Voice quality and f_0 cues for affect expression: implications for synthesis

Irena Yanushevskaya, Christer Gobl & Ailbhe Ní Chasaide

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

Abstract

Synthesised stimuli were used to investigate how two notionally separable dimensions of tone-of-voice – voice quality and fundamental frequency – are involved in the expression of affect. Listeners were presented with three series of stimuli: (1) stimuli exemplifying different voice qualities, (2) stimuli all with modal voice quality but with different affect-related f_0 contours, and (3) stimuli incorporating variation in both voice quality and affect-related f_0 contours. A total of 15 stimuli were rated for 12 different affective attributes. Voice quality differentiation appears to account for the highest affect ratings overall, as indicated by the scores obtained for stimuli series (1) and (3). The relatively weaker affect signalling of stimuli differentiated by f_0 alone corroborates findings in [2]. It also suggests that for the generation of expressive, affectively coloured speech synthesis, it is not sufficient to manipulate only f_0; we also need to capture the voice quality dimension of the voice source.

1. Introduction

In interpersonal communication, the vocal dimension combines with the visual signalling of facial expression and gesture to infuse the informational content of utterances with the complex expressive nuances of our true or simulated feelings and attitudes. Even when the visual information is not available, these nuances are largely retained, as listeners are finely attuned to the speaker's tone-of-voice and to subtle differences in the temporal structure.

The present paper looks at two notionally separable dimensions of tone-of-voice, voice quality and f_0, and through synthesis-based experiments aims to explore how they collaborate in expressing affect. There is to date little empirical knowledge on how voice quality variation signals affect, although it is widely thought to be of central importance. In comparison, there has been a substantial literature on the role of f_0 and temporal variation, e.g., [6, 8], at least as concerns the expression of strong emotions such as fear, anger, joy, etc.

This work builds on earlier research [3], which demonstrates how synthetic stimuli differing in voice quality can bring about different affective colouring in a single utterance. In a subsequent experiment [2], stimuli differing in voice quality were further modified to incorporate large f_0 differences, as described by Mozziconacci for a range of affects [6]. Listeners rated the affective strength of these 'combined' stimuli in comparison to stimuli which included only the f_0 modification (i.e. all having modal voice quality). The former 'combined' stimuli received much higher ratings for affect than the latter. It was not clear however, what the affective contribution of the affect-related f_0 contours was: it could be that the high ratings for the 'combined' stimuli were largely a consequence of the voice quality variation. This was not possible to assess, as there was no series varying in voice quality only, without the affective f_0 contours. Therefore, in the present experiment, three sets of stimuli are presented to include such an option.

As in the earlier studies, a 'broad-palette' approach is adopted, whereby the listener is confronted with a range of stimuli, and rates them in terms of a rather wide selection of affective attributes. These incorporate not only the 'big' emotions, but also milder affective states and attitudes (such as *relaxed* and *formal*). Furthermore, the voice qualities generated were based on prior analyses of different voice qualities and not on a prior investigation of affective speech.

A further aim of this work is to provide insights into how expressive synthesis might be achieved. Given the lack of empirical data on the use of voice quality, this approach of synthesis and perceptual testing offers a way of guiding our attempts to synthesise affect in speech, providing initial, experimentally derived hypotheses on the mapping of the two.

2. Synthetic stimuli

The perception test involved 15 synthetic stimuli of a Swedish utterance – "ja adjö [ˈjɑː aˈjøː] – generated using the KLSYN88 formant synthesiser [4]. These stimuli divide into three groups of five stimuli, which are labelled as 'VQ only', 'f_0 only', 'VQ + f_0' (see Table 1). The 'VQ only' group is made up of stimuli which are differentiated in terms of voice quality; the 'f_0 only' stimuli are differentiated by having affect-related f_0 contours, based on the contours in Mozziconacci [6]; and the 'VQ + f_0' stimuli combine these affect-related f_0 contours with the voice quality which was deemed to be the most appropriate for these affects.

'VQ only' stimuli. The synthesised voice qualities include modal voice, breathy voice, whispery voice, lax-creaky voice and tense voice. These stimuli aim to simulate voice qualities according to the voice quality classification system outlined by Laver [5]. The exception is lax-creaky voice, which is conceptually an extension of the Laver framework (for further discussion, see [3]). The stimuli are essentially a subset of those used in [3] where harsh voice and creaky voice were also used. They were omitted here in order to reduce the overall number of stimuli for the perception test.

One voice quality that is somewhat different here than in [3] is whispery voice. Note that this quality was problematic in the earlier experiment and was therefore modified to provide a more satisfactory rendition. Changes were made to the dynamics and the level of the aspiration noise: on average the aspiration noise level was lowered by about 4 dB.

Note that the 'VQ only' series of stimuli do in fact incorporate some f_0 differences. These differences were deemed as 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. Mozziconacci [6] provides quantitative data for f_0 contours associated with the following affective states: indignation, anger, joy, fear, boredom, sadness and neutral (see Figure 1). As the f_0 contour of our original modal utterance is very similar to that of Mozziconacci's neutral f_0 contour, Mozziconacci's f_0 contours could easily be adapted to our synthetic stimuli by a simple proportional scaling of the values in Figure 1.

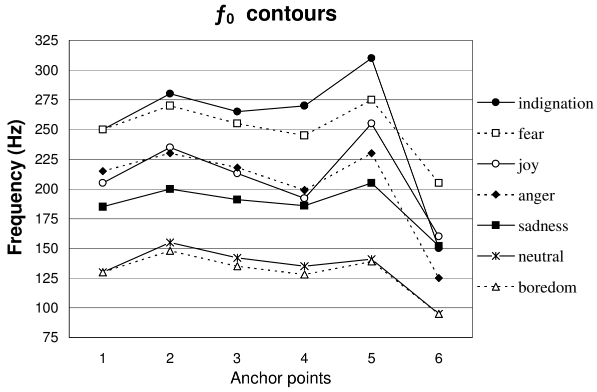

Figure 1: Affect-related fundamental frequency contours as described in [6].

Again, to limit the number of stimuli, we decided to exclude the f_0 contour for anger in favour of the f_0 contour for indignation which displays more extreme f_0 excursions. Thus for this group of stimuli, five non-neutral f_0 contours were generated – by modifying the f_0 values of the modal stimulus – relating to the following affects: indignation, joy, fear, sadness and boredom.

Note that the affect-related contours involve an f_0 level which is raised relative to the neutral, with the exception of the one for boredom, which is very slightly lowered.

'VQ + f_0' stimuli. For this set of stimuli, each of the five non-neutral f_0 contours was combined with one of the voice qualities of the 'VQ only' group. In other words, these stimuli were generated by modifying the f_0 values of a non-modal quality from the 'VQ only' group. The five stimuli of this group were obtained by combining voice quality and f_0 contours as follows: the f_0 contour of indignation and joy were combined with tense voice, f_0 'fear' was combined with whispery voice, f_0 'boredom' with lax-creaky voice, and f_0 'sadness' with breathy voice. 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 is worth pointing out that the lax-creaky voice quality (in the 'VQ only' series) has an intrinsically lower f_0 than any of the affect-related contours derived from [6], and lower therefore than any of the f_0 contours in the 'f_0 only' series or the 'VQ + f_0' series. In other words, one should note that for the two stimuli with lax-creaky voice, the 'VQ only' stimulus is the one whose f_0 contour deviates the most from the neutral f_0 contour.

'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'

Table 1: Synthesised stimuli.

3. The perception test

The perception test was conducted as a series of six sub-tests according to the procedure described in [3] with 20 native speakers of Hiberno-English 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 participants were asked to judge for each stimulus whether the speaker sounded more sad or happy, etc., and mark their response on the answer sheet where the opposite affective labels were placed on each side with seven boxes in between. Subjects were asked to choose the centre box if they considered the utterance to bear no affective load; checking the boxes to the left or right to the centre box was meant to indicate the presence and strength of a particular affect, the most extreme ratings being further from the centre box. The pairs of affective attributes tested were *sad-happy*, *intimate-formal*, *relaxed-stressed*, *bored-interested*, *apologetic-indignant*, and *fearless-scared*.

A one-way ANOVA with voice quality as a factor as well as the Tukey's TSD test were conducted to explore the difference in perception of various voice quality stimuli. The significance level was set at p < .05.

4. Results and discussion

Figure 2 illustrates the highest mean ratings obtained for each affect according to the three stimulus types: 'f_0 only' (grey bars – affect related f_0 contours coupled with modal voice), 'VQ + f_0' (black bars – stimuli incorporating distinct voice qualities coupled with affect related f_0 contours) and 'VQ only' (white bars – voice quality only). The lines through the bars show the estimated standard error of the mean. Table 2 complements Figure 2, indicating for each affect, which stimulus yielded the highest rating within each stimulus group. Particular affects marked with an asterisk are those for which we have specific f_0 contours (see Figure 1), and thus the 'f_0 only' and the 'VQ + f_0' stimuli were expected to be the most highly rated for these particular affects.

Certain groupings emerged in the results: within each series of stimuli, a particular stimulus appeared to be associated with a cluster of affects, e.g., whispery voice is associated with *scared*, *intimate*, and *apologetic*. This is very much in keeping with the findings in our earlier studies, where it is clear that there is no simple one-to-one mapping between quality and affect [2, 3]. Coming at the same question from a different angle, Campbell [1] has likewise emphasised that utterances in naturalistic corpora tend not to be associated with single affect labels, but rather with constellations of affective labels. To facilitate the discussion of results they are presented in terms of these groupings.

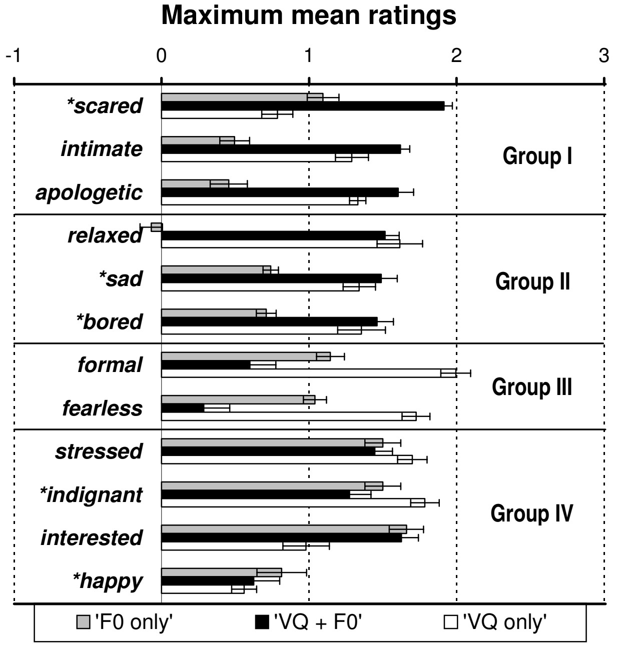

Figure 2: For each affect is shown the maximum mean rating and estimated standard error of the mean for the three groups of stimuli 'f_0 only' (grey), 'VQ + f_0 (black), and 'VQ only (white). Affect ratings: 0 = none, 3 = max.

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

Table 2: Stimuli yielding the highest rating within each stimulus group.

Overall, the stimuli which incorporate distinct voice qualities (whether 'VQ only' or 'VQ +f_0') are most effective in yielding the highest ratings. The 'f_0 only' stimuli yield generally low ratings, excepting the affects *indignant*, *stressed* and *interested*. Only for *interested* was the rating significantly higher than for the 'VQ only' stimulus. Conversely, ratings for the 'VQ only' stimuli were significantly higher for the 'f_0 only' ones in seven of the affects tested. Comparing the ratings for the 'VQ only' and the 'VQ + f_0' stimuli, only in the case of the affect *scared* did the addition of a non-neutral f_0 contour yield a significantly higher rating.

Group I: *scared*, *intimate* and *apologetic*. In this group, the 'VQ + f_0' stimuli gave the highest ratings. Whispery voice is the preferred voice quality, particularly when combined with the f_0 contour for *fear*. For the affect *scared* it is clear that the combination of voice quality and f_0 is greatly more potent than either of these dimensions on their own. In the case of *intimate* and *apologetic*, it is again the case that the combined stimulus 'VQ +f_0' produces the highest ratings, even though as mentioned the f_0 enhancement of the 'VQ only' does not reach significance.

Traditionally, phoneticians have associated a breathy voice quality with intimacy, whereas in this study, the whispery quality yielded much higher ratings.

For the affects *intimate* and *apologetic*, the contribution of the f_0 contour is minor, as can be seen from the very low ratings obtained for the 'f_0 only' stimuli, and from the fact that the difference between the 'VQ + f_0' and the 'VQ only' stimuli is not statistically significant. For these two affects, the whispery voice quality appears to be the dominating cue.

We would hesitate to conclude, however, that pitch is irrelevant in the cueing of *intimate* or *apologetic*. The high pitch contour associated with fear was the only contour combined with whispery voice. If one were to have a greater variety of pitch contours combined with whispery voice, it is conceivable that one would find a more definite contribution from pitch. There is no *a priori* reason to expect *intimate* and *apologetic* to be associated with high pitch: intuitively one might expect low pitch. On the other hand, Ohala's "frequency code" [7] could be interpreted to suggest that a raised f_0 level would be associated with these affects. Support for this is suggested by the fact that among the 'f_0 only' stimuli, listeners opted for the relatively high f_0 contour of *fear*, rather than for one of the lower pitched contours, even if the ratings were low.

Group II: *relaxed, sad* and *bored*. For these affects, the lax-creaky voice quality is favoured and appears to be chiefly responsible for the strength of rating. Lax-creaky voice was also found to produce high ratings with these affects in [3]. Traditionally, phoneticians have associated creaky voice with boredom and breathy voice with sadness. The lax-creaky quality combines creakiness with an underlying lax/breathy phonatory setting, and appears to be considerably more potent in cueing these affects than breathy voice or creaky voice on their own.

The present results also throw light on a question arising from [3] concerning the contribution of the inherently low f_0 of lax-creaky voice to the high ratings obtained for these affects. In the present experiment, the 'VQ only' lax-creaky stimulus has similarly the low f_0 deemed to be associated with this quality, but this time, listeners are also presented with a relatively higher pitch contour for the lax-creaky + f_0 'boredom' stimulus. However, this f_0 difference does not yield a significant difference in the ratings. Given that the stimulus with the higher f_0 achieved relatively higher ratings for both *sad* and *bored* in this test, we would conclude here that a low f_0 is not a necessary correlate of these affects. This conclusion is further corroborated by the relatively low scores of the 'f_0 only' stimuli here, as well as by the fact that the highest rating among the 'f_0 only' stimuli was for that with the f_0 'sadness' contour, which is considerably higher than the neutral.

Group III: *formal* and *fearless*. Here a tense voice quality was the only stimulus to yield strong ratings. From the combined 'VQ + f_0' series, the highest ratings obtained (tense + f_0 'indignation') were very low relative to tense voice on its own, suggesting that high pitch is distinctly disfavoured for these affects. Such a conclusion is also suggested by the highest rating obtained for the 'f_0 only' series, where the lowest available f_0 contour was chosen (f_0 'boredom').

Group IV: *stressed*, *indignant*, *interested* and *happy*. For the affects *stressed* and *indignant*, the tense voice quality yielded the highest rating. For these affects the very high f_0 contour of the f_0 'indignation' is also effective in cueing these states. The difference in the ratings between these two stimulus types is not statistically significant. Curiously, the addition of the f_0 'indignation' contour to tense voice significantly reduces the ratings for *indignant*.

For the affect *interested*, the highest ratings were obtained for the 'f_0 only' stimulus (f_0 'indignation'). Although tense voice was the most highly rated of the 'VQ only' series, the rating is nonetheless low. We can also see from the most highly rated combined stimulus (tense + f_0 'indignation') there is no enhancement of what the high pitch contour can achieve on its own.

Finally, for the affect *happy* none of the present stimuli achieved high ratings. This finding echoes many studies where happiness or joy has not been easy to elicit. It is likely that visual cues (facial expression) as well as the formant shifts associated with smiling are important in signalling this affect.

5. Implications for synthesis

As mentioned in the introduction, one of our goals is to formulate more experimentally based hypotheses as regards how tone-of-voice in synthesis might be shaped to the affective requirements of the context.

The results do allow us to refine on the otherwise default assumptions that tend to be made about the voice qualities that are associated with specific affects. For example, rather than the traditional association of boredom with creaky voice and intimacy or sadness with breathy voice, the present results (along with the results in [3]) point to a lax-creaky quality as being considerably more effective.

The present approach also illustrates how we might garner information on how the milder affective states, such as formal, relaxed, apologetic, etc., might be cued in synthesis. Virtually all the literature in the field has tended to focus on strong emotions, while in real life synthesis applications these may be the least required.

The present results also indicate that voice quality differentiation is likely to be crucially important for expressive synthesis. The relatively low affect ratings obtained in this and in a previous study [2] for the 'f_0 only' stimuli when compared to stimuli that have voice quality differentiation underscore the fact that f_0 manipulation alone is not likely to be a successful approach.

The reason for the generally low ratings obtained for the 'f_0 only' stimuli may also have to do with the fact that there is a natural covariation between f_0 and other source parameters, which may be violated when f_0 is manipulated on its own, resulting in a decrement in its affective potency. The rating of the 'f_0 only' stimuli is in some ways not very surprising in view of the fact that researchers who have observed f_0 correlates of emotions in production data have often failed to demonstrate their effectiveness in perception. This gap between production and perception may partly be due to the voice quality deficit.

In current state-of-the-art synthesis systems, there is limited control of voice source parameters. The f_0 contour can in principle be readily manipulated, and as most of the available information on vocal affect concerns f_0 dynamics, this would at first glance appear to be the obvious way to proceed. However, in current synthesis systems there is a reluctance to manipulate f_0, because of the potential loss in quality of the output. The present results suggest strongly that f_0 manipulation alone will not work, and indicate why this might be.

6. Conclusions

Voice quality differentiation appears to account for the highest affect ratings overall, as indicated by the scores obtained for the 'VQ only' and 'VQ + f_0' stimuli. On the other hand, the 'f_0 only' stimuli yielded relatively weak affect cueing, with the exception of the affects *indignant*, *interested* and *stressed*. This relatively weaker affect signalling of f_0 alone was also found in the earlier study [2].

The low affective ratings generally obtained for the 'f_0 only' series suggest that f_0 manipulation will simply not deliver adequate signalling of affectively coloured synthetic speech. In order to generate expressive synthesis we will need to capture the voice quality dimension of the source, and understand how it combines with pitch variation.

7. Acknowledgments

This work is supported by the EU-funded Network of Excellence on Emotion, HUMAINE, and by the project Prosody of Irish Dialects, funded by the Irish Research Council for Research in the Humanities and Social Sciences.

8. References

[1] Campbell, N., "Perception of affect in speech – towards an automatic processing of paralinguistic information in spoken conversation", *Proc. of INTERSPEECH 2004*, Jeju Island, Korea, Vol. 2, pp. 881-884.

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

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

[4] 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*, Vol. 87, 820-857, 1990.

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

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

[7] Ohala, J., An ethological perspective on common cross-language utilization of f_0 of voice", Phonetica, Vol. 41, 1-16, 1984.

[8] Scherer, K.R., "Vocal affect expression: A review and a model for future research", *Psychological Bulletin*, Vol. 99, 143-165, 1986.