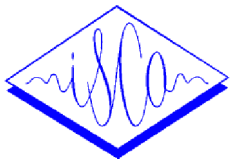

Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

Voice Quality and Loudness in Affect Perception

Irena Yanushevskaya, Christer Gobl and Ailbhe Ní Chasaide

Phonetics and Speech Laboratory
Centre for Language and Communication Studies
School of Linguistic, Speech and Communication Sciences
Trinity College Dublin
{yanushei; cegobl; anichsid}@tcd.ie

10.21437/SpeechProsody.2008-157

Abstract

Different voice qualities tend to vary in terms of their intrinsic loudness. Perceptual experiments have shown that voice quality variation can be strongly associated with the affective colouring of an utterance. The question addressed in this paper concerns the role that the intrinsic loudness variation might play in this voice quality-to-affect mapping. To test the hypothesis that the intrinsic loudness variation is *not* a major determinant of the perceived affective colouring, listeners rated the affective colouring of two series of stimuli: one series varied in voice quality and contained intrinsic loudness variation; the other series were of a constant voice quality, but matched loudness variations of the first series. The results overall support the hypothesis that loudness contributes relatively little to the perceived affective colouring of specific voice qualities. But variation in loudness (in the absence of voice quality variation) is not entirely irrelevant: some contribution of loudness to certain high activation affects was also found.

1. Introduction

Prosodic features such as pitch, voice quality, loudness, and timing organisation of speech play a fundamental role in conveying emotions and attitudes in human communication. Specific emotions are communicated by particular combinations of pitch and loudness as well as the speaking rate. For example, Scherer [1] suggests that happiness, elation, fear and rage are indicated by an increase in pitch, intensity and speech rate, whereas sadness and dejection as well as boredom are signalled by a decrease in all these parameters. In other words, there are affects that are expressed loudly and others of which a low intensity is typical. As was noted by Frick [2], 'contempt is loud and grief and boredom are soft'.

When speakers vary their voice quality, there is typically a concomitant variation in loudness. Likewise, in synthesised speech, if we alter parameters of the glottal source pulse to effect voice quality differences changes in amplitude ensue, effecting differences in loudness.

Experiments reported in [3, 4], show a clear mapping between specific voice qualities and perceived differences in affect. For example, a lax-creaky voice quality tends to yield high ratings for affects such as boredom, sadness and a relaxed state, whereas a tense voice quality tends to be rated as signalling anger, happiness and stress. The stimuli used in those experiments involved manipulations to the glottal source parameters, and as mentioned, such manipulations engender variations in loudness. Although in real life one assumes that there will be a tendency for loudness and voice quality to covary, these can also be independently controlled to some extent: thus, for example, we can produce modal voice at different loudness levels.

A question arises in interpreting the results of these studies, as to whether the loudness variation of the stimuli might account for the differences in perceived affect colouring. Our initial hypothesis is that affective cueing is *not* simply a consequence of the loudness variation in these voice quality stimuli.

In order to test this hypothesis an experiment was designed where affective ratings were elicited using two series of synthetic stimuli. The first series, the 'voice quality' stimuli, differed in terms of voice quality and included intrinsic loudness variations. The second series, the 'loudness' stimuli, involved stimuli with matching loudness to those of the voice quality series, but whose voice quality was modal for the whole series.

2. Synthesised stimuli

11 synthesised stimuli were used in all. The base stimulus was a high quality copy synthesis of the Swedish utterance "ja adjö ['jɑː a'jøː], which was also used in [3]. This utterance was deemed affectively neutral for the participants of the experiment, all speakers of Irish-English. The stimuli were generated using the KLSYN88a formant synthesiser [5] and exploiting the modified version of the LF voice source model [6], which is available as an option in this synthesiser.

'Voice quality' stimuli. The synthesised voice qualities include modal voice, whispery voice, breathy voice, lax-creaky voice, harsh voice and tense voice. The source parameters manipulated were OQ (open quotient), TL (spectral tilt) SQ (speed quotient) AH (aspiration noise) and AV (amplitude of voicing). B1 and B2 (bandwidth of the first and second formants) were also manipulated.

These stimuli aim to simulate voice qualities according to the voice quality classification system outlined by Laver [7]. The exception is lax-creaky voice, which is conceptually an extension of the Laver framework. For further discussion, and for details concerning the manipulations, see [3]. One voice quality that is somewhat different here than in [3] is whispery voice. This quality was problematic in the earlier experiment and was therefore modified to provide a more satisfactory rendition.

'Loudness' stimuli. On the basis of the modal voice quality stimulus in the first series, five new stimuli were generated. Each of the new stimuli matched the level of loudness of one of the original non-modal voice quality stimuli (whispery, breathy, lax-creaky, tense, and harsh). E.g., there was a modal voice stimulus with the loudness matching that of tense voice, a modal stimulus with the loudness matching that of breathy voice, and so on. (See Section 3 for details.)

Loudness is a perceptual attribute of a sound; it could be defined as the subjective strength of a sound. According to Scharf [8], "loudness resides in the listener, not in the stimulus". Perception of loudness is influenced by not only the intensity of the signal, but by the signal's frequency components and bandwidth, as well as the background against which the sound is presented [8]. For example, spectral tilt will influence the listener's perception of loudness [9]. Differences in spectral tilt are important in differentiating among voice qualities: thus when we change voice quality, we do tend to vary the loudness. But as pointed out earlier, we can also produce a particular voice quality at different loudness levels.

Simply adjusting the intensity level of the modal stimulus so as to match the intensity level of the original non-modal stimuli might not be satisfactory in generating the desired match in loudness. An auditory experiment was therefore carried out.

3. Preliminary test: loudness matching

A preliminary listening test was carried out using stimuli with modal voice quality, but where the loudness was systematically changed. The purpose of the test was to find the stimuli that would best match in terms of loudness each of the original voice quality stimuli.

The modal voice quality was chosen as the basic stimulus, and its intensity level was changed in steps of 1 dB to provide a selection of sounds, which could then be compared to the original voice quality stimuli.

A set of 24 stimuli was thus prepared each stimulus being given a numeric value corresponding to the change in dB. The 'quietest' stimulus (Stimulus –12) had an intensity level that was 12 dB less than that of the original modal voice stimulus, and the 'loudest' stimulus (Stimulus +12) had an intensity level that was 12 dB higher. As the original modal voice stimulus (Stimulus 0) was also included in the set, the total number of stimuli was 25.

To obtain the desirable intensity values, the waveform of the original modal stimulus was multiplied by appropriate scaling factors to effect an increase/decrease of the intensity level of 1 dB, 2 dB, etc. The resulting stimuli were arranged according to increasing intensity from the lowest intensity to the highest intensity, with the original modal voice in the middle of the range. This order was kept constant as the range of stimuli was presented to the listeners.

16 native speakers of Hiberno-English participated in the listening test. They were instructed to listen in turn to each of the five original non-modal voice quality stimuli, labelled Sound A, Sound B, etc., and to select for each voice quality stimulus the best loudness match out of the range of 25 stimuli.

The stimuli were played through a high quality speaker in a quiet room. The participants were allowed to listen to the stimuli as many times as they needed to make a decision, and then to mark the responses on an answer sheet. The five original voice quality stimuli were presented 5 times in a randomised order (a total of 16 x 5 = 80 responses for each non-modal voice quality, or 400 in all). The average measures Intraclass correlation coefficient (ICC) calculated to test the overall consistency of the stimuli ratings by the participants in the experiment was found to be relatively high at 0.992.

For the stimuli that were selected by the participants of the auditory test as best matching each of the non-modal voice qualities, the mean dB value was calculated. This mean value represents the change required to match the loudness of each of the non-modal voice qualities. These values are shown in Table 1, together with the corresponding scaling factors. Standard deviation of the mean values in dB is shown in brackets. The five 'loudness' stimuli were generated by scaling the amplitude of the original modal waveform with each of the five scaling factors. These 'loudness' stimuli, together with the 'voice quality' stimuli (11 in total) were used in the subsequent series of perception experiments to test our hypothesis that loudness per se is not the main determinant of the affective colouring associated with voice quality.

Table 1: *Scaling factors (and corresponding dB differences) used to generate the matched 'loudness' series. Values in brackets show standard deviations.*

Stimuli	Scaling factor	Difference in dB
L_whispery	0.43	-7.3 (1.17)
L_breathy	0.63	-4.0 (1.34)
L_lax-creaky	0.72	-2.8 (0.86)
L_harsh	1.35	+2.6 (1.47)
L_tense	1.43	+3.1 (1.18)

4. Affect-mapping experiment

The 11 stimuli were presented to 16 subjects, native speakers of Hiberno-English. The perception test was conducted according to the procedure described in [3] and [4] as a series of six subtests. In each subtest, 10 randomisations of the 11 stimuli were presented to the participants, and responses were obtained for a pair of opposite affective attributes. The pairs of affective attributes tested were *sad-happy*, *intimate-formal*, *relaxed-stressed*, *bored-interested*, *apologetic-indignant*, and *fearless-scared*.

The participants were asked to judge for each stimulus whether the speaker sounded more sad or happy, etc., and to mark their response on the answer sheet. The ratings were interpreted as a seven point scale ranging from -3 to + 3, where 0 corresponded to no affect perceived, and plus or minus 1, 2 or 3 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 voice quality stimuli. The significance level was set at $p < .05$. The intraclass correlation

coefficient was calculated to test interrater agreement and consistency in voice quality-to-affect association.

5. Results and discussion

Fig. 1 shows for each of the two stimulus series ('voice quality' stimuli in grey, 'loudness' stimuli in black), the maximum mean rating obtained for each of the affects tested, i.e. the mean rating for the most highly scored stimulus within each series.

It is clear from Fig. 1 that for the majority of the affects tested, the 'voice quality' series generated much higher affective ratings than the 'loudness' series. For eight affects out of 12, the ratings yielded by the former were markedly higher than the latter. In the remaining four affects *formal*, *interested*, *happy* and *fearless*, differences were smaller and not statistically significant. Note for *happy* and *fearless* the ratings were low regardless of stimulus type.

In the 'loudness' series, only for 4 of the 12 affects (*formal, stressed*, *interested* and *indignant*) was the maximum mean rating 1 or higher. In contrast, maximum mean ratings for the 'voice quality' stimuli were above 1 with a single exception, *fearless*.

These results support our hypothesis that the contribution of loudness is not the main determinant of the affective colouring which different voice qualities impart. This is not to say that loudness has no role to play: for certain affects – particularly *formal* – loudness alone yields a fairly high rating. It may well be that for these specific affects loudness is an important cueing factor, and it is worth noting that for *formal* the addition of tense voice quality did not yield higher affective ratings.

Furthermore, it is worth noting that the results of the present study are in keeping with those obtained for Hiberno-English speakers in [4], and support the broad findings reported with regard to voice quality-to-affect mapping.

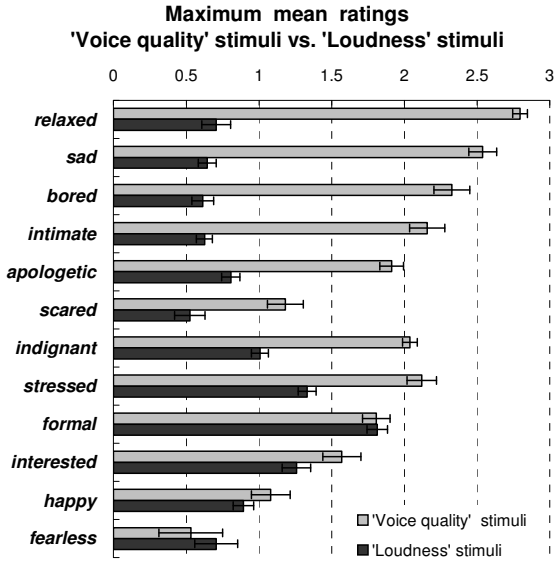

Figure 1. *Maximum mean rating and estimated standard error of the mean. Affect ratings: 0 = none, 3 = max.*

Table 2 indicates which of the stimuli in either series was most potent in cueing the individual affects. In the case of the 'voice quality' series, each of the voice qualities tested yielded highest ratings for at least one affective state. In the case of the 'loudness' stimuli, the highest mean ratings were always associated with one of two stimuli: the L_whispery and the L_tense. As can be ascertained from Table 1, these were at the extreme ends of the loudness continuum, with L_whispery being the quietest, and L_tense being the loudest. And in the case of the former stimulus, L_whispery, it is the most highly rated for the affects *relaxed*, *sad*, *bored*, *intimate*, *apologetic* and *scared*, i.e., the top six affects shown in Fig. 1 above. Note, however, that the actual rating is in every case very low. For these affects, the voice quality appears to be of crucial importance and the difference between the loudness and voice quality stimuli is very pronounced. In the case of the L_tense stimulus (modal voice with the loudness of the original tense voice), the loudness level as such does appear to have some affect cueing function, particularly for *formal, stressed* and *interested*. For these affects, the contrast between the two types of stimulus is less dramatic. Although not surprising in itself, it is worth noting that the L_tense stimulus is in all cases associated with affects that have high activation. It would seem therefore that loudness as such (without a necessary contribution of voice quality) does contribute to the perception of high activation. The converse is not true: L_whispery (the quietest stimulus) does not appear to contribute much to the perception of low activation states, where voice quality appears to be essential.

Table 2. *Stimuli in both series yielding the highest rating for each affect.*

Affect	'Voice quality' stimuli	'Loudness' stimuli
relaxed, sad, bored, intimate	lax-creaky	L_whispery
apologetic	breathy	L_whispery
scared	whispery	L_whispery
indignant	harsh	L_tense
stressed, formal, interested, happy, fearless	tense	L_tense

Interrater agreement. Intraclass correlation coefficients (ICC) were calculated to show listeners' agreement in the stimulus-to-affect association. Table 3 presents a broad indication of the results, for each stimulus in each test. Results for the voice quality stimuli are grouped in the upper part of the table, the loudness stimuli – in the lower part. For each stimulus/test, a capital letter in bold type in a particular cell indicates an affect for which a high degree of agreement was found – an ICC > 0.8. The choice of letter in this cell indicates which of the pair of affects was perceived. In the case where the actual mean rating was rather low (i.e., below 1 on the rating scale in Fig. 1) the letters are shown in brackets.

Table 3 demonstrates that listeners are more consistent in their responses when rating the affect-strength of 'voice quality' stimuli than when rating 'loudness' stimuli. Lax-creaky voice, harsh voice and tense voice are the voice qualities that demonstrate the best interrater agreement, while the 'loudness' stimuli based on modal voice, and the original modal voice stimulus are characterised by relatively low ICC. As can be seen from the density of letters in the cells in the upper part of the table, it is clear that there is good interrater agreement for the voice quality stimuli in how they are mapped to affect. For each of the voice qualities tested, there seems to be a consistent mapping to one or more affects.

Table 3. *Voice quality stimulus-to-affect association, only the stimuli with ICC ≥ 0.8 are shown; affects yielding average rating < 1 are shown in brackets.*

Test / Stimuli	apologetic-indignant	bored-interested	fearless-scared	intimate-formal	relaxed-stressed	sad-happy
whispery			**S**		**R**	
breathy			(S)		**R**	
lax-creaky	**A**	**B**	(S)	**I**	**R**	**S**
harsh	**I**	**I**	(F)	**F**	**S**	(H)
tense			(F)	**F**	**S**	**H**
modal						
L_whispery			(S)		(R)	
L_breathy						
L_lax-creaky						
L_harsh						
L_tense			(S)			

6. Summary and conclusions

Overall, the results show that loudness variation on its own is relatively ineffective for affect cueing. Stimuli incorporating voice quality variations yield relatively high maximum mean ratings: only for one affect, *fearless*, was the rating below 1, whereas for the 'loudness' stimuli, the ratings were below 1 for eight out of the 12 affects tested.

Apart from getting lower ratings, the 'loudness' stimuli also produce a significantly lower degree of agreement in voice quality-to-affect association.

The results also show, however, that loudness level (in the absence of voice quality variation) is not entirely irrelevant to affect cueing. High loudness levels it can play a role in the cueing of high activation states, particularly *formal*, *stressed* and *interested*. On the other hand, for low activation states such as *relaxed, sad, bored, intimate* and *apologetic*, a reduction in the loudness level (in the absence of voice quality variation) has little effect.

7. Acknowledgements

This research is supported by the EU-funded Network of Excellence on Emotion, HUMAINE.

8. References

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

[2] Frick, R. W., 1985. Communicating emotion: the role of prosodic features. *Psychological Bulletin* 97 (3), 412-429.

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

[4] Yanushevskaya, I.; Gobl, C.; Ní Chasaide, A., 2005. Voice quality and f_0 cues for affect expression: implications for synthesis. *Proceedings of the 8th International Conference on Spoken Language Processing, INTERSPEECH 2005*. Lisbon, 1849-1852.

[5] 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.

[6] Fant, G.; Liljencrants, J.; Lin, Q., 1985. A four-parameter model of glottal flow. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 4, 1-13.

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

[8] Scharf, B., 1978. *Loudness. Handbook of Perception*. Vol. IV Hearing. New York: Academic Press.

[9] Sluijter, A.M.C.; van Heuven, V.J.; Pacilly, J.J.A., 1997. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America* 101 (1), 503-513.