UNIVERSAL AND LANGUAGE-SPECIFIC PERCEPTION OF AFFECT FROM VOICE

Irena Yanushevskaya, Ailbhe Ní Chasaide & Christer Gobl

Phonetics and Speech Laboratory, Centre for Language and Communication Studies,
Trinity College Dublin, Ireland
`yanushei@tcd.ie; anichsid@tcd.ie; cegobl@tcd.ie`

ABSTRACT

This paper outlines the general results of a cross-language study of perception of affect from voice. The study aims to clarify how variations in voice quality (in synthesized stimuli) can evoke different affective coloring for subjects from four different language/cultural backgrounds: Irish English, Russian, Spanish and Japanese. This study furthermore addresses, by including major f_0 differences in some stimuli, some aspects of the role of f_0 in affect cueing, particularly for the signaling of strong emotions. The results suggest both universal and language/culture-specific trends in voice to affect association.

Keywords: voice quality, affect, cross-language variation, perception

1. INTRODUCTION

The voice source and its dynamic variation are integral to the prosodic dimension of spoken communication [3]. Voice source variation is crucial to the two conceptually separable aspects of spoken prosody. Firstly, tone-of-voice is a crucial dimension in the prosodic signaling of the speaker's attitude, emotion or mood (the paralinguistic dimension). Secondly, less well known or appreciated is the role played by dynamic variation of voice source in the more narrowly linguistic signaling functions of prosody (e.g., intonational phrase, stress, accentuation, focus etc.).

Despite significant advances in the study of vocal expression/communication of affect (a review in [8]), still relatively little is known about how the voice source varies in affect signalling. Most data on vocal expression have been collected for methodologically less problematic aspects of the tone-of-voice, such as f_0, intensity and timing. Relatively little empirical data exist concerning what is universal and what is language/culture-specific in the use of the vocal cues in affect signalling.

In general, culture-specific influence on emotion expression/communication is opposed to universal biological factors [7, 9]. On the one hand, vocal expression of biologically based ('hard-wired', 'basic') emotions represents a universal (often involuntary) behaviour related to distinct physiological changes such as muscle tension and sympathetic arousal (fight-or-flight response). These emotions tend to be expressed similarly across different cultures and universally recognised. Thus, the biological component is universal. On the other hand, there is the vocal communication/expression of more complex cognitive emotions (affective states), often for manipulative purposes, when affect communicated is not always affect experienced. Here cultural factors govern the conventions of affect expression. Thus, the cognitive component is culturally informed.

This study aims at throwing light at what is universal and what may be culture-specific in the way speakers with different language/culture background perceive affective colouring from different voice qualities.

2. MATERIALS AND METHOD

Synthesised stimuli The cross-language perception study follows an approach in [1] and [2]. Rather than record and analyse affectively coloured speech, the approach aims at eliciting listeners' affective attributions to a range of different voice qualities. This involves having subjects listen to an utterance synthesised with a variety of voice qualities and having them rate whether and to what extent these stimuli impart affective overtones. The voice quality stimuli were based on prior analyses [3] as well as on the broader literature on the acoustic characteristics of different voice qualities. The main objective of the study is to explore the mapping of voice quality to affect in each of the four language groups (Hiberno-English, Russian, Spanish and Japanese). The listeners' reactions were elicited not to individual dimensions of voice quality, such as spectral slope or breathiness, but rather to the holistic voice quality entity, such as breathy voice or tense voice. A short utterance

['jɑ a'jø] was synthesized in a range of distinct voice qualities. This involved complex manipulations to the stimulus utterance in ways that would render holistic impressions of particular voice qualities according to the framework in [5]. A detailed description of stimuli generation is given in [2]. The synthesized voice quality stimuli: whispery, breathy, lax-creaky, tense and modal were then combined with affect-related f_0 contours described in [6] which varied in the magnitude of f_0 excursions and dynamics (f_0 contours fear, sadness, boredom, joy and indignation) in order to ascertain whether voice quality cues (particularly to strong emotions) become more effective when major f_0 perturbations are included. Furthermore, it was of interest to evaluate the contribution of the same f_0 contours on their own, without voice quality variations. Finally, it was of interest to test whether and to what extent the different language groups might differ in their handling of these different possibilities. Thus, the material for the listening test included three groups of stimuli. The first of these, 'VQ only', included a range of distinct voice qualities. The second, 'f_0 only' group simply involved manipulations to the modal voice stimulus, so as to incorporate a variety of f_0 contours. A further group of stimuli combined specific f_0 contours and voice qualities most likely to co-occur. Thus, for example, whispery voice was combined with f_0 fear, breathy voice with f_0 sadness etc. The stimuli could be further grouped into five 'Affect groups' each of which included three different types of stimuli manipulation, 'VQ only', 'VQ+ f_0' and 'f_0 only'. Thus, 'Affect group fear' included (i) whispery voice, (ii) whispery voice combined with f_0 fear and (iii) f_0 fear combined with modal voice (see Table 1 below).

Table 1: Voice quality stimuli and affect-related f_0 contours used in the cross-language study.

'Affect group'	Voice quality stimuli	Affect–related f_0 contours
fear	whispery	f_0 fear
sadness	breathy	f_0 sadness
boredom	lax-creaky	f_0 boredom
joy	tense	f_0 joy
indignation	tense	f_0 indignation
neutral	modal	neutral

Listening test The listening test was conducted as outlined in [2] as six subtests. In each subtest, the participants were asked to listen to the stimuli in random order and to assess the affective coloring of each stimulus on a 7-point rating scale with the polar opposite affective labels placed on each

side of the scale. The scale allowed to rate the affective coloring of each stimulus as strong (+/-3), moderate (+/-2), mild (+/-1) or no affect (0). The pairs of affective labels were *apologetic-indignant, bored-interested, intimate-formal, sad-happy relaxed-stressed,* and *scared-fearless.*

Participants and translation of affective labels The participants were speakers of Irish-English, Russian, Spanish and Japanese, 20-21 subjects in each language group; the groups were gender balanced. The selection of speakers was motivated by what is known about the use of voice quality and f_0 in each of these languages as well as by impressionistic observations. The instructions to the participants were given in their respective languages, and they were given an opportunity to ask for clarification as regards the test procedure prior to the test. The translation of the labels from English was undertaken 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 and the purpose of the scales. The translators were asked to discuss the translation possibilities and to come to consensus regarding the best possible choice in order to keep the translation accurate and to maintain the polarity of the affective labels.

Statistical analysis A mixed type repeated measures ANOVA was conducted to assess the statistical significance of the differences in voice to affect associations that emerged in the course of the cross-language study. A complex mixed design 4x3x5 was used. The independent variables were Language (4), Stimulus Type (3), and Affect Group (5) while the dependent variable was the affective rating that each stimulus yielded. The interrater agreement was assessed using the single measures Intraclass Correlation Coefficient.

3.   RESULTS AND DISCUSSION

The discussion of results will primarily focus on ratings above +/-1 (above the grey area in Figs. 1-2). This threshold is admittedly arbitrary, but it allows us to focus on more robust and consistent voice to affect associations. Due to space constraints, only a few examples will illustrate to what extent we find agreement among the languages and what distinct language-specific trends emerge.

Languages agree A considerable cross-language agreement was found in the way synthesised stimuli signal a number of affects. There are of course some differences in the strength of affec-

tive ratings demonstrating disparity in the relative potency of the stimulus in cueing the affect for a particular language. This agreement is seen for the affects representing emotions proper (*sad, scared, indignant*), and also attitudes and interpersonal stances (*apologetic, interested, bored, fearless*). The following stimuli were similarly rated by listeners from all the language groups tested, with differences showing up in the strength of the affective rating: whispery + f_0 'fear' signalling *apologetic* and *scared*, lax-creaky and lax-creaky + f_0 'boredom' signalling *bored* and *sad*; tense signalling *indignant* and *fearless*; tense + f_0 'indignation' and modal + f_0 'indignation' signalling *interested;* modal + f_0 'fear' signalling *scared*. This is illustrated in Fig. 1, for the *bored-interested* subtest, for a selection of stimuli.

Figure 1: Languages agree: *bored-interested* (selection). Asterisks show statistically significant differences in stimuli ratings across language groups.
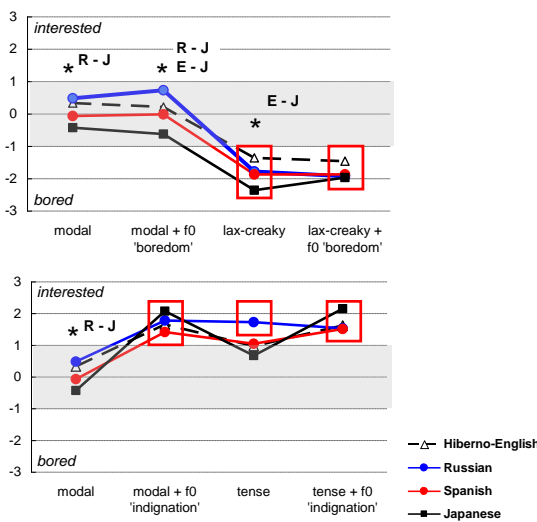
Languages disagree In two subtests, *intimate-formal* and *relaxed-stressed*, the cross-language difference was substantial. The *intimate* was associated for Spanish and Japanese subjects with tense/modal voice combined with f_0 'indignation', a stimulus which was associated rather with *formal* for the Irish-English and Russian subjects. On the other hand, *intimate* was cued by lax-creaky voice and lax-creaky + f_0 'boredom', whispery voice and whispery + f_0 'fear' for Irish-English and Russian, qualities that did not evoke *intimate* for the Spanish and Japanese listeners. There was a conspicuous gap in the cueing of *formal* with any of the stimuli presented here for the Japanese listeners (Fig. 2, panels A-B). There was no affective response from the Japanese in the *stressed-relaxed*

test for any of the stimuli presented here (Fig. 2, panels C-D).

Figure 2: Languages disagree: *intimate-formal* (selection, panels A-B) and *relaxed-stressed* (selection, panels C-D). Asterisks show statistically significant differences in stimuli ratings across language groups.
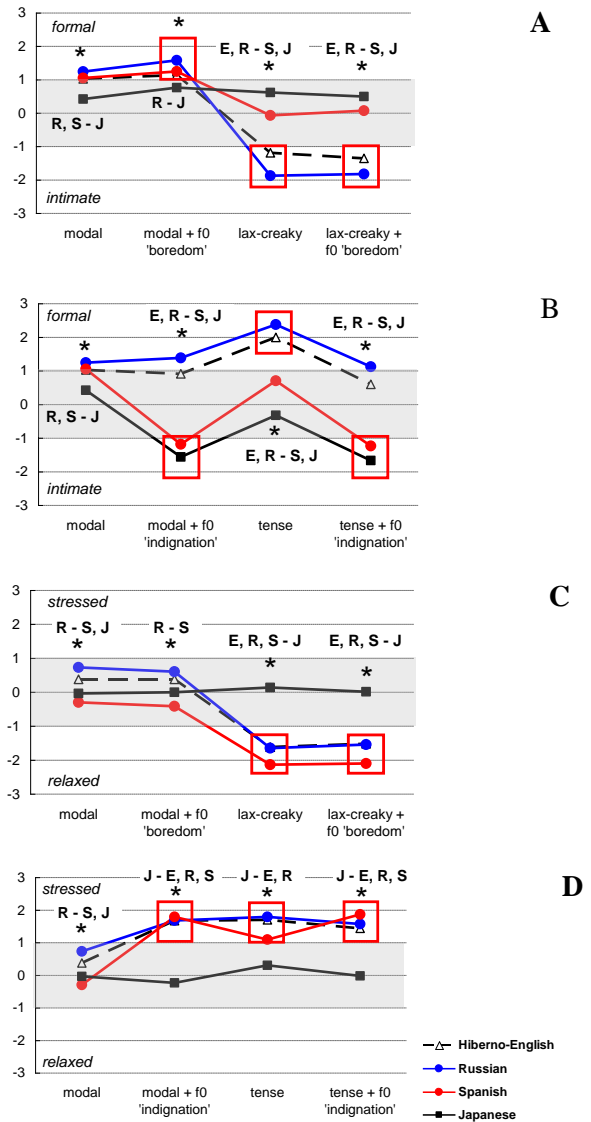
These differences are likely to be linked to underlying cross-cultural differences. The major differences that have shown up here involve mainly attitudes or interpersonal stances (rather than emotions as such): *intimate-formal, relaxed-stressed*, which are more likely to reflect culture-shaped learned behaviour. There may be culture-specific differences in the display rules for certain affects which would affect the extent to which there would be iconic voice-to-affect associations. E.g., it could be that none of the stimuli available signalled *stressed* or *relaxed* for the Japanese because the expressions of stress are frowned upon by this

collectivist culture and are therefore not expressed through vocal cues. The same is true for *formal*. The expression of formality is done in Japanese through the system of honorifics *keigo*, and vocal expression may very well be only of minor relevance for this affect. There is a possibility that the affective labels are not equivalent for the different languages. Intuitively, we did not feel that the semantic differences were likely to explain the very different voice to affect mappings emerging in the present experiment. However, this issue does suggest interesting future research where one could try to elicit how stimuli such as these map to a richer range of affective states. Another factor that could contribute to this kind of cross-language differences would be the neutral voice (baseline) prevalent in these different languages. Impressionistically, the neutral voice of Spanish (and indeed Japanese) is very different from that of English or Russian. If the neutral voice in language A is different from language B and if the neutral voice is similar to a quality which has affective overtones in language B, then cross-language differences have to be expected in the voice to affect association for these languages. This is an interesting area in itself which intersects with how voice is exploited in a particular language to signal affect.

4. CONCLUSIONS

The stimuli presented in this study are necessarily a very limited subset of the infinite variety that exists in real life. Not only have certain voice qualities been omitted (e.g., harsh voice and falsetto), but extreme versions of these qualities have not been included. As for the f_0 contours, although a reasonable sampling of the types of differences in f_0 level and f_0 dynamics often described in the literature were included, these are also a limited selection of the endless possibilities that exist. The same points have to be made about the combined stimuli.

Where, as with intimacy, the affects are strongly signalled, but the different languages appear to use very different qualities in the signalling, we can be fairly certain that we are dealing with a major difference in voice to affect mapping, regardless of the factors that might contribute to this difference. Where, as with the Japanese responses to *stressed/relaxed* and to *formal,* there is a gap in the affective signalling, the conclusions have to be rather different. It is likely that the gaps arise because the stimulus set did not include the qualities

that would be used in this language for the signalling of these affects. One would need eventually to look at production data, e.g., from spontaneous speech corpora [4], which might prompt some more appropriate voice qualities (and f_0 contours) which could be tested within the kind of perception experiment presented here.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] Gobl, C., Bennett, E., N íChasaide, A. 2002. Expressive synthesis: how crucial is voice quality? *Proc. of the IEEE Workshop on Speech Synthesis* Santa Monica, California, USA.

[2] Gobl, C., N íChasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, 189-212.

[3] Gobl, C., N íChasaide, A. 2010. Voice source variation and its communicative functions. In Hardcastle, W.J., Laver, J., Gibbon, F.E. (eds.), *The Handbook of Phonetic Sciences* (2nd ed.). Oxford: Blackwell, 378-423.

[4] Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M. 1998. Acoustic nature and perceptual testing of corpora of emotional speech. *Proc. of the 5th International Conference on Spoken Language Processing* Sydney, Australia.

[5] Laver, J. 1980. *The Phonetic Description of Voice Quality.* Cambridge: Cambridge University Press.

[6] Mozziconacci, S. 1995. Pitch variations and emotions in speech. *Proc. of the 13rd International Congress of Phonetic Sciences* Stockholm, 178-181.

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

[8] Scherer, K.R. 2003. Vocal communication of emotion: a review of research paradigms. *Speech Communication* 40, 227-256.

[9] Zinken, J., Knoll, M., Panksepp, J. 2008. Universality and diversity in the vocalisation of emotions. In Izdebski, K. (ed.), *Emotions in the Human Voice*. San Diego, CA: Plural Publishing, 185-202.