (**1989**). "A preliminary study of acoustic voice quality correlates," STL-QPSR **30**, 9-22.

Gobl, C., Bennett, E., and Ní Chasaide, A. (**2002**). "Expressive synthesis: how crucial is voice quality?," in *IEEE Workshop on Speech Synthesis* (Santa Monica, California, USA), pp. 1-4.

Gobl, C., and Ní Chasaide, A. (**1992**). "Acoustic characteristics of voice quality," Speech Comm. **11**, 481-490.

Gobl, C., and Ní Chasaide, A. (**1999**). "Techniques for analysing the voice source," in *Coarticulation: Theory, Data and Techniques*, edited by W. J. Hardcastle, and N. Hewlett (Cambridge University Press, Cambridge), pp. 300-321.

Gobl, C., and Ní Chasaide, A. (**2000**). "Testing affective correlates of voice quality through analysis and resynthesis," in *ITRW on Speech and Emotion* (Newcastle, Northern Ireland), pp. 178-183.

Gobl, C., and Ní Chasaide, A. (**2003a**). "Amplitude-based source parameters for measuring voice quality," in *VOQUAL'03* (Geneva, Switzerland), pp. 151-156.

Gobl, C., and Ní Chasaide, A. (**2003b**). "The role of voice quality in communicating emotion, mood and attitude," Speech Comm. **40**, 189-212.

Gobl, C., and Ní Chasaide, A. (**2010**). "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle, J. Laver, and F. E. Gibbon (Blackwell Publishing Ltd, Oxford), pp. 378-423.

Goudbeek, M., and Scherer, K. (**2010**). "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," J. Acoust. Soc. Am. **128**, 1322-1336.

Graham, C. R., Hamblin, A. W., and Feldstein, S. (**2001**). "Recognition of emotion in English by speakers of Japanese, Spanish and English," IRAL - International Review of Applied Linguistics in Language Teaching **39**, 19-37.

Grandjean, D., Bänziger, T., and Scherer, K. R. (**2006**). "Intonation as an interface between language and affect," Prog. Brain Res. **156**, 235-268.

Gussenhoven, K. (**2004**). *The Phonology of Tone and Intonation* (Cambridge University Press, Cambridge), p.355.

Guzman, M., Correa, S., Muñoz, D., and Mayerhoff, R. (**2013**). "Influence on spectral energy distribution of emotionsl expression," J. Voice **27**, 129.e121-129.e110.

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., and Wedin, L. (**1980**). "Perceptual and acoustic correlates of abnormal voice qualities," Acta Otolaryngol. **90**, 441-451.

Hanson, H. M. (**1997**). "Glottal characteristics of female speakers: acoustic correlates," J. Acoust. Soc. Am. **101**, 466-481.

Hanson, H. M., and Chuang, E. S. (**1999**). "Glottal characteristics of male speakers: acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**, 1064-1077.

Heldner, M. (**2003**). "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish," Journal of Phonetics **31**, 39-62.

Iseli, M., Shue, Y.-L., Epstein, M. A., Keating, P., Kreiman, J., and Alwan, A. (**2006**). "Voice source correlates of prosodic features in American English," in *Interspeech 2006 - ICSLP* (Pittsburgh, Pennsylvania, USA), paper 1933-Thu1931A1933O.1931.

Ishi, C. T., Ishiguro, H., and Hagita, N. (**2008**). "The role of breathy/whispery voice qualities in dialogue speech," in *Speech Prosody 2008* (Campinas, Brazil).

Ito, M. (**2004**). "Politeness and voice quality: The alternative method to measure aspiration noise," in *Speech Prosody 2004* (Nara, Japan), pp. 213-216.

Ito, M. (**2005**). "The contribution of voice quality to the expression of politeness: an experimental study" (unpublished doctoral dissertation), (University of Edinburgh, Edinburgh).

Johnstone, T., and Scherer, K. R. (**2000**). "Vocal communication of emotion," in *Handbook of Emotions*, edited by M. Lewis, and J. M. Haviland-Jones (Guilford Press, New York), pp. 220-235.

Juslin, P., Laukka, P., and Bänziger, T. (**2017**). "The mirror ot our soul? Comparisons of spontaneous and posed vocal expression of emotion," Journal of Nonverbal Behavior. Advance online publication.

Juslin, P. N., and Laukka, P. (**2001**). "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," Emotion **1**, 381-412.

Juslin, P. N., and Laukka, P. (**2003**). "Communication of emotions in vocal expression and music performance: different channels, same code?," Psychol. Bull. **5**, 770-814.

Juslin, P. N., and Scherer, K. R. (**2005**). "Vocal expression of affect," in *The New Handbook of Methods in Nonverbal Behavior Research*, edited by J. Harrigan, R. Rosenthal, and K. R. Scherer (Oxford University Press, Oxford), pp. 65-135.

Kappas, A., Hess, U., and Scherer, K. R. (**1991**). "Voice and emotion," in *Fundamentals of Nonverbal Behavior*, edited by R. S. Feldman, and B. Rimé (Cambridge University Press, Cambridge), pp. 200-238.

Keller, E. (**2005**). "The analysis of voice quality in speech processing," in *Nonlinear Speech Modeling and Applications* (Springer Berlin / Heidelberg), pp. 54-73.

Kitayama, S., and Ishii, K. (**2002**). "Word and voice: spontaneous attention to emotional utterances in two languages," Cognition and Emotion **16**, 29-59.

Klasmeyer, G., and Sendlmeier, W. F. (**2000**). "Voice and emotional states," in *Voice Quality Measurement*, edited by R. D. Kent, and M. J. Ball (Singular Publishing Group, San Diego), pp. 339-357.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820-857.

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., and Okubo, Y. (**2013**). "Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners," Front. Psychol. **4**, 105.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (**2014**). "Toward a unified theory of voice production and perception," Loquens **1**, e009.

Kreiman, J., Gerratt, B. R., and Ito, M. (**2007**). "When and why listeners disagree in voice quality assessment tasks," J. Acoust. Soc. Am. **122**, 2354-2364.

Kreiman, J., and Sitdis, D. (**2011**). "Perception of emotion and personality from voice," in *Foundations of Voice Studies: an Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell, United Kingdom), pp. 302-360.

Ladd, D. R., Scherer, K. R., and Silverman, K. (**1986**). "An integrated approach to studying intonation and attitude," in *Intonation in Discourse*, edited by K. Johns-Lewis (Croom Helm, London), pp. 125-138.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation contour type, voice quality, and *F*0 range in signaling speaker affect " J. Acoust. Soc. Am. **78**, 435-444.

Landis, J. R., and Koch, G. G. (**1977**). "The measurement of observer agreement for categorical data," Biometrics **33**, 159-174.

Laukka, P. (**2008**). "Research on vocal expression of emotion: state of the art and future directions," in *Emotions in the Human Voice*, edited by K. Izdebski (Plural Publishing, San Diego, CA), pp. 153-169.

Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., and Althoff, J. (**2016**). "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," J. Pers. Soc. Psychol. **111**, 686-705.

Laukka, P., Juslin, P., and Bresin, R. (**2005**). "A dimensional approach to vocal expression of emotion," Cognition and Emotion **19**, 633-653.

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., and Elenius, K. (**2011**). "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation," Computer Speech & Language **25**, 84-104.

Laukkanen, A.-M., Alku, P., Airas, M., and Waaramaa, T. (**2008**). "The role of voice quality in the expression and perception of emotion," in *Emotions in the Human Voice*, edited by K. Izdebski (Plural Publishing, San Diego, CA), pp. 171-184.

Laukkanen, A.-M., Vilkman, E., Alku, P., and Oksanen, H. (**1996**). "Physical variations related to stress and emotional state: a preliminary study," Journal of Phonetics **24**, 313-335.

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge), p.186.

McCluskey, K. W., and Albas, D. C. (**1981**). "Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults," Int. J. Psychol. **16**, 119-132.

McCluskey, K. W., Albas, D. C., Niemi, R. R., Cuevas, C., and Ferrer, C. A. (**1975**). "Cross-cultural differences in the perception of the emotional content of speech: a study of the development of sensitivity in Canadian and Mexican children," Dev. Psychol. **11**, 551-555.

McGraw, K. O., and Wong, S. P. (**1996**). "Forming inferences about some intraclass correlation coefficients," Psychol. Methods **1**, 30-46.

Mesquita, B., Frijda, N. H., and Scherer, K. R. (**1997**). "Culture and emotion," in *Handbook of Cross-Cultural Psychology: Basic Processes and Human Development*, edited by J. W. Berry, P. R. Dasen, and T. S. Saraswathi (Longwood Professional Books).

Mesquita, B., and Walker, R. (**2003**). "Cultural differences in emotions: a context for interpreting emotional experiences," Behav. Res. Ther. **41**, 777-793.

Mozziconacci, S. (**1995**). "Pitch variations and emotions in speech," in *XIIIth International Congress of Phonetic Sciences* (Stockholm), pp. 178-181.

Mozziconacci, S. (**1998**). *Speech Variability and Emotion: Production and Perception. Doctoral Thesis* (Technische Universiteit Eindhoven, Eindhoven), p. 210.

Mozziconacci, S. (**2002**). "Prosody and emotions," in *Speech Prosody 2002* (Aix-en-Provence, France).

Mozziconacci, S. J. L., and Hermes, D. (**1999**). "Role of intonation patterns in conveying emotion in speech," in *The XIVth International Congress of Phonetic Sciences* (San Francisco, USA), pp. 2001-2004.

Murphy, P. J., and Laukkanen, A.-M. (**2009**). "Electroglottogram analysis of emotionally styled phonation," in *Multimodal Signals: Cognitive and Algorithmic Issues* (Springer Berlin/Heidelberg), pp. 264-270.

Murray, I. R., and Arnott, J. L. (**1993**). "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," J. Acoust. Soc. Am. **93**, 1907–1108.

Ní Chasaide, A., and Gobl, C. (**1993**). "Contextual variation of the vowel voice source as a function of adjacent consonants," Lang. Speech **36**, 303-330.

Ní Chasaide, A., and Gobl, C. (**1995**). "Towards acoustic profiles of phonatory qualities," in *XIIIth International Congress of Phonetic Sciences* (Stockholm), pp. 6-13.

Ní Chasaide, A., and Gobl, C. (**2004a**). "Decomposing linguistic and affective components of phonatory quality," in *Interspeech 2004* (Jeju Island, Korea), pp. 901-904.

Ní Chasaide, A., and Gobl, C. (**2004b**). "Voice quality and f_0 in prosody: towards a holistic account," in *Speech Prosody 2004* (Nara, Japan), pp. 189-196.

Ní Chasaide, A., Yanushevskaya, I., Kane, J., and Gobl, C. (**2013**). "The Voice Prominence Hypothesis: the interplay of F0 and voice source features in accentuation," in *Interspeech 2013* (Lyon, France), pp. 3527-3531.

Oatley, K., Keltner, D., and Jenkins, J. M. (**2006**). *Understanding Emotions* (Blackwell Publishing Ltd, Oxford), p.536.

Ofuka, E., McKeown, J. D., Waterman, M. G., and Roach, P. J. (**2000**). "Prosodic cues for rated politeness in Japanese speech," Speech Comm. **32**, 199-217.

Ogarkova, A., Borgeaud, P., and Scherer, K. R. (**2009**). "Language and culture in emotion research: a multidisciplinary perspective," Social Science Information **48**, 339-357.

Paeschke, A. (**2004**). "Global trend of fundamental frequency in emotional speech," in *Speech Prosody 2004* (Nara, Japan).

Paeschke, A., Kienast, M., and Sendlmeier, W. F. (**1999**). "F0-contours in emotional speech," in *XIVth International Congress of Phonetic Sciences* (San Francisco, USA).

Pakosz, M. (**1983**). "Attitudinal judgments in intonation: some evidence for a theory," J. Psycholinguist. Res. **12**, 311-326.

Patel, S., Scherer, K. R., Björkner, E., and Sundberg, J. (**2011**). "Mapping emotions into acoustic space: The role of voice production," Biol. Psychol. **87**, 93-98.

Pell, M. D., Monetta, L., Paulmann, S., and Kotz, S. A. (**2009a**). "Recognizing emotions in a foreign language," Journal of Nonverbal Behaviour **33**, 107-120.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., and Kotz, S. A. (**2009b**). "Factors in the recognition of vocally expressed emotions: a comparison of four languages," Journal of Phonetics **37**, 417-435.

Pell, M. D., and Scorup, V. (**2008**). "Implicit processing of emotional prosody in a foreign versus native language," Speech Comm. **50**, 519-530.

Pittam, J., Gallois, C., and Callan, V. (**1990**). "The long-term spectrum and perceived emotion," Speech Comm. **9**, 177-187.

Russell, J. A. (**1991**). "Culture and the categorisation of emotions," Psychol. Bull. **110**, 426-450.

Russell, J. A., Bachorowski, J.-A., and Fernández-Dols, J.-M. (**2003**). "Facial and vocal expression of emotion," Annu. Rev. Psychol. **54**, 329-349.

Ryan, C., Ní Chasaide, A., and Gobl, C. (**2003**). "Voice quality variation and the perception of affect: continuous or categorical?," in *XVth International Congress of Phonetic Sciences* (Barcelona, Spain), pp. 2409-2412.

Sabini, J., and Silver, M. (**2005**). "Why emotion names and experiences do not neatly pair," Psychol. Inq. **16**, 1-10.

Sadanobu, T. (**2004**). "A natural history of Japanese pressed voice," Journal of the Phonetic Society of Japan **8**, 29-44.

Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. (**2010a**). "Perceptual cues in nonverbal vocal expressions of emotion," Quaterly Journal of Experimental Psychology **63**, 2251-2272.

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (**2010b**). "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," Proc. Natl. Acad. Sci. U. S. A. **107**, 2408-2412.

Scherer, K. R. (**1986**). "Vocal affect expression: a review and a model for future research," Psychol. Bull. **99**, 143-165.

Scherer, K. R. (**2000**). "A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology," in *6th International Conference on Spoken Language Processing* (Beijing, China), pp. 379-382.

Scherer, K. R. (**2003**). "Vocal communication of emotion: a review of research paradigms," Speech Comm. **40**, 227-256.

Scherer, K. R. (**2013**). "Vocal markers of emotion: Comparing induction and acting elicitation," Computer Speech & Language **27**, 40-58.

Scherer, K. R., Banse, R., and Wallbott, H. G. (**2001**). "Emotion inferences from vocal expression correlate across languages and cultures," J. Cross Cult. Psychol. **32**, 76-92.

Scherer, K. R., Clark-Polner, E., and Mortillaro, M. (**2011**). "In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion," Int. J. Psychol. **46**, 401-435.

Scherer, K. R., Ladd, D. R., and Silverman, K. E. A. (**1984**). "Vocal cues to speaker affect: testing two models," J. Acoust. Soc. Am. **76**, 1346-1356.

Schröder, M. (**2001**). "Emotional speech synthesis," in *Eurospeech 2001* (Aalborg, Denmark), pp. 561-564.

Shrout, P. E., and Fleiss, J. L. (**1979**). "Intraclass correlations: uses in assessing rater reliability," Psychol. Bull. **86**, 420-428.

Streiner, D. L., and Norman, G. R. (**2008**). *Health Measurement Scales* (Oxford University Press, Oxford), p. 431.

Sundberg, J., and Nordenberg, M. (**2006**). "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," J. Acoust. Soc. Am. **120**, 453-457.

Sundberg, J., Patel, S., Björkner, E., and Scherer, K. R. (**2011**). "Interdependencies among voice source parameters in emotional speech," IEEE Transactions on Affective Computing **2**, 162-174.

van Bezooijen, R., Otto, S. A., and Heenan, T. A. (**1983**). "Recognition of vocal expressions of emotion: a three-nation study to identify universal characteristics," J. Cross Cult. Psychol. **14**, 387-406.

Waaramaa, T. (**2014**). "Perception of emotional nonsense sentences in China, Egypt, Estonia, Finland, Russia, Sweden, and the USA," Logopedics Phoniatrics Vocology, 1-7.

Waaramaa, T., Laukkanen, A. M., Alku, P., and Väyrynen, E. (**2008**). "Monopitched expression of emotions in different vowels," Folia Phoniatrica et Logopaedica **60**, 249-255.

Waaramaa, T., and Leisiö, T. (**2013**). "Perception of emotionally loaded vocal expressions and its connection to responses to music. A cross-cultural investigation: Estonia, Finland, Sweden, Russia, and the USA," Front. Psychol. **4**, 344.

Westermann, R., Spies, K., Stahl, G., and Hesse, F. W. (**1996**). "Relative effectiveness and validity of mood induction procedures: a meta-analysis," Eur. J. Soc. Psychol. **26**, 557-580.

Williams, C. E., and Stevens, K. N. (**1972**). "Emotions and speech: some acoustical correlates," J. Acoust. Soc. Am. **52**, 1238-1250.

Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. (**2009**). "Voice parameter dynamics in portrayed emotions," in *6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA 2009)* (Florence, Italy), pp. 21-24.

Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. (**2013**). "Voice quality in affect cueing: does loudness matter?," Front. Psychol. **4:335**, 1-14.

Yanushevskaya, I., Tooher, M., Gobl, C., and Ní Chasaide, A. (**2007**). "Time- and amplitude-based voice source correlates of emotional portrayals," in *Affective Computing and Intelligent Interaction: Proceedings of the ACII 2007*, edited by A. Paiva, R. Prada, and R. W. Picard (Springer-Verlag, Lisbon, Portugal), pp. 159-170.

Yuan, J., Shen, L., and Chen, F. (**2002**). "The acoustic realisation of anger, fear, joy and sadness in Chinese," in *7th International Conference on Spoken Language Processing* (Denver, Colorado, USA), pp. 2025-2028.

Zinken, J., Knoll, M., and Panksepp, J. (**2008**). "Universality and diversity in the vocalisation of emotions," in *Emotions in the Human Voice*, edited by K. Izdebski (Plural Publishing, San Diego, CA), pp. 185-202.

Table I. The synthesised stimuli used in the cross-language study. All F0 stimuli have modal voice. All VQ stimuli have neutral F0. Additionally, the stimulus *Modal+F0 neutral* is included for baseline comparison. For the stimuli with * see text.

Stimulus Group	**Stimulus Type** (type of manipulation)		
	VQ	F0	VQ+F0
WHISPERY‖FEAR	*Whispery*	*F0 fear*	*Whispery+F0 fear*
BREATHY‖SADNESS	*Breathy*	*F0 sadness*	*Breathy+F0 sadness*
LAX-CREAKY‖BOREDOM	*Lax-creaky**	*F0 boredom*	*Lax-creaky+F0 lowered**
TENSE‖JOY	*Tense*	*F0 joy*	*Tense+F0 joy*
TENSE‖INDIGNATION	*Tense*	*F0 indignation*	*Tense+F0 indignation*
Baseline Stimulus: ***Modal+F0 neutral***			

Table II. Stimuli that obtained the highest ratings (absolute values shown as subscripts) within each stimulus type for the affects tested. Stimuli yielding the highest ratings across stimulus types for at least one language are shown in bold type.

Affect	VQ	F0 (all with Modal Voice)	VQ+F0
indignant	**Tense *R* 2.18, *E* 1.78, *S* 1.64, *J* 1.53**	F0 indignation *R* 1.68, *E* 1.50, *S* 1.02	Tense+F0 indignation *R* 1.67, *E* 1.27
interested	Tense *R* 1.72, *S* 1.05	**F0 indignation**	**Tense+F0 indignation**
		J 2.08, ***R* 1.78, *E* 1.66,** *S* 1.42	***J* 2.15,** *E* 1.62, *R* 1.53, ***S* 1.51**
formal[a]	**Tense *R* 2.38, *E* 1.99**	**F0 boredom** *R* 1.58, ***S* 1.25,** *E* 1.14	Tense+F0 joy *R* 1.43
	Modal *S* 1.05		
stressed[b]	**Tense *R* 1.79, *E* 1.70,** *S* 1.09	F0 indignation *S* 1.79, *R* 1.68, *E* 1.67	**Tense+F0 indignation *S* 1.87,** *R* 1.57, *E* 1.44
happy	**Tense *R* 1.31**	F0 indignation *R* 1.21	
fearless	**Tense *R* 2.22, *E* 1.72, *S* 1.61, *J* 1.30**	F0 boredom *R* 1.41, *E* 1.04	Tense+F0 joy *R* 1.00
apologetic	**Lax-creaky *R* 1.63, *S* 1.37**	F0 fear *J* 1.37	Lax-creaky+F0 lowered *R* 1.57, *S* 1.23
	Whispery *E* 1.33		**Whispery+F0 fear *J* 1.82, *E* 1.60**
bored	**Lax-creaky**		**Lax-creaky+F0 lowered**
	J 1.96, ***R* 1.92, *S* 1.87, *E* 1.46**		***J* 2.35, *S* 1.87,** *R* 1.77, *E* 1.35
intimate	Lax-creaky *R* 1.81		**Lax-creaky+F0 lowered *R* 1.87**
	Whispery *E* 1.29	F0 fear *S* 1.19	**Whispery+F0 fear *E* 1.62**
		F0 indignation *J* 1.55	**Tense+F0 indignation *J* 1.66, *S* 1.23**
relaxed[c]	Lax-creaky *S* 2.09, *R* 1.53, *E* 1.51		**Lax-creaky+F0 lowered *S* 2.13, *R* 1.64, *E* 1.61**
sad	**Lax-creaky *R* 1.82, *S* 1.74, *E* 1.49, *J* 1.49**		Lax-creaky+F0 lowered *S* 1.69, *R* 1.65, *J* 1.38, *E* 1.34
scared		F0 fear *J* 1.27, *R* 1.25, *S* 1.16, *E* 1.09	**Whispery+F0 fear *E* 1.91, *R* 1.83, *J* 1.67, *S* 1.37**

[a] None of the stimuli signalled *formal* to Japanese.
[b] None of the stimuli signalled *stressed* to Japanese.
[c] None of the stimuli signalled *relaxed* to Japanese.

Table III. Summary of the stimulus to affect association for the languages tested (**R**=Russian; **E**=Irish-English; **S**=Spanish; **J**=Japanese). Only the stimuli yielding the highest rating are shown.

	Indign.	Interest.	Happy	Fearless	Formal	Stressed	Apologet	Bored	Sad	Scared	Intimate	Relaxed
Tense	**R E S J**		**R**	**R E S J**	**R E**	**R E**						
F0 indignation		**R E**										
Tense+F0 indignation		**S J**				**S**					**S J**	
Lax-creaky/ Lax-creaky+F0 lowered							**R S**	**R E S J**	**R E S J**		**R**	**R E S**
Modal/F0 boredom					**S**							
Whispery+F0 fear							**E J**			**R E S J**	**E**	
Gaps			**E S J**		**J**	**J**						**J**

The list of figure captions

Figure 1 (color online). Parameter variation in the synthesised voice quality stimuli (after Gobl & Ní Chasaide, 2003).

Figure 2 (color online). f_0 contours used in the F0 stimuli. The lowest contour shows the f_0 contour used with the lax-creaky voice quality (see text).

Figure 3 (color online). Affective ratings obtained in the six subtests: each row represents a listening test (e.g., *indignant-apologetic*), each column represents a particular stimulus group (e.g., WHISPERY||FEAR). Languages: Irish-English (E); Russian (R), Spanish (S), Japanese (J). Statistically significant cross-language differences: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The most highly rated stimulus for a particular affect is marked by a surrounding box.

Figure 4 (color online). Ratings for the three stimulus types, from the three most effective stimulus groups in each language.

Figure 5 (color online). Summary plots of voice-to-affect associations for the most highly rated stimuli. The data for *Lax-creaky* (solid line) and *Lax-creaky+F0 lowered* (dotted line) stimuli are superimposed. Larger data points show cases when a stimulus yielded highest ratings compared to other stimuli for a particular affect.