INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Voice-to-Affect Mapping: Inferences on Language Voice Baseline Settings

Ailbhe Ní Chasaide, Irena Yanushevskaya, Christer Gobl

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences
Trinity College Dublin, Ireland
anichsid@tcd.ie, yanushei@tcd.ie, cegobl@tcd.ie

Abstract

Modulations of the voice convey affect, and the precise mapping of voice-to-affect may vary for different languages. However, affect-related modulations occur relative to the baseline affect-neutral voice, which tends to differ from language to language. Little is known about the characteristic long-term voice settings for different languages, and how they influence the use of voice quality to signal affect. In this paper, data from a voice-to-affect perception test involving Russian, English, Spanish and Japanese subjects is reexamined to glean insights concerning likely baseline settings in these languages. The test used synthetic stimuli with different voice qualities (modelled on a male voice), with or without extreme f_0 contours as might be associated with affect. Cross-language differences in affect ratings for modal and tense voice suggest that the baseline in Spanish and Japanese is inherently tenser than in Russian and English, and that as a corollary, tense voice serves as a more potent cue to high-activation affects in the latter languages. A relatively tenser baseline in Japanese and Spanish is further suggested by the fact that tense voice can be associated with *intimate*, a low activation state, just as readily as with the high-activation state *interested.*

Index Terms: affect, voice source, glottal, voice quality, perception, cross-language, f_0 contour, prosody, paralinguistic

1. Introduction

The affective prosody of the voice, how we convey emotion, mood and interpersonal attitude through tone of voice is an area relatively poorly understood. It is fundamental to our understanding of how speech communication works, and has important implications in speech technology, as being able to incorporate it would greatly extend the usability and range of application of systems involving speech synthesis and recognition [1-3].

Some past studies have explored how shifts to voice quality in synthetic stimuli can evoke different affects for English speaking listeners [1] and how differences in the f_0 contour may combine with voice quality cues in the signaling of affect [4]. Extending this work, a cross-language experiment was conducted where voice quality and the f_0 contours were varied separately and together, and affective ratings elicited for speakers of Russian, English, Spanish and Japanese [5]. Some of the results of this study have been reported in [6, 7], and [8]. Broadly speaking, although there are many points of convergence in how these language cohorts associated voice to affect, some striking cross language differences did emerge.

The voice source carries complex strands of information. It encodes on the one hand the affective and linguistic prosody, which determine many aspects of the linguistic and affective meaning of the message. However, these prosodic modulations occur relative to the speaker's baseline 'affect-neutral' voice, and are interpreted relative to the characteristic long term phonatory setting of the particular language [9]. This language-intrinsic, characteristic baseline voice setting remains an elusive concept. Although intuitively obvious when one listens to a foreign language, and although often commented on by foreign-language teachers, it can be difficult to define, describe or teach in an explicit way. As listeners, we are also apt to interpret it in terms of our own language's affective voice code – something that probably contributes at a subliminal level to racial stereotyping.

Some of the data from the earlier cross-language experiment are revisited here, but with a focus on the potential insights they may yield on the likely neutral voice settings in these languages, and how such differences might influence the mapping of voice to affect in the specific language.

2. Cross-language perception experiment

The aim of the perception test was to explore voice-to-affect mapping for subjects with differing linguistic backgrounds, by presenting a range of voice-varying synthetic stimuli and eliciting whether and to what extent these were associated with different affects.

Three types of stimuli were presented. Reflecting the fact that the primary interest was to establish the affective colorings that different voice qualities can impart, the first series consisted of exemplars of some distinct voice qualities, in all of which the f_0 contour was the same, i.e. an overall falling contour with two high falling accents. As many studies have shown shifts in the level, range and dynamics of f_0 to be associated with affect signaling in production (though not alone very effective in signaling affect in perception [2, 10]), a second set of stimuli was included, all with modal voice quality, but varied in terms of the level, range and dynamics of the f_0 contour, to mirror affect-related shifts described in [11]. A third series combined the voice quality settings of the first series with the f_0 contour settings of the second. The principal aim was to chart the voice-to-affect patterns in each language with a view to gaining insights into the similarities (likely to be universal) and the differences across the languages (see [5, 8]). In the present paper, results are reexamined for a subset of these stimuli.

2.1. Stimulus preparation

The generation of the three series of stimuli is described here, and outlined also in [1, 7, 8]: the first entailing voice quality differences; the second entailing differences only to level, range and dynamics of the f_0 contour (the tune, i.e. contour shape, remained the same). In the third series, the specific voice qualities of the first set were paired with particular f_0 contours of the second series.

http://dx.doi.org/10.21437/Interspeech.2017-1181

Voice Quality Stimuli (VQ): These were based on the set used in [1], with only minor modifications. The starting point was the *Modal Voice* stimulus from that experiment. It was based on a detailed source-filter decomposition of the Swedish utterance [ˈjɑː aˈjøː], elicited in an affect-neutral context from a male speaker. As none of the subjects spoke Swedish, this was effectively a nonsense utterance for all. The non-modal stimuli involved further voice source manipulations using the KLSYN88a formant synthesizer [12]. The non-modal voice quality stimuli – *Whispery Voice*, *Breathy Voice*, *Lax-creaky Voice* and *Tense Voice* – represented a selection of voice qualities according to the classification system of Laver [9], with one addition, *Lax-creaky Voice*, which is conceptually an extension of the Laver framework. For a detailed description of parameter manipulations, see [1]. The VQ stimuli of interest to this paper are *Tense Voice* and *Modal Voice.* All the VQ stimuli retained the same neutral f_0 contour.

F0 stimuli (F0): These stimuli were all based on the *Modal Voice* stimulus with the original f_0 contour, i.e. the neutral f_0 starting point. Further contours were generated on the basis of quantitative data from a production study in [11] which provided f_0 contours elicited for indignation, anger, joy, fear, boredom, sadness as well as for a neutral affective state. These productions were based on Dutch, but were adopted – not with an expectation that affective correlates would necessarily correspond to the affect-related elicitation terms – but rather, as a set of widely different contours in terms of level, range and dynamics. To generate stimuli with non-neutral f_0 contours, the f_0 values of the (original) neutral f_0 contour were modified by proportional scaling of the values in [11], to generate additional non-neutral contours for indignation, joy, fear, boredom and sadness. As mentioned, the shape of the contour was the same in every case: what varied was its level, range and dynamics. In the present paper, the F0 stimulus of particular interest the one with the indignation f_0 contour referred to as *Modal Voice+IndignationF0*. It was the most extreme in terms of its deviation from neutral.

Combined stimuli (VQ+F0): In this series, the specific voice quality stimuli were combined with one of the above non-neutral f_0 contours. The pairing of voice quality with a particular f_0 contour (deemed likely to co-occur) was guided by the results in [1, 4], as well as by comments in the literature. The only combined stimulus of interest in this paper is that of *Tense Voice+IndignationF0.*

Of the total 15 stimuli used in the test (5 VQ, 5 F0, 5 VQ+F0), results for the following four are examined in detail here:

Modal Voice (VQ stimulus)
Tense Voice (VQ stimulus)
Modal Voice+IndignationF0 (F0 stimulus)
Tense Voice+IndignationF0 (VQ+F0 stimulus)

2.2. The perception test

In the perception test, the 15 stimuli were randomized and presented to listeners 10 times in a series of mini-tests. In each test listeners rated the stimuli in terms of whether and to what degree they were associated with one of a pair of opposite affects. The seven-point rating scale was used with polar opposite affective labels placed on each side of the scale, and it allowed for rating the stimuli as having strong (±3), moderate (±2), mild (±1) or no affective coloring (0). The affect pairs tested were the following: *apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, sad-happy* and *scared-fearless*.

2.3. Participants: language groups

The participants (20-21 per language, gender balanced) were native speakers of Irish-English (E), Russian (R), Spanish (S) and Japanese (J). Instructions were given in their own language and carried out in the respective countries.

2.4. Affect labels

The affective labels were translated from English by at least two native speakers of the respective languages who had a good command of English and who were also familiar with the nature of the research. The translators discussed the translation options and selected the version that was both accurate and best suited to maintain the polarity of affective labels.

3. Results

The broad findings of the test revealed that the stimuli varying in voice quality (with or without non-neutral f_0 contours) were dominant in the signaling of affect. *Tense Voice* was the principal voice quality associated with high-activation/arousal affects. *Lax-creaky Voice* was particularly effective in the signaling of low-activation/arousal affects, while whispery voice was also effective, particularly when combined with the fear f_0 contour. Of the stimuli with non-neutral f_0 contours and modal voice, only one, *Modal Voice+IndignationF0,* yielded strong affective ratings: the others were relatively ineffective and undifferentiated in terms of their affect ratings.

Figure 1 shows results obtained across the languages for the four stimuli of interest here. Each plot in the figure is arranged so that high activation/power states are in the upper half, and low activation/power states in the lower half. The affective states are also arranged so that the more negative affects are located to the left, and the more positive to the right (note that this only works approximately: as there were more negative than positive states, the distribution of the latter is not symmetrical). The innermost 'doughnut' in the spidergram represents 0 affect: the further towards the perimeter the value, the higher the absolute rating for individual affects.

Looking first at *Modal Voice*, one can observe a cross-language difference. Although this stimulus is rated in each language as the one closest to having a neutral affective coloring, it is clear that there is a bias towards high activation, but with differences across these languages. High-activation states are the most strongly associated in Russian and least in Japanese, where it is deemed virtually neutral across all the affects tested. The order in terms of the degree to which high-activation is associated with this stimulus is: R>E>S>J.

Responses to *Tense Voice* generally show a distinct shift towards high-activation, and the cross-language differences mirror to a large extent the differences observed for the *Modal Voice* stimulus. The language profiles differ both in terms of the range of affects associated with this quality and in terms of the strength of those ratings. Although, *Tense Voice* yields the highest ratings obtained across the board for the affects *indignant* and *fearless,* for Russian and English it is more widely and strongly associated with many more high-activation states than for Spanish and Japanese, where there is little or only weak signaling of the states *formal, interested*, and *stressed* (though responses for *stressed* in Spanish differ from the

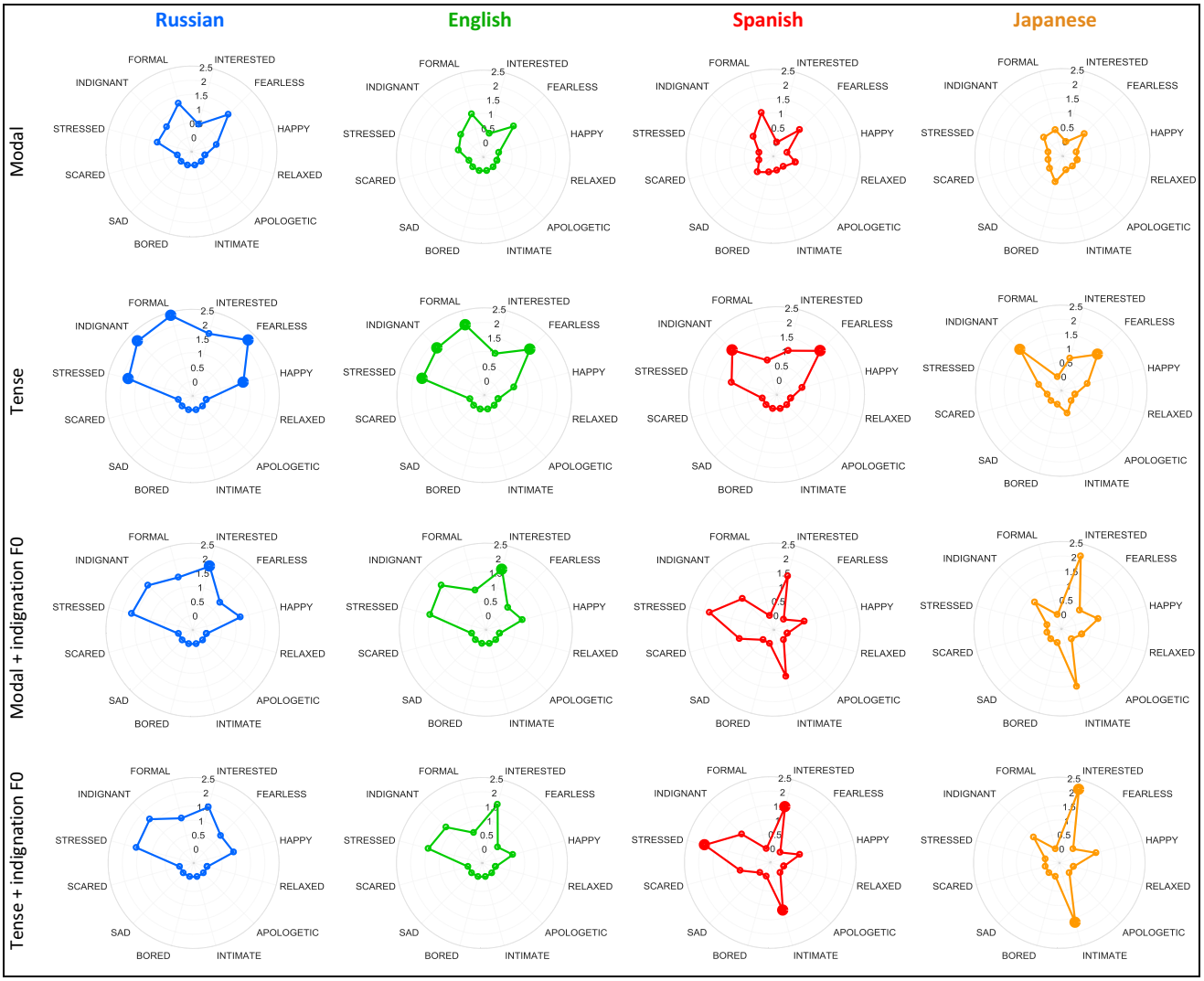

Figure 1. *Affective absolute ratings (0 to 2.5) for the stimuli Modal Voice, Tense Voice, Modal Voice+IndignationF0, Tense Voice+IndignationF0, for Russian, English, Spanish and Japanese. Filled in markers = highest rating across tests for a given affect.*

Japanese). Again, overall, in terms of the strength and range of ratings, the order is R>E>S>J.

As mentioned above, of the F0 stimulus set, i.e. where f_0 was manipulated but a modal voice quality retained, the only one that yielded strong affective ratings was *Modal+IndignationF0*. As can be seen in Figure 1, it yielded the strongest ratings obtained for the affect *interested* across the four languages. Other than that, there is considerable cross-language variation in the range of affective associations it conjures. In Russian and English, it is generally associated with high-activation responses for a wide range of affective states, with a rating strength somewhere between that obtained for the *Modal Voice* and *Tense Voice* stimuli. In Japanese and Spanish, a rather different profile emerged: it yielded fewer instances and relatively weaker signaling of high-activation states – and unexpectedly, a strong association with the low-activation state *intimate*. In Spanish, it is also associated with *stressed* (similar, though more weakly than for Russian and English), an affect not at all associated for Japanese.

In the combined stimulus *Tense Voice+IndignationF0*, the same cross language differences emerged as found for the stimulus *Modal Voice+IndignationF0*. In all the languages, the response pattern is clearly dominated by this extreme f_0 contour. However, the effectiveness of combining a tense phonatory setting with this contour differs for the two language groups. In Russian and English, it *weakens* its affective strength (for all affective ratings) relative to the modal phonatory setting in the *Modal Voice+IndignationF0* stimulus. In contrast, for Spanish and Japanese, this combination *enhances* the affective potency, yielding the highest overall ratings for both the low-activation *intimate* and the high-activation *interested* (as well as *stressed* in Spanish).

4. Discussion

Taken together, these findings can be explained, at least in part, as reflecting differences in the affectively neutral baseline voice settings for these languages. The difference in ratings obtained for the *Modal Voice* stimulus does suggest

that this stimulus is located at a different point relative to the language's affectively neutral voice. The fact that ratings in Russian and English veer towards the high-activation end of the scale suggests that the present *Modal Voice* stimulus is perceived as somewhat tense – and that in these languages, the affect-neutral baseline may be somewhat laxer in quality than the present *Modal Voice*. This is less indicated for Spanish and virtually not at all for Japanese, suggesting that the present *Modal Voice* stimulus comes closer to the affect-neutral setting in these languages. Laver [9] has pointed out that the modal voice quality, as he defined it, may or may not coincide with the habitual setting used by a given language or speaker. It should also be noted that the modal voice defined by Laver [9] entails full glottal closure in the closed phase of the glottal setting, and that research in the intervening years suggests that such full closure may be rather less frequently attested than was then thought.

A difference in the tension setting of the characteristic affect-neutral phonation mode in these languages would also explain the differences emerging in responses to *Tense Voice*. The relatively strong affective responses for *Tense Voice* found for Russian and English – substantially increased rating of all high activation states relative to the modal setting (excepting *happy* for English listeners) makes sense if the affect-neutral setting is rather lax to begin with. If the affect-neutral baseline is relatively lax, then the present *Tense Voice* stimulus would be perceived as deviating more from neutral, and so should generate a stronger perception of high arousal states. In the case of Spanish and Japanese, a more tense affect-neutral setting would militate towards *Tense Voice* being less likely to be perceived as a marker of high arousal – something that tallies with the lower ratings on high-activation states found here.

Cross-language responses to the *Modal+IndignationF0* stimulus are very different, suggesting that the role of f_0 in signaling affect may be different across these languages. The association with intimacy in Japanese and Spanish was a rather unexpected finding (diametrically different from Russian and English). Note that despite the large similarities there were some differences in the Spanish and Japanese responses. The lack of association of this stimulus with *indignant* and *stressed* in Japanese is not found in Spanish which for these affects yields responses more like Russian and English. Thus, although the Spanish and Japanese perceive this stimulus as *intimate*, there is an essential difference, and this stimulus evokes more negative connotations in Spanish.

The cross-language difference in the affective range of the *Tense+IndignationF0* stimulus further supports the idea that tense phonation may be relatively close to the neutral baseline in Spanish and Japanese. Although *Tense Voice* is not in itself associated with intimacy (except minimally in Japanese) which appears rather to be crucially linked to the extreme f_0 dynamics of the indignation f_0 contour, the fact that the tense phonatory setting does not detract from, and indeed enhances the *intimate* coloring (relative to a modal voice setting) suggests that tense voice in itself is relatively affect-neutral. Hence it can associate with both high- and low-activation states (*intimate* and *interested* in the case of Japanese, *intimate, interested, stressed, indignant* in the case of Spanish). *Tense Voice* is never associated with a low activation state in either Russian or English.

These results suggest that the affect-neutral baseline voice setting in Russian and English is more lax than in Spanish and Japanese, and that as a corollary, increasing phonatory tension may be less available for the signaling of high-activation affective states in the latter languages. The differences in sensitivity to and interpretation of the extreme indignation f_0 contour may well be related to this. Pitch and voice quality can be separately controlled, but as they work synergistically in prosodic signaling, they tend to covary. This cross-language perception data highlights their independence and potential for differential exploitation in different languages, suggesting also that the language baseline voice quality may be a determining factor.

5. Conclusions

The results of a cross-language perception test on the mapping of voice-to-affect suggest differences in the baseline affect-neutral voice in these languages. They further suggest that this baseline may influence how the voice quality and pitch dimensions of the voice are exploited in affective prosody.

In view of the fact that the 'rules' of affective prosody can differ across languages, the term *paralinguistic* may not be an optimal term to describe them, suggesting they are *peripheral* to the linguistic system. When we learn a language, we learn the differing levels of the 'speech code', segmental and prosodic. The latter encompasses not simply the *linguistic prosody*, but also the closely related *affective prosody.* Although the latter is relatively little understood, it is a crucial component of language competence to be able to appropriately encode and decode this dimension of the meaning of a spoken language.

This perception-based route to exploring the neutral baseline, would lend itself also to the exploration of within-language divergences, associated with different social groupings. Just as with other meaning-carrying dimensions of spoken language, specific social subgroups evolve differences which mark their identity as a subgroup, based on class, region, education, gender, age, etc. [9, 13-16]. Bearing this in mind, it is worth noting here that the stimuli were of a male voice. Given the well-known differentiation of Japanese male/female speech, it is likely that in that the female voice would require separate investigation. In this context, note also that the English-speaking subjects in this experiment were Irish, a fact that may well influence both the neutral voice setting, relative to other varieties, such as American or British and the voice-to-affect mapping.

These findings have implications for affective speech technologies, whether involving generation or recognition of affect. They underline the risks associated with any assumptions of universality: an intended friendly, intimate voice, appropriate for one language might be interpreted as angry or indignant in another.

6. Acknowledgments

This research was supported by funding from the EU Sixth Framework Network of Excellence HUMAINE and from the Dept. of Arts, Heritage, Regional, Rural and Gaeltacht Affairs (*ABAIR* project).

7. References

[1] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication,* vol. 40, pp. 189-212, 2003.

[2] K. R. Scherer, "Vocal communication of emotion: a review of research paradigms," *Speech Communication,* vol. 40, pp. 227-256, 2003.
[3] C. Gobl and A. Ní Chasaide, "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds., 2 ed Oxford: Blackwell Publishing Ltd, 2010, pp. 378-423.
[4] C. Gobl, E. Bennett, and A. Ní Chasaide, "Expressive synthesis: how crucial is voice quality?," in *IEEE Workshop on Speech Synthesis*, Santa Monica, California, USA, 2002, pp. 1-4.
[5] I. Yanushevskaya, "Vocal correlates of affective states. Unpublished PhD thesis," Trinity College Dublin, Dublin, 2010.
[6] I. Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Voice quality and f0 cues for affect expression: implications for synthesis," in *Interspeech 2005 - Eurospeech*, Lisbon, Portugal, 2005, pp. 1849-1852.
[7] I. Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Mapping voice to affect: Japanese listeners," in *Speech Prosody 2006*, Dresden, Germany, 2006, pp. 1-4.
[8] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Universal and language-specific perception of affect from voice," in *XVIIth International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 2208-2211.
[9] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press, 1980.
[10] K. R. Scherer, "Vocal affect expression: a review and a model for future research," *Psychological Bulletin,* vol. 99, pp. 143-165, 1986.
[11] S. Mozziconacci, "Pitch variations and emotions in speech," in *XIIIth International Congress of Phonetic Sciences*, Stockholm, 1995, pp. 178-181.
[12] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America,* vol. 87, pp. 820-857, 1990.
[13] P. Trudgill, *The Social Differentation of English in Norwich*. Cambridge: Cambridge University Press, 1974.
[14] D. Crystal, *English tone of voice. Essays on intonation, prosody and paralanguage*: Edward Arnold, 1975.
[15] J. Stuart-Smith, "Glasgow: Accent and voice quality," in *Urban Voices: Accent Studies in the British Isles*, P. Foulkes and G. Docherty, Eds., London: Arnold, 1999, pp. 203-222.
[16] J. H. Esling, "Crosslinguistic aspects of voice quality," in *Voice Quality Measurement*, R. D. Kent and M. J. Ball, Eds., San Diego: Singular Publishing Group, 2000, pp. 25-35.