# Mapping Voice to Affect: Japanese listeners

*Irena Yanushevskaya, Christer Gobl & Ailbhe Ní Chasaide*

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

## Abstract

This paper reports the results of perception tests administered to speakers of Japanese as part of a cross-language investigation of how voice quality and $f_0$ combine in the signalling of affect. Three types of synthesised stimuli were presented: (1) 'VQ only' involving variations in voice quality and a neutral $f_0$; (2) '$f_0$ only', with different $f_0$ contours and modal voice; and (3) combined 'VQ + $f_0$' stimuli, where combinations of (1) and (2) were employed. Overall, stimuli involving voice quality variation (1 and 3) proved to be most consistently associated with affect. In series (2) only stimuli with very high $f_0$ yielded high affective ratings. Some striking differences emerge in the ratings obtained for Japanese subjects compared to those obtained for speakers of Hiberno-English [7], suggesting that the generation of expressive speech synthesis will need to be sensitive to language specific uses of the voice.

## 1. Introduction

Voice quality and $f_0$ characteristics are known to be crucial in signalling affect. Except in the case of a speaker in the grip of extreme emotion whose vocal characteristics may not be under voluntary control, we know that speakers exploit tone-of-voice in conventional ways to signal their real or feigned emotion, mood and attitude. This is an aspect of speech communication about which relatively little is understood. One would assume, however, that while some aspects of this signalling may be universal, it is more than likely that some aspects are specific to the language/culture.

Building on earlier work [2,3,7], the experiment reported here is part of a cross-language investigation of the way voice quality and $f_0$ contribute to affect perception. In [7], speakers of Hiberno-English rated the affective colouring of a short utterance, synthesised to provide three series of stimuli, varying in terms of (1) voice quality, (2) $f_0$ contour and (3) specific combinations of the qualities and contours of (1) and (2). The present paper reports on a similar experiment carried out on Japanese subjects.

A 'broad-palette' approach is adopted in this as in the earlier studies, whereby the listeners are confronted with a range of stimuli varying in vocal characteristics, and rate them in terms of a selection of affective attributes. These include not only strong emotions, but also many milder affective states and attitudes. Furthermore, the voice qualities generated were guided by earlier analyses of different voice qualities, and were not based on any specific investigation of affective speech.

The issues explored here are not only important for our understanding of human communication, but also for practical objectives of producing synthesis and recognition systems capable of dealing with this dimension of human interaction. This work contributes to these goals which are being jointly pursued within the EU-funded HUMAINE Network of Excellence.

## 2. Synthesised stimuli

15 synthesised stimuli of a Swedish utterance "ja adjö" ['jɑ: a'jø:] (male voice) generated using the KLSYN88 formant synthesiser [4] were used in the listening tests. The stimuli can be grouped into three types according to the parameters systematically varied in the stimuli generation: 'VQ only', '$f_0$ only', 'VQ + $f_0$' (see Table 1). The 'VQ only' stimuli are differentiated in terms of voice quality; the '$f_0$ only' stimuli incorporate affect-related $f_0$ contours based on the data in Mozziconacci [6]; and the 'VQ + $f_0$' stimuli are a combination of these affect-related $f_0$ contours and the most appropriate voice quality for each of these affects.

Detailed description of the synthesised stimuli is given in [3], and the changes introduced to the stimuli for the present experiment are outlined in [7].

**'VQ only' stimuli.** The synthesised voice qualities include modal voice, breathy voice, whispery voice, lax-creaky voice and tense voice. The modal stimulus was 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 [5], with one addition, lax-creaky voice, which is conceptually an extension of the Laver framework. These stimuli are essentially a subset of those used in [3] where they are described in full.

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 (see Fig. 2).

**'$f_0$ only' stimuli.** Affect-related $f_0$ contours used in the synthesis of the '$f_0$ only' and the 'VQ + $f_0$' stimuli were taken from [6] which provides quantitative data based on Dutch production data, for $f_0$ contours associated with *indignation, anger, joy, fear, boredom, sadness* as well as for a *neutral* affective state. These $f_0$ contours (except the one for *anger* which was not used in the present experiment) were adapted to our synthetic stimuli by a proportional scaling of the values in [6]. By modifying the $f_0$ values of the modal stimulus (neutral $f_0$), five '$f_0$ only' stimuli were generated with non-neutral pitch variations related to indignation, joy, fear, boredom and sadness (see Fig. 1). 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.

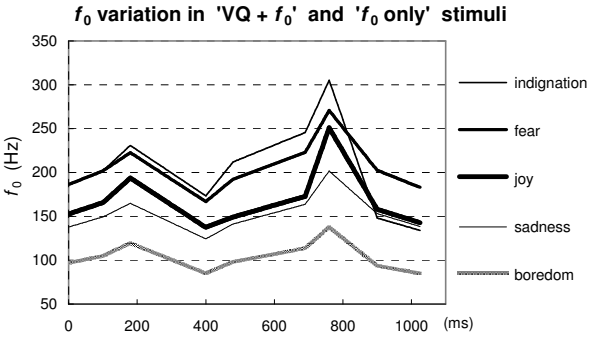


Figure 1: *Variation in $f_0$ for 'VQ + $f_0$' and '$f_0$ only' stimuli (scaled after the $f_0$ values in [6])*

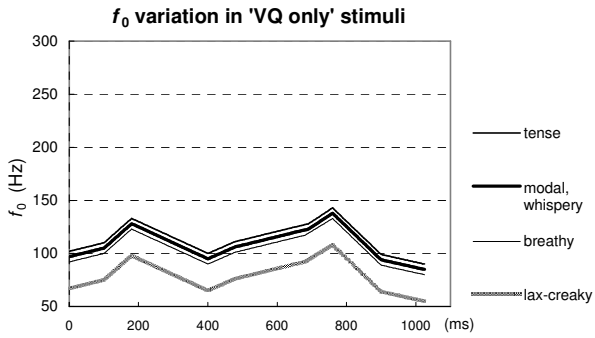


Figure 2: *Intrinsic $f_0$ variation in 'VQ only' stimuli.*

For **'VQ + $f_0$' stimuli,** non-neutral $f_0$ contours were combined with the voice qualities of the 'VQ only' group as shown in Table 1. The choice of voice quality to be combined with a particular $f_0$ contour was guided by the results in [3] as well as by comments in the literature.

It should be borne in mind that the lax-creaky voice quality in the 'VQ only' series has the lowest $f_0$ of all the stimuli: thus, of the two stimuli with lax-creaky voice, 'VQ only' and 'VQ + $f_0$', the former is the one whose $f_0$ contour deviates the most from the neutral $f_0$ contour of modal voice (see Fig. 2).

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' |

## 3. Listening tests

The perception test was conducted according to the procedure described in [3] and [7] as a series of six subtests with 21 native speakers of Japanese as participants. 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 by two native speakers of Japanese.

The participants judged each stimulus for the presence and strength of affect, and marked 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 the utterance had no affective load; 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 ratings were then interpreted as 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 TSD test were conducted to explore the difference in perception of various stimuli. The significance level was set at $p < .05$.

## 4. Results

Table 2 shows the stimuli yielding the highest rating for each stimulus type '$f_0$ only', 'VQ + $f_0$' and 'VQ only', for each of the affects tested. In Fig. 3 is shown the mean rating for the most highly scored stimulus for each affect (maximum mean rating) – again, for each stimulus type. As was noted in the previous experiment [7], a particular stimulus tended to be associated with more than one affect, e.g., lax-creaky voice is associated with *bored* and *sad*. These clusters of affects are shown in Fig. 3 as Groups A, B and C. Note that Group D is made up of the remaining affects, which did not group into such clusters, and for which generally very low affective ratings were obtained. The discussion and comparison of results are presented with regard to these groupings.

Table 3 presents a somewhat different perspective on results, showing for each stimulus in the test, where an affective mean rating of 1 or more was obtained for at least 16 of the 21 subjects (more than 76%). Although this is a somewhat ad hoc filter, it does serve the purpose here of focussing on the stimuli which most consistently evoked a particular affective colouring.

**Group A: *interested, intimate, *indignant, *happy*.** In this group, the stimuli that received the highest ratings were tense + $f_0$ 'indignation', modal + $f_0$ 'indignation' and tense voice. For the first two of these affects, *interested* and *intimate*, the combined stimulus yields the highest rating, though not significantly higher than for the '$f_0$ only' stimulus. The tense voice on its own has little effect, which would indicate that the affect is principally cued by the very high and dynamically varying contour of $f_0$ 'indignation'.

The association of *intimacy* with the $f_0$ 'indignation' contour was somewhat unexpected, as was the fact that tense voice appears to enhance the effect, if only slightly. These results are unlike our results for Hiberno-English listeners, for whom the very different whispery voice quality was the pre-

ferred option. It is often assumed that breathy voice quality is associated with *intimacy*, but these results highlight the fact that such associations may be highly language-specific.

Table 2: *Stimuli yielding the highest rating within each group. * = affects for which specific $f_0$ contours were available.*

| Affect | '$f_0$ only' | 'VQ + $f_0$' | 'VQ only' |
|---|---|---|---|
| *interested*<br>*intimate*<br>**indignant*<br>**happy* | modal + $f_0$ 'indignation' | tense + $f_0$ 'indignation' | tense |
| **bored* | modal + $f_0$ 'boredom' | lax-creaky + $f_0$ 'boredom' | lax-creaky |
| **sad* | modal + $f_0$ 'sadness' | | |
| *apologetic*<br>**scared* | modal + $f_0$ 'fear' | whispery + $f_0$ 'fear' | whispery |
| *fearless* | modal + $f_0$ 'boredom' | tense + $f_0$ 'indignation' | tense |
| *formal* | modal + $f_0$ 'boredom' | lax-creaky + $f_0$ 'boredom' | lax-creaky |
| *stressed* | modal + $f_0$ 'fear' | whispery + $f_0$ 'fear' | tense |
| *relaxed* | modal + $f_0$ 'indignation' | tense + $f_0$ 'indignation' | breathy |

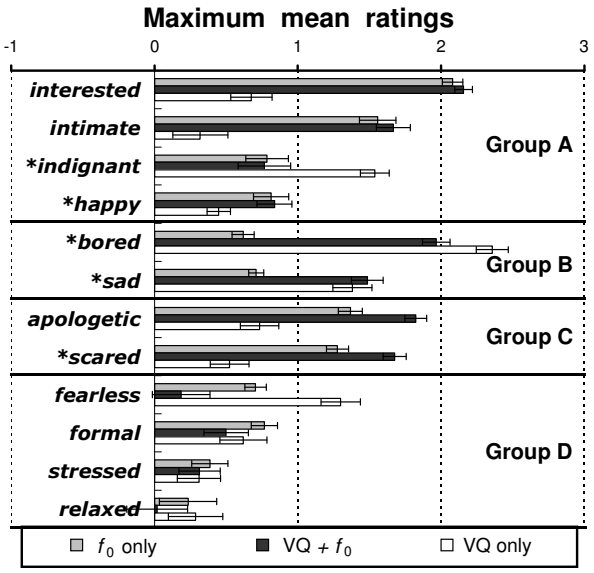


Figure 3: *Maximum mean rating and estimated standard error of the mean for the three groups of stimuli. Affect ratings: 0 = none, 3 = max.*

For the affect *indignant*, although the stimuli achieving highest ratings are the same as for the other affects in the group, the response pattern is rather different. Here, a tense voice quality yields considerably higher ratings than either the combined stimulus or the '$f_0$ only' stimulus. It seems to be the case that the very high $f_0$ level and the large dynamic range of the $f_0$ 'indignant' are not effective in cueing *indignation* for these Japanese listeners. This is a further point of contrast with Hiberno-English listeners [7]. Of course, given the strong association of $f_0$ 'indignation' with the very different affect, *intimacy*, it is hardly surprising that indignation must involve some other means of expression: intuitively one would expect indignation and intimacy to rely on rather different voice characteristics.

Ratings for the *happy* affect were very low regardless of stimulus type. This was also the finding of our earlier experiments [3,7]. Happiness has often been reported as an affect accurately recognised in the facial expression, and notoriously difficult both to portray and to identify in vocal expressions, e.g. [1].

Table 3: *Stimuli for which mean ratings ≥ 1 are obtained by more than 76% of subjects.*

| Stimuli \ Test | Sad-Happy | Relaxed-Stressed | Intimate-Formal | Fearless-Scared | Bored-Interested | Apologetic-Indignant |
|---|---|---|---|---|---|---|
| **'$f_0$ only'** | | | | | | |
| modal+$f_0$'fear' | – | – | – | – | – | **A** |
| modal+$f_0$'sadness' | – | – | – | – | – | – |
| modal+$f_0$'boredom' | – | – | – | – | – | – |
| modal+$f_0$'joy' | – | – | – | – | – | – |
| modal+$f_0$'indignation' | – | – | **I** | – | **I** | – |
| **'VQ + $f_0$'** | | | | | | |
| whispery+$f_0$'fear' | – | – | – | **S** | – | **A** |
| breathy+$f_0$'sadness' | – | – | – | – | – | **A** |
| lax-creaky+$f_0$'boredom' | – | – | – | – | **B** | – |
| tense+$f_0$'joy' | – | – | – | – | **I** | – |
| tense+$f_0$'indignation' | – | – | **I** | – | **I** | – |
| **'VQ only'** | | | | | | |
| whispery | – | – | – | – | **B** | – |
| breathy | – | – | – | – | **B** | – |
| lax-creaky | – | – | – | – | **B** | – |
| tense | – | – | – | **F** | – | – |
| modal | – | – | – | – | – | – |

**Group B: *bored, *sad.** Highest ratings for these two affects were obtained with a lax-creaky voice quality, either on its own or in combination with the $f_0$ 'bored' contour. As is clear from Fig. 3, the '$f_0$ only' stimuli do not appear to be effective. These results are broadly similar to those of our Hiberno-English subjects.

A certain amount of caution is needed in interpreting the role of $f_0$: it should be remembered that the lax-creaky voice quality had an intrinsically low $f_0$ – considerably lower than in the "combined" stimulus. Thus the fact that higher ratings were obtained for *boredom* with the 'VQ only' than with the 'VQ + $f_0$' stimulus suggests that the very low $f_0$ typically associated with creaky voice *does* contribute something to the signalling of *boredom* for these listeners. For similar reasons, we can deduce that a very low $f_0$ does *not* contribute to cueing *sadness*.

**Group C: *apologetic, *scared.** Similar ratings as well as similar overall pattern of responses were found for the stimuli most effective in cueing these affects (Fig. 3, Table 2). As for our Hiberno-English results in [7], the combined stimulus whispery + $f_0$ 'fear' received considerably higher ratings than those obtained for any of the components – '$f_0$ only' or 'VQ

only' – on their own. Nonetheless, the high ratings obtained for the '$f_0$ only' stimulus (modal + $f_0$ 'fear') implies a considerable contribution of the intonation variable in conveying the affects *apologetic* and *scared*. Perception of the 'VQ only' stimulus (whispery voice) was ineffective in cueing these affects.

**Group D:** ***fearless, formal, stressed, relaxed.*** This group includes the affects for which no systematic 'clustering' was found. Other than for *fearless* – which patterned rather like *indignant* (in Group A) and yielded a maximum mean rating higher than 1 for tense voice ('VQ only') – the ratings obtained for the stimuli in this group were very low.

Results for *formal, stressed* and *relaxed* are very different to those of the Hiberno-English listeners in [7], who gave relatively higher ratings for these affects, particularly with regard to stimuli involving voice quality cues.

## 5. Discussion and conclusions

These results accord with the results of earlier experiments [2,3,7] in showing that there is no clear one-to-one mapping between voice attributes and affect. Particular stimuli may be correlated with a cluster of affective attributes, some being more strongly signalled than others. Furthermore, a particular affect may be associated with more than one type of stimulus.

Overall, analysis of the data obtained from the Japanese subjects showed that the stimuli incorporating distinct voice qualities (with or without $f_0$ variations) always yielded the highest ratings. These findings are broadly in keeping with results obtained for Hiberno-English speakers [7]. Similarly, in Table 3 we see that – while all stimuli incorporating non-modal voice qualities ('VQ + $f_0$' and 'VQ only') are consistently associated with at least one affect – of the '$f_0$ only' stimuli, only two appear to have a consistent affective signalling potential: $f_0$ 'indignation' and $f_0$ 'fear'. These are of course the stimuli involving very high $f_0$ levels: as mentioned above, not only were they found to be potent cues to affect, but in some important respects their signalling role appears to be different to that found for the Hiberno-English listeners. Note also that the affects for which the study in [6] provides $f_0$ contours did not on the whole yield high affective ratings.

Comparing the ratings for the 'VQ only' and '$f_0$ only' stimuli, we note that these manipulations on their own are only effective in a minority of (rather different) affects. '$f_0$ only' stimuli got relatively higher ratings for the affects *interested, intimate, apologetic* and *scared.* 'VQ only' got relatively higher ratings for the affects *bored, indignant, sad* and *fearless.* This is quite a different picture to that found for our Hiberno-English listeners for whom 'VQ only' stimuli virtually always yielded considerably higher ratings. For the Japanese subjects, the $f_0$ contour appears to be of relatively greater importance, but note that in all cases, the addition of some non-modal voice quality results in higher ratings for the $f_0$ manipulated stimuli.

There are three cases where the 'VQ only' yield considerably higher ratings than the combined 'VQ + $f_0$' stimuli: for the affects *indignant*, *fearless* and *bored*. We would suggest that for the first two affects this is likely to reflect a non-optimal matching of voice quality to pitch contour. For *bored*, the optimal $f_0$ would appear to be in any case the very low $f_0$ of lax-creaky voice.

In this context, it is important to emphasise the exploratory nature of these experiments. The 'broad-palette' approach has involved giving listeners a palette of vocal colours and letting them indicate what their affective associations might be. It is not assumed that we have necessarily covered all the voice qualities that might be important, all the $f_0$ differences that might signal affect, nor indeed are the combinations of the two necessarily optimal.

While results accordingly cannot provide definitive statements concerning the roles of voice quality and affect, they offer many insights, and point to similarities as well as striking differences between the Japanese and Hiberno-English listener groups.

Although space limitation precludes a full discussion, some observations are worth noting. First of all, there is the relatively greater importance of the $f_0$ contour, at least when it involves very high levels. Secondly, there are the striking gaps in affective signalling for affects *formal, stressed* and *relaxed* which Hiberno-English listeners rated relatively highly, responding particularly to specific voice qualities. Thirdly, there are cases of a strikingly different attribution of affect to the same stimulus by the two language groups: for example, the combined stimulus tense + $f_0$ 'indignation', which appears to be readily associated with indignation for the Hiberno-English subjects, is associated with intimacy by the Japanese.

As a next step, our objective is to carry out a fuller comparison of the two groups, and to extend the study to further language groups. Insofar as one motivation for this research is to guide the generation of affective synthetic speech, these results do indicate that this objective will need to be pursued in a way that is sensitive to the language/cultural groupings involved.

## 6. Acknowledgments

This work has been carried out as part of the EU-funded Network of Excellence on Emotion, HUMAINE. The authors would like to thank Mika Ito for her kind assistance in collecting the Japanese data.

## 7. References

[1] Elfenbein, H.A.; Ambady, N., 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* 128, 203-235.

[2] Gobl, C.; Bennett, E.; Ní Chasaide, A., 2002. Expressive synthesis: how crucial is voice quality?. *Proc. of the IEEE Workshop on Speech Synthesis*, Santa Monica, California, paper 52, 1-4.

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

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

[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 XIIIth International Congress of Phonetic Sciences*, Stockholm, 178-181.

[7] Yanushevskaya, I.; Gobl, C.; Ní Chasaide, A., 2005. Voice quality and $f_0$ cues for affect expression: implications for synthesis. *Proc. of INTERSPEECH 2005*, Lisbon, 1849-1852.