interspeech 2008 incorporating SST 08

Cross-Language Study of Vocal Correlates of Affective States

Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl

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

yanushei@tcd.ie, anichsid@tcd.ie cegobl@tcd.ie

Abstract

This paper is concerned with a cross-cultural study of vocal correlates of affect. Speakers of 4 languages, Irish-English, Russian, Spanish and Japanese, were asked to judge affective content of synthesised stimuli of three types: (1) stimuli varying in voice quality, with a neutral pitch contour; (2) stimuli with affect-related f_0 contours and modal voice; and (3) stimuli in which specific voice qualities and affect-related f_0 contours were combined. Some of the main results are illustrated and point to similarities among the language groups as well as some striking cross-language/culture differences in how these stimuli map to affect.

Index Terms: voice quality, f_0 contour, affect, emotion, perception, cross-language, cross-cultural

1. Introduction

There is a growing interest in how the voice carries affective information, signalling to the listener how the speaker feels and her/his attitude to the interlocutor, the situation, etc. This interest is particularly fuelled by the need to provide speech technologies that are more suited to the needs of the user in particular applications. In principle, these technologies should be competent to handle the complex affective nature of spoken language, where the communicative meaning of utterances depend, not only on their verbal content, but also on the ongoing modulation of tone of voice.

Tone of voice can largely be equated with voice quality, but it is also the case that differences in the dynamics, range and sometimes the specific shape of the pitch contour can play an important part in affect signalling. The role of pitch variation has featured in many studies, e.g. [1-3]. Despite its central importance, relatively little is understood about how voice quality variation may be used to communicate affect. Using synthesis and perceptual testing, the authors have in the past explored how voice quality maps to affect, and additionally, how f_0 cues may combine with voice quality. These experiments were in the first instance carried out on subjects who are speakers of Irish-English (see, for example, [4]).

Clearly, the inter-language and cultural dimension is of crucial importance, not only to our fundamental understanding of this aspect of the communicative process, but also to the possible harnessing of this knowledge in speech technology for languages other than English. For example, if we aspire to multilingual synthesisers which are capable of rendering affective speech, we need to know not only how the voice carries affect, but also the particular voice-to-affect mapping of the language in question. We would expect that many aspects may be universal and that certain affects tend to be signalled by more-or-less the same vocal features, but we also know, if only intuitively, that there are also some important cross-language/cultural differences.

K. R. Scherer comments in [5] that empirical research on intercultural emotion recognition from the voice is extremely rare. The studies reported in [5-9] show that speakers of different languages recognise emotions from vocal cues with better than chance accuracy, but that at the same time there is considerable evidence for culture-specific patterns of emotion recognition. The accuracy of emotion recognition has been shown to depend on the particular emotion, on the experimental design (being lower in studies that used a balanced research design) and on the types of stimuli used (accuracy being higher in studies involving imitation rather than acted or spontaneous emotional expressions) [9]. For example, [8], comparing perception of acted emotional speech by speakers of Spanish and Swedish, showed that sadness is interpreted best cross-linguistically and that Spaniards recognise Swedish anger well, but not the other way around. In general, Spaniards recognise emotions expressed by Swedish speakers more accurately than the other way around. In [5], where emotion perception by speakers from nine different countries was compared, language is named as a major factor in differences in emotion recognition: overall recognition accuracy decreases with the decreasing similarity of languages, with segmental and suprasegmental aspects of language clearly affecting encoding and decoding of emotion.

Burkhardt and colleagues [10] reported the results of another study involving cross-cultural comparison where listeners from France, Germany, Greece and Turkey judged how adequately, in terms of affect expression, synthesised utterances with systematically manipulated prosodic features (pitch, duration, jitter) matched the semantics of the utterances. The authors conclude that, although the effect of the language was not found to be significant, certain differences among the speakers of different countries demonstrate that a cross-cultural global emotion simulation in synthesis will not suffice.

Much more research is needed to establish to what extent the vocal cues to affect are common to particular languages, and to what extent they are not. The present paper outlines research which has been undertaken to explore cross-language/cultural differences and similarities in how tone of voice maps to affect. It involves perception experiments – using synthesised stimuli where both voice quality and f_0 parameters are varied – on subjects from four different language backgrounds: Irish-English, Russian, Spanish and Japanese. The experimental paradigm is essentially the same as in earlier investigations on Irish-English subjects, and both the stimuli and the procedures are described in detail in [4].

Simply put, this research is aiming to provide some initial indicators on how particular voice qualities and pitch contours are associated with affective states in these languages. It aims eventually to provide at least some firm hypotheses on what might be universal and draw attention to where there might be major pitfalls for communication (in human as in synthesised speech generation and perception), due to the

Accepted after peer review of full paper

cross-language/cross-cultural differences in the way in which the voice is exploited to communicate affect.

2. Method

Synthesised stimuli were presented in a series of listening tests to speakers of four different languages, Irish-English, Russian, Spanish and Japanese. As mentioned above, the methodology follows that used in some earlier investigations, and more extended descriptions of the stimuli and the procedure are to be found in [4].

2.1. Stimuli

15 synthesised stimuli of a Swedish utterance "ja adjö" [ˈjɑː aˈjøː] (male voice) were generated using the KLSYN88 formant synthesiser [11]. This utterance was deemed to be semantically neutral, being in a language foreign to all the listener groups. The stimuli were of three types according to the parameters systematically varied in the stimulus generation (see Table 1): the **'VQ only'** stimuli were differentiated in terms of voice quality; the **'f_0 only'** stimuli incorporate affect-related f_0 contours based on the data in Mozziconacci [1]; the **'VQ + f_0'** stimuli are a combination of these affect-related f_0 contours and voice quality appropriate for each of these affects. A detailed description of the synthesised stimuli is given in [4], and the changes introduced to the stimuli for the present experiment are outlined in [12].

'VQ only' stimuli: the synthesised voice qualities include modal voice, breathy voice, whispery voice, lax-creaky voice and tense voice. The modal stimulus is based on a detailed source-filter decomposition of the original utterance, spoken by a male speaker. The non-modal stimuli involved further voice source manipulations aimed at simulating a selection of voice qualities according to the classification system of Laver [13], with one addition, lax-creaky voice, which is conceptually an extension of the Laver framework.

Note that the 'VQ only' series of stimuli do in fact incorporate some f_0 differences. These differences were deemed to be intrinsic aspects of voice quality differentiation, and we decided to include them. They are very minor for the most part: f_0 is marginally higher (5 Hz) for tense voice and marginally lower for breathy voice (5 Hz) compared with modal voice. The one quality where there is a more substantial intrinsic f_0 difference is the lax-creaky quality, where there is a lowering of 30 Hz relative to modal voice.

'f_0 only' stimuli: these incorporated affect-related f_0 contours as described in [1] which provides quantitative data based on Dutch production data. These f_0 contours were in that study found to be associated with indignation, anger, joy, fear, boredom, sadness as well as with a neutral affective state. The affective f_0 contours included some rather substantial deviations from the neutral contour and were adapted to our synthetic stimuli by a proportional scaling of the values in [1]. The neutral f_0 contour was used for the modal voice quality of our 'VQ only' stimuli. By modifying the f_0 values of the modal stimulus, five 'f_0 only' stimuli were generated with non-neutral pitch variations in [1]. Our expectation was that these contours might tend to evoke the affects with which they were associated in the Dutch study, *indignation, joy, fear, boredom* and *sadness*. Given the considerations that have prompted the cross-language study, there is, however, no presumption that these f_0 contours should necessarily be associated with those particular affective states.

'VQ + f_0' stimuli: these involved matching one of the non-modal qualities of the f_0 series to one of the non-neutral f_0 contours above. The choices regarding which voice quality might be matched to a particular f_0 contour was guided by the results of our previous experiments as well as by past findings of other researchers or more general claims in the phonetic literature. As the intonation patterns of the 'f_0 only' series were based on data from a Dutch study, we started with a default hypothesis that these contours might also be cross-linguistically associated with affects similar to those of that study. Therefore, for the combined stimuli, we simply tried to match them with the likeliest voice qualities. It goes without saying that our choices in matching voice quality to f_0 contour reflect a Western European bias. Furthermore, in the possible mapping of the combined (or indeed any) stimuli to particular affects, although we had certain expectations, we were amply aware of the fact that such expectations might not be met. Indeed, the central interest of the experiment was to allow the listeners to 'sort' these stimuli, guiding us to potential cross-language differences and similarities.

Consequently, certain caveats and necessary limitations of this study must be borne in mind when interpreting results. Firstly, the sampling of the potential range of vocal variation (voice quality and f_0 contour) is necessarily limited. For the voice qualities, not all potentially important qualities could be included, and furthermore, extreme exemplars of the individual qualities were avoided. As regards the combined stimuli it is particularly the case that sampling of the potential space is not only sparse, but it may miss important combinations. There may be combinations that would be particularly relevant, say, to a Japanese ear, which are not approximated by any of the stimuli included. Thus, when looking at results particularly for the combined stimuli, it is important to bear in mind that the lack of a clear affective rating may simply indicate that the optimal combination was not represented in the stimulus set. Where a combined stimulus yields a high affective rating, one can safely infer that the particular combination of VQ and pitch contour does yield an affective colouring. When a high rating is not achieved, we cannot infer the opposite, but must simply conclude that we do not yet know whether there is some such combination that is required to imprint affect on the speech output.

Table 1. *Synthesised stimuli.*

VQ only	**f_0 only**	**VQ + f_0**
breathy	modal + f_0 'sadness'	breathy + f_0 'sadness'
whispery	modal + f_0 'fear'	whispery + f_0 'fear'
lax-creaky	modal + f_0 'boredom'	lax-creaky + f_0 'boredom'
tense	modal + f_0 'joy'	tense + f_0 'joy'
modal	modal + f_0 'indignation'	tense + f_0 'indignation'

2.2. Listening tests

The perception test was conducted according to the procedure described in [4] as a series of six subtests with the native speakers of Irish English (n=20), Russian (n=21), Spanish (n=20) and Japanese (n=21). In each sub-test, 10 randomisations of 15 stimuli were presented to the participants, and responses were obtained for a pair of opposite affective attributes (e.g., *sad-happy*). The pairs of affective attributes tested were *sad-happy*, *intimate-formal*, *relaxed-stressed*, *bored-interested*, *apologetic-indignant* and *fearless-*

scared. The affective labels were translated from English into respective languages by native speakers.

For each pair of affective labels, the participants judged each stimulus for the presence and strength of affect, on a seven point scale ranging from -3 to + 3, with 0 corresponding to no perceived affect, and plus or minus 1, 2 or 3 corresponding to mild, moderate and strong presence of an affect respectively. For each stimulus within each subtest, mean ratings were calculated across 10 randomisations for every subject. The results for every stimulus within each subtest were further averaged across all subjects' responses.

A one-way ANOVA with stimulus-type as a factor as well as the Tukey's HSD test were conducted to explore the difference in perception of various stimuli for each subtest. The significance level was set at $p < .05$.

3. Results: some illustrations

Comprehensive coverage of results is not possible here, but we will illustrate some of the main trends which have emerged in the data. Figure 1 illustrates the maximum ratings achieved for four affects: *indignant, bored, intimate* and *formal*. The three horizontal lines show results for the three stimulus groups as shown in Table 1. The initial letters (E, R, S, J) refer to the four language groups tested (Irish-English, Russian, Spanish and Japanese respectively) and these are located to indicate the strength of the rating obtained (i.e., showing the maximum rating obtained for any stimulus within either stimulus group).

Note for the affects *indignant* and *bored* that the trends are the same across the four languages. In the case of *indignant*, the highest rating for the 'f_0 only' stimuli was as expected for f_0 'indignation'. For the 'VQ only' stimuli, tense voice yielded the highest ratings, and for each language group the voice quality signalling was stronger than the f_0 signalling. Curiously, the combination of tense + f_0 'indignation' was less effective than tense voice with the neutral f_0 contour.

For the *bored* affect, the broad trends are also similar across the four language groups. In this affect, the 'f_0 only' stimuli were largely ineffective, whereas for the 'VQ only' stimuli, the lax-creaky voice quality yielded rather strong ratings for boredom.

The maximum ratings for *intimate* and *formal* show that there are also some striking cross-language differences in how voice quality and f_0 signal affect.

In the case of *intimate*, it is clear that f_0 ***or*** voice quality can be potent in evoking this affect, but that the language groups differ in their choices. It is clear that the high level and extensive dynamic range of the 'modal + f_0 'indignation' stimulus was potent in signalling intimacy, especially for the Japanese subjects, but also for the Spanish. None of the 'f_0 only' stimuli yielded high ratings for the Irish-English or Russian subjects.

The opposite is true of the voice quality ('VQ only') stimuli. Here, a lax-creaky voice quality was highly rated for intimacy by the Russian group, with whispery voice also rating highly (this last fact is not inferable from the figure). For the Irish-English subjects, whispery voice yielded the highest ratings, but ratings for lax-creaky were almost as high. For the Japanese and Spanish subjects, none of the 'VQ only' stimuli yielded high ratings. Clearly for these groups, f_0 cues are vital for affect signalling.

This last is evident also from the results for the combined stimuli 'VQ + f_0'. Here the preferred stimuli were those with tense voice and f_0 'indignation'. It is also striking that tense

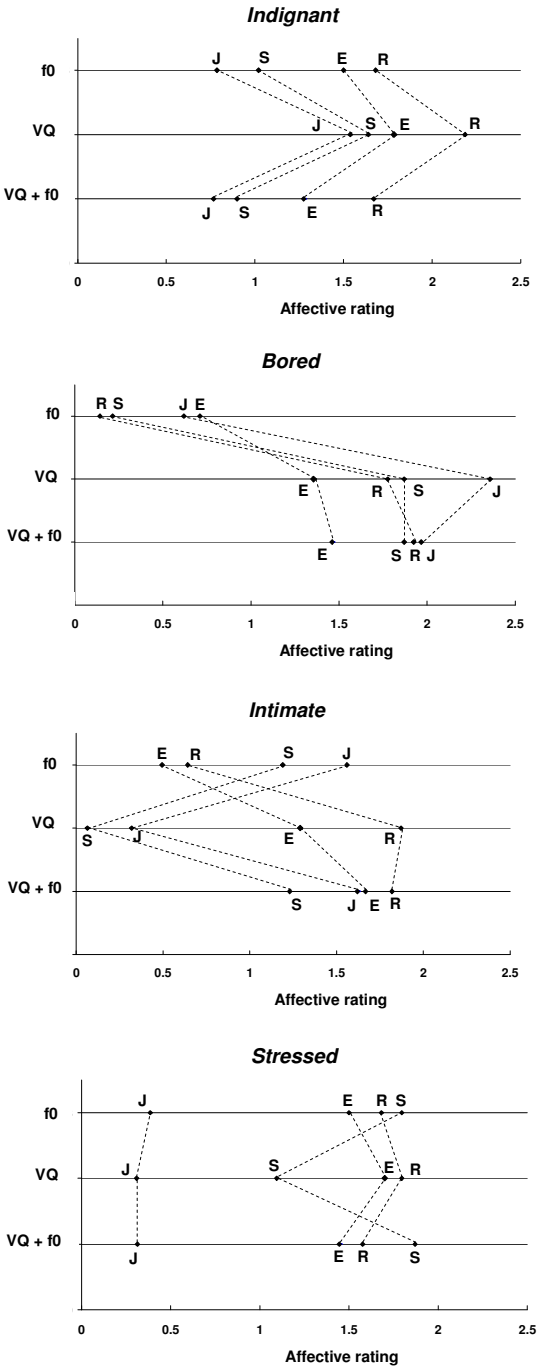

Figure 1. *Affective rating for 3 stimulus types compared across language groups*

voice was preferred over modal voice, as both these qualities were combined in this test with f_0 'indignation'. Given that choice, those stimuli with tense voice were rated more highly, even though the difference in ratings was small.

Not surprisingly, in the combined stimuli 'VQ + f_0' lax-creaky + f_0 'boredom' was the most highly rated stimulus for the Russian subjects, followed by whispery + f_0 'fear'. The Irish-English made the same choices, but in reversed rank order. The finding that lax-creaky and whispery voice signal intimacy for Russian/Irish-English was in keeping with our expectations in this experiment. Note that both these voice qualities have a breathy-whispery quality in common, and that

traditionally in the phonetics literature, breathy or whispery voice have been associated with intimacy. This was also an expectation, given the results of some our earlier experiments [4], which were based on Irish-English subjects.

The finding that tense voice (in combination with the f_0 contour which we expected to be associated with indignation) is for Japanese and Spanish subjects linked to intimacy is a rather surprise finding which runs contrary to our initial expectations. The two language groups differ not only in terms of the relative weight they appear to attach to f_0 vs. voice quality cues to this affect, but also in terms of the particular selection of vocal qualities they favour. Russian and Irish-English subjects were overall rather similar in terms of their responses in that both lax-creaky voice and whispery voice were both potent in signalling intimacy. For the Japanese and Spanish subjects, a high, dynamically varying pitch contour appears to be crucial, and the preferred vocal quality veers towards the tense.

What is striking is the divide between Russian/Irish-English on the one hand, and Japanese/Spanish on the other. These results for intimacy underline two important aspects of cross-language differentiation. First of all, languages may differ in the relative importance they attach to voice quality as compared to f_0 cues. Secondly, languages may differ in how specific voice qualities (and/or intonation contours) map to affect.

In the final panel of Figure 1, results for *stressed* are shown. These also show cross-language differences, with the Japanese group clearly separating out from the Russian, Spanish and Irish-English. In the case of the latter three languages, f_0 'indignation', tense voice, or a combination of tense voice and f_0 'indignation' are associated with the stressed affect. The only prominent difference among these three languages is that the tense voice quality on its own is less effective for the Spanish than for the other two. The striking difference is of course that none of the stimuli presented in this experiment generated high ratings for *stressed* in the case of the Japanese subjects.

4. Conclusions

As pointed out in the introduction, the present experiment is but a first pass at a very big question, and there are necessarily limitations that must be borne in mind when interpreting the results. The selection of voice qualities tested is limited, and does not include extreme deviations from modal voice. Some additional qualities, e.g. falsetto voice, as well as more extreme examples might well be needed to get a more comprehensive coverage. Similarly, the range of f_0 contours is limited. Perhaps more importantly, the particular combinations of voice qualities and f_0 contours chosen for the 'VQ+f_0' series are likely not to have been optimal in certain cases. This fact is brought home when we consider that the particular combinations were guided by the results of earlier studies and by comments in the phonetics literature, which have tended to be biased towards English. The unexpected results discussed above for the affect *intimate* in the case of Japanese and Spanish subjects highlight the fact that our intuitions are necessarily biased by our language background. They also illustrate, however, that the broad approach adopted here of allowing subjects to freely associate affect to a range of stimuli does serve to draw attention to major cross-language differences in how voice and pitch signal affect.

There are clear implications for speech technology, whether directed at the recognition or synthesis of affective speech in different languages. The kinds of differences illustrated here are likely to be crucial if we want to synthesise an intimate voice in a multi-lingual synthesis system. Similarly, if we want to recognise affects in speech, it is clear that affects such as *intimate* or *stressed* will need language specific adaptations. Although the present study tackles only a small portion of the affective signalling landscape, it does at least serve to highlight the potential dangers of any prior assumptions that languages do roughly the same things when it comes to vocally signalling affective states.

5. Acknowledgements

The research was supported by the EU Sixth Framework Network of Excellence HUMAINE.

6. References

[1] Mozziconacci, S., "Pitch variations and emotions in speech", *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, 1995, 178-181.

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

[3] Bänziger, T. and Scherer, K. R., "The role of intonation in emotional expression", *Speech Communication,* 46, 252-267, 2005.

[4] Gobl, C. and Ní Chasaide, A., "The role of voice quality in communicating emotion, mood and attitude", *Speech Communication,* 40 (1-2), 189-212, 2003.

[5] Scherer, K. R., Banse, R. and Wallbott, H. G., "Emotion inferences from vocal expression correlate across languages and cultures", *Journal of Cross-Cultural Psychology,* 32 (1), 76-92, 2001.

[6] McCluskey, K. W. and Albas, D. C., "Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults", *International Journal of Psychology,* 16 (2), 119-132, 1981.

[7] van Bezooijen, R., Otto, S. A. and Heenan, T. A., "Recognition of vocal expressions of emotion: a three-nation study to identify universal characteristics", *Journal of Cross-Cultural Psychology,* 14 (4), 387-406, 1983.

[8] Abelin, Å., "Spanish and Swedish interpretations of Spanish and Swedish emotions – the influence of facial expressions", in *FONETIK 2004, the XVII Swedish Phonetics Conference*, Stockholm University, Sweden, 2004.

[9] Elfenbein, H. A. and Ambady, N., "On the universality and cultural specificity of emotion recognition: a meta-analysis", *Psychological Bulletin,* 128 (2), 203-235, 2002.

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

[11] Klatt, D. H. and Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of the Acoustical Society of America,* 87, 820-857, 1990.

[12] Yanushevskaya, I., Gobl, C. and Ní Chasaide, A., "Voice quality and f_0 cues for affect expression: implications for synthesis", in *Interspeech 2005 – Eurospeech*, Lisbon, Portugal, 2005, 1849-1852.

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