INTERSPEECH 2009 BRIGHTON

Perceived Loudness and Voice Quality in Affect Cueing

Irena Yanushevskaya, Christer Gobl, Ailbhe Ní Chasaide

Centre for Language and Communication Studies, Trinity College Dublin, Ireland

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

Abstract

The paper describes an auditory experiment aimed at testing whether the intrinsic loudness of a stimulus with a given voice quality influences the way in which it signals affect. Synthesised voice quality stimuli in which intrinsic loudness was systematically manipulated were presented to listeners to test the effect of this manipulation on the affective colouring of the stimuli. The results showed that even when devoid of intrinsic loudness variation, non-modal voice quality stimuli were capable of communicating affect. However, changing the loudness of a non-modal voice quality stimulus towards its intrinsic loudness resulted in the increase of affective ratings.

Index Terms: perceived loudness, voice quality, emotion and affect.

1. Introduction

Loudness is defined as the subjective auditory sensation of the magnitude of the sound. The subjective nature of perceived loudness makes its measurement extremely challenging. Assumptions of perceived loudness as subjective auditory sensation have to be based on the results of listening tests using psychoacoustic procedures such as magnitude estimation and magnitude production. Objective methods of estimation of perceived loudness include the use of loudness models and loudness meters. Perceived loudness is not equivalent to, but is most closely related to the intensity of the signal. Other properties of the signal, such as its spectral characteristics and bandwidth, its duration and the conditions in which the sound is presented to the listener (e.g., monaural or binaural, with or without background noise) as well as the characteristics intrinsic to the listener, influence the signal's perceived loudness [1-3]

In speech research, loudness has been studied primarily as a perceptual correlate of syllable prominence and stress through acoustic measures related to overall intensity of the speech signal, spectral properties of the signal (spectral slope, spectral balance or spectral emphasis) as well as through the studies of vocal effort [4-6]. In emotional speech research, loudness, together with other prosodic features, has been considered an important cue in communicating high activation affects, such as anger and elation [7, 8].

The term loudness has been used somewhat loosely across studies and perceived loudness has not infrequently been identified with intensity or explained through the changes in spectral balance of the speech signal. It is well established that increased vocal effort leads to a decrease in the spectral slope as measured by LTAS [9], but it does not necessarily imply that spectral slope is the sole feature responsible for the increase in perceived loudness. Intrinsic perceived loudness of different voice qualities is related to the shape of the glottal pulse and therefore to the spectral slope. Conversely, changing the spectral slope of the voice source will inevitably entail changes in voice quality. Although there is a tendency for vocal effort, voice quality and perceived loudness to co-vary, it is not impossible to produce essentially the same voice quality at different loudness levels (e.g., 'ordinary' and stage whisper or soft but 'angry' tense voice).

The paper follows on the studies reported in [10] and [11]. In [10], an utterance synthesised with different voice qualities was tested to explore the extent to which the different voice qualities might alter its affective colouring. In changing the glottal pulse shape to synthesise the different voice qualities, the loudness is concomitantly altered. Thus while the experiment in [10] reported distinct affective associations with particular voice qualities, the question arose as to whether this might partially be due to the intrinsic loudness differences in the stimuli. In [11], the hypothesis was tested, using synthetic voice quality stimuli, that affective cueing is *not* simply a consequence of the loudness variation in these stimuli. Two series of synthetic stimuli were used to elicit affective ratings for a number of pairs of affective labels: (1) the 'Voice quality' stimuli, that incorporated distinct voice quality features including intrinsic loudness variations and (2) the 'Loudness' stimuli, where voice quality (modal voice) was kept constant, but in which loudness was systematically modified to match the intrinsic loudness of the original voice quality stimuli. The loudness matching stimuli were obtained in the course of a preliminary auditory test in which the listeners selected the sounds best matching the loudness of the original voice quality from a range of 25 modal stimuli differing by relatively fine 1 dB steps. The results suggested that loudness variation on its own is relatively ineffective for affect cueing as the 'Voice quality' stimuli consistently yielded higher affective ratings compared to the 'Loudness' ones, although high loudness level (in the absence of voice quality variation) appeared to play a role in the cueing of high activation states (e.g., *stressed*).

The above study demonstrated that loudness variation *alone* is rather ineffective in cueing affect, and clearly by manipulating loudness we cannot generate affectively coloured output. However, this does not necessarily mean that the intrinsic loudness differences which tend to be correlated with particular voice qualities are not playing an important role. Given that in human speech production there is a natural tendency towards co-variation of voice quality and loudness, we hypothesised that loudness differences do play an important role – but only when these loudness variations occur with the appropriate voice quality. In a further test of this issue, the present experiment tested the hypothesis that equalising the perceived loudness of voice quality stimuli while maintaining intrinsic voice quality variations will have relatively little impact on affective ratings. New stimuli were generated (see description below) in which the loudness of all stimuli was set to that of the original modal voice stimulus, but in which the inherent voice quality characteristics were maintained. The stimuli with voice quality differences were firstly normalised to the loudness level of the original modal voice (Series M), and two further series were generated from this set: Series L involved a simple amplification by 2 dB of all stimuli, and Series Q – attenuation by 2 dB of all stimuli.

10.21437/Interspeech.2009-479

Note that for each series we had a range of stimuli differing in voice quality but with normalised loudness levels.

2. Method

2.1. Synthesised stimuli

The stimuli used in this experiment were based on the original voice quality stimuli used in a number of earlier experiments, e.g., [10]. The original voice quality stimuli (modal, whispery, breathy, lax-creaky, and tense voice) were generated based on a high quality copy synthesis of the Swedish utterance "ja adjö [ˈjɑː aˈjøː] using the KLSYN88a formant synthesiser [12] which incorporates the modified version of the LF voice source model [13]. 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. For detailed description of the stimuli, see [10].

The new set of stimuli for the experiment consisted of three series. The first series (Series M) was generated by attenuation or enhancement of the loudness level of all original non-modal voice quality stimuli to match that of the modal voice. The amplitude of each of the original voice quality signal was multiplied by the reciprocal of the corresponding scaling factors used in [11] resulting in a uniform gain or decrease in overall intensity level. The scaling factors used to synthesise the voice quality stimuli used to obtain the necessary intensity increase/attenuation in dB are presented in Table 2.1. [Scaling factor = $10^{(-7.27/20)}$ where -7.27 dB difference is required. Difference in dB = 20 $\log_{10}$ (I_1/I_2)]. The resulting non-modal voice quality stimuli all had equal loudness, that of the modal voice. This allowed us to control for loudness in the following auditory experiments.

Table 2.1 *Scaling factors used to generate voice quality stimuli of Series M with the loudness matching that of the modal voice.*

Series M stimuli	Scaling factor	Difference in dB relative to the original voice quality stimuli
whispery	2.309	+7.27
breathy	1.592	+ 4.04
lax-creaky	1.376	+2.77
modal	1	0
tense	0.701	-3.09

Subsequently, from the stimuli in Series M (in which the voice quality was kept distinct but the loudness was set to match that of the modal voice) two more series were generated. In Series L ('louder' versions) all stimuli were amplified by 2 dB and in Series Q ('quieter' versions) all stimuli were attenuated by 2 dB. Thus, in each series there were stimuli with different voice quality but with normalised loudness level.

Overall, 15 synthesised stimuli were used in the present experiment falling into three groups: 1) Series M: non-modal voice quality stimuli whose intrinsic loudness was set to that of the modal voice stimulus, 2) Series Q: their 'quieter' versions, and 3) Series L: their 'louder' versions.

2.2. Listening test

The 15 stimuli were presented to 16 native speakers of Hiberno-English in a series of perception tests according to the procedure described in [10]. In each test, the 15 stimuli were presented to the participants in randomised order; responses were obtained for a pair of opposite affective attributes *apologetic-indignant, bored-interested, fearless-scared*, *intimate-formal, relaxed-stressed,* and *sad-happy*. The subjects were asked to judge the affective colouring of each stimulus and to mark their response on the answer sheet. The ratings were interpreted as a scale ranging from -3 to + 3, where 0 corresponded to no affect perceived, and plus/minus 1, 2 or 3 to mild, moderate and strong presence of an affect respectively.

A two-way repeated measures ANOVA with two within-subject factors: 'voice quality' (5 levels) and 'loudness' (3 levels) was conducted for each of the six scales separately; the alpha level was set to 0.05. Significant main effects were found for both voice quality and loudness level, as well as for the two-way interaction for all scales except *fearless-scared* (here the voice quality and the two-way interaction effects were not statistically significant). The intraclass correlation coefficient and percent agreement were used to measure the raters' consistency in voice quality-to-affect association.

3. Results and discussion

The mean ratings obtained for the stimuli tested in this experiment for each pair of affective labels are shown in Figure 3.1; the stimuli which yielded the highest mean ratings for a particular affect are marked with stars. As evident from Figure 3.1, three voice qualities, tense, lax-creaky and whispery, no matter what level their loudness was set to, emerged as the most potent for all affect signalling. They were associated with clusters of affects that are summarised in Figure 3.2.

The ratings of Series M plotted in Figure 3.1 as grey bars show the effect of loudness normalisation (increase or attenuation of the intrinsic non-modal voice quality loudness to that of the modal voice) on voice-to-affect association. Note that intrinsic loudness level of tense was lowered by 3 dB, the loudness level of lax-creaky was increased by 2.8 dB and the loudness level of whispery voice was increased by about 7 dB (Table 2.1). Grey bars in Figure 3.1 show what voice quality can achieve when devoid of intrinsic loudness. It is clear that even with loudness normalisation, non-modal voice qualities are still effective in affect cueing as each is associated with at least one affect with the rating above 1 (which we interpret here as mild to moderate affective colouring). Thus, whispery voice from Series M gets associated with *apologetic* and *intimate*, breathy voice – with *apologetic*; lax-creaky – with *apologetic*, *bored*, *intimate*, *relaxed* and *sad*; and tense voice – with *indignant, interested*, *formal, stressed* and *happy*. Only two affects failed to be rendered by these stimuli: *fearless* and *scared*. The ratings of the modal voice are particularly conspicuous in this respect compared to the non-modal voice qualities as in no case did it achieve the rating above 1. As the original voice quality stimuli with intrinsic loudness differences were not used in the present experiment, it is impossible to directly compare them and the loudness normalised stimuli and, although informative, the broad indication of voice to affect association is only partially useful. Nonetheless, it is clear that in [11] and the present experiment

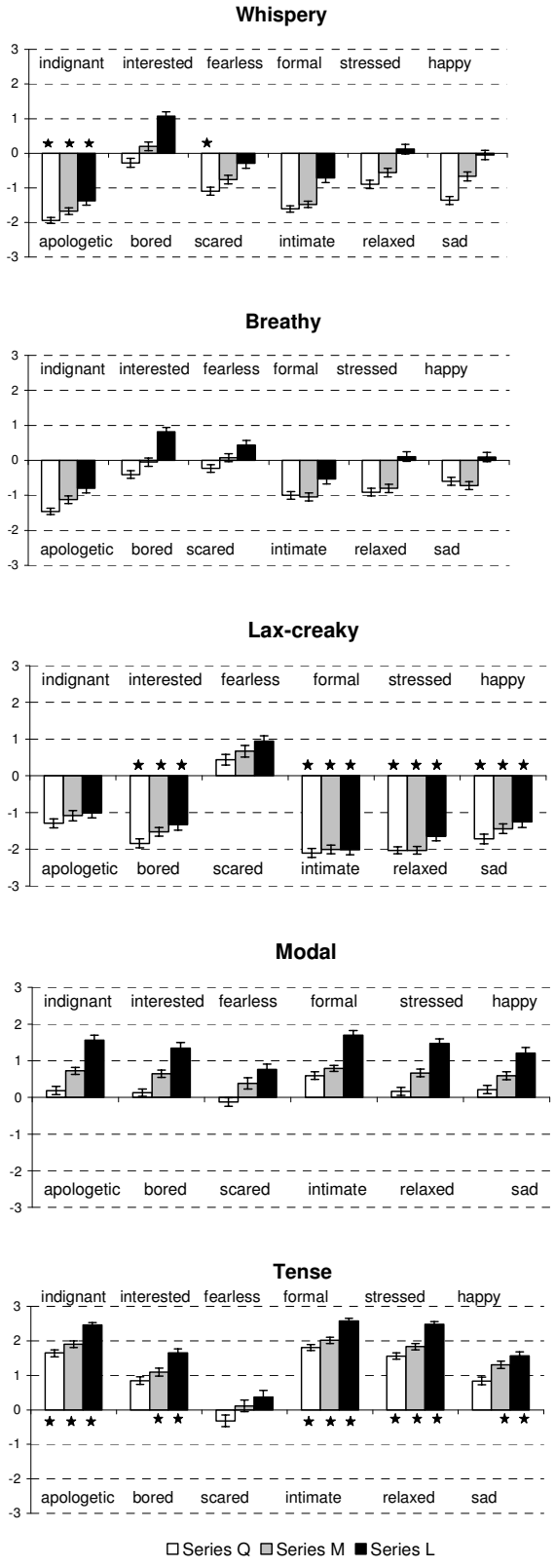

Figure 3.1 *Ratings obtained by the three groups of stimuli: Series M (grey bars), Series Q (white bars) and Series L (black bars). Stars show the stimuli that obtained maximum mean rating for a particular affect; maximum ratings below 1 (=no affective colouring perceived) are not marked.*

the listeners associated the same affect with the same voice quality stimuli, although there is obviously a reduced range of affective ratings where the loudness-normalised stimuli are concerned. It could be tentatively concluded that although non-modal voice qualities are still potent in affect signalling, changing their intrinsic loudness level to that of modal voice does influence their potential in communicating affect as the ratings are somewhat lower.

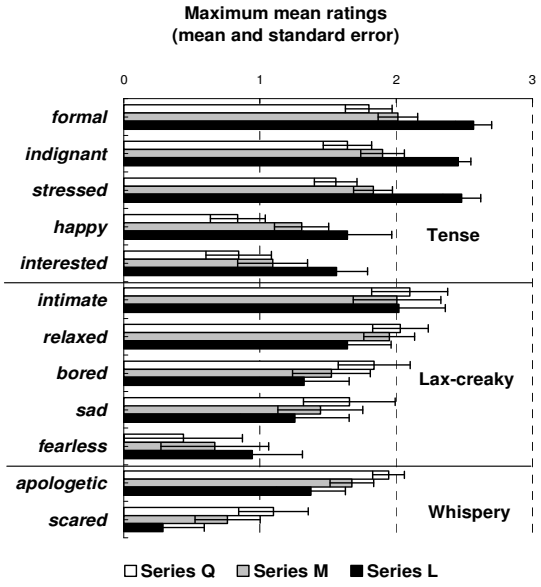

Figure 3.2 *Maximum mean ratings of the voice quality stimuli. Series Q (white bars), Series M (grey bars), and Series L (black bars). Affect rating: 0=none; 3=max.*

Comparison of the ratings of Series Q, M and L in Figure 3.1 shows that the listeners are sensitive to the increase or decrease of loudness in voice quality stimuli. The effect of loudness attenuation (Series Q, white bars in Figure 3.1) is as follows: 1) for whispery, breathy, lax-creaky voice qualities that are mostly associated with low activation affects, decreasing loudness level results in an increase in ratings; 2) for tense and modal voice qualities that are associated mostly with high activation states, the opposite is true – decrease of the loudness level entails decrease in affective ratings. Loudness enhancement (Series L, black bars in Figure 3.1) shows the opposite for the majority of affects: with the loudness level increased by 2 dB, tense voice quality receives significantly higher ratings for high activation affects whereas the ratings of lax-creaky and whispery voice decrease. (The only exception here appears to be the *intimate* affect for which the increase in loudness level of the lax-creaky voice resulted in virtually no change in affective ratings.) This is in a way a non-surprising finding as enhancing the loudness of tense voice from Series M by 2 dB or attenuating the loudness of lax-creaky or breathy voice from Series M by 2 dB essentially brings the loudness level of these stimuli closer to their original intrinsic loudness. The results demonstrate that the closer the stimulus loudness gets to the intrinsic loudness of the original voice quality, the higher the affective ratings, and therefore the more successful affect signalling. An increase or decrease of loudness level will only result in the increase of affective ratings for certain voice qualities. On the other hand, when loudness level is not set to extreme values but to that of

modal voice, voice quality alone proves to be sufficient for successful affect cueing.

The data on interrater agreement are summarised in Table 3.1. For each stimulus/test, a capital letter in a particular cell indicates an affect for which a high degree of agreement was found (an ICC ≥ 0.8 and percent agreement of 75% or more). The choice of letter in this cell indicates which of the pair of affects was perceived.

Table 3.1 *Interrater agreement of voice to affect association: shown are the cases with ICC ≥ 0.80 and percent agreement of 75% or more.*

Stimuli \ Test		apologetic-indignant	bored-interested	fearless-scared	intimate-formal	relaxed-stressed	sad-happy
Series Q	whispery			S			
	breathy	A					
	lax-creaky	A	B		I	R	S
	modal						
	tense				F	S	
Series M	whispery				I		
	breathy				I		
	lax-creaky	A	B		I	R	S
	modal						
	tense		I				H
Series L	whispery	A	I				
	breathy		I				
	lax-creaky	A	B		I	R	S
	modal						
	tense		I			S	H

Table 3.1 shows that lax-creaky voice yields the highest agreement across all subtests compared to other voice qualities. It is confidently associated with *apologetic, bored, intimate, relaxed* and *sad*, irrespective of its loudness settings. In comparison, breathy and whispery voice qualities show lower interrater agreement, but they are associated with at least one affect tested in the 6 subtests. Tense voice was associated with *interested, formal, stressed* and *happy*, and the agreement varied somewhat for different affects. Note that none of the modal voice loudness variants was associated with any particular affect.

4. Conclusions

The results point towards a conclusion that loudness does contribute positively in affect cueing. This can be seen in terms of the ratings given the three loudness levels (Series Q, M, and L). Furthermore, Figures 3.1 and 3.2 do suggest that increased loudness with high activation states may play a fairly important role. It is particularly striking in the ratings of the modal voice, as increasing its loudness by 2 dB results in significant increase in affective ratings for such affects as *indignant*, *interested*, *formal*, *stressed* and *happy*. Here loudness alone appears to achieve affective colouring.

In view of these results and the results reported in [11] we should conclude that whereas loudness on its own does little to cue affect, but when combined with appropriate voice quality it may be important in the signalling of high activation states. Furthermore, when the loudness approximates the intrinsic loudness of a particular voice quality, it is, perhaps unsurprisingly, at its most effective. The results further suggest that even without intrinsic loudness variation, voice quality stimuli, and in particular whispery, lax-creaky and tense, prove effective in affect cueing.

The study supports the hypothesis suggested in [14] that manipulating loudness of a synthesised stimulus while keeping voice quality constant should have a less prominent impact on the stimulus perception than varying the voice quality and keeping absolute loudness unchanged. Loudness appears to work in affect cueing in conjunction with intrinsic voice quality variations.

5. Acknowledgements

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

6. References

[1] B. Scharf, "Loudness," in *Handbook of Perception*. vol. 4. Hearing, E. C. Carterette and M. P. Friedman, Eds. New York: Academic Press, 1978, pp. 187-242.

[2] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London: Academic Press, 2003.

[3] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2 ed. Berlin: Springer-Verlag, 1999.

[4] M. Heldner, "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish," *Journal of Phonetics,* vol. 31, pp. 39-62, 2003.

[5] A. M. C. Sluijter, V. J. van Heuven, and J. J. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *The Journal of the Acoustical Society of America,* vol. 101, pp. 503-513, 1997.

[6] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America,* vol. 107, pp. 3438-3451, 2000.

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

[8] R. W. Frick, "Communicating emotion: the role of prosodic features," *Psychological Bulletin,* vol. 97, pp. 412-429, 1985.

[9] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *The Journal of the Acoustical Society of America,* vol. 120, pp. 453-457, 2006.

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

[11] I. Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Voice quality and loudness in affect perception," in *Speech Prosody 2008*, Campinas, Brazil, 2008.

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

[13] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR* vol. 4, pp. 1-13, 1985.

[14] M. Schröder, "Speech and Emotion Research: an Overview of Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis," *Phonus 7. Research Report of the Institute of Phonetics, Saarland University,* 2004.