Voice Quality Variation and the Perception of Affect: Continuous or Categorical?

Colin Ryan, Ailbhe Ní Chasaide and Christer Gobl
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
E-mail: cryan10@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

ABSTRACT

This paper explores the mapping of voice quality to affect, for a synthesised tense – lax voice continuum. Two questions are of interest. Firstly, over such a continuum, do listeners' attributions of affect change in a continuous or in a more categorical way? Secondly, might discreetly different affects emerge at different points in the continuum: specifically, might a moderately tense voice cue *happy*? Note that extremely tense does not appear to, but seems rather to be associated with the affect *angry*. A continuum of stimuli were synthesised ranging from a very tense quality at one end to a very lax quality at the other. Listeners rated their affective colouring in terms of the pairs of affective attributes: stressed/relaxed, angry/content, happy/sad, interested/bored, formal/intimate, and indignant/apologetic. Results suggest that, for these voice qualities, listeners' attribution of affect is essentially continuous. Furthermore, moderate tension does not evoke *happy.*

1 INTRODUCTION

It is widely assumed that paralinguistic communication involves continuous, gradient variation of phonetic parameters, whereas linguistic meaning relies more heavily on categorical decisions. The present study looks at the nature of the mapping of voice quality to affect, building on a study by Gobl and Ní Chasaide [1], where listeners judged the affective colouring of an utterance synthesised with a variety of different voice qualities. The strongest ratings for most of the affective attributes emerged for just two of the qualities tested. On the one hand tense voice yielded high ratings for affects such as *angry*, *stressed*, *formal*, which have high activation and power. On the other hand, lax-creaky voice yielded high ratings for low activation affects such as *relaxed*, *intimate* and *bored.*

The principal aim here was to see whether a continuum of stimuli ranging from tense to lax voice maps to affect in a continuous or categorical fashion. For example, do listeners detect increasing degrees of anger when stimuli deviate from modal voice towards tense voice in a stepwise fashion? Or might the association of anger depend on tense voice passing a specific threshold? In this case, ratings for *anger* should be more categorical. Our initial hypothesis is that the mapping is a continuous, gradual one.

In the original experiment reported in [1], tense voice was also associated more weakly with the attribute *happy*. The possibility that both *anger* and *happiness* are associated with tense voice has been suggested by Scherer [2]. It is conceivable that a moderately tense voice might be more potent in evoking the *happy* affect, whereas a very tense voice might be counterproductive and serve to undermine it. The second hypothesis here is therefore that in the modal-to-tense part of the continuum, ratings for the affect *happy* will be higher when the quality is moderately tense than when it is extremely tense.

2 CONSTRUCTION OF STIMULI

The stimuli used in this study were an adaptation of previously generated stimuli, used in [1]. The aim was to produce a continuum of voice qualities ranging from very tense to very lax voice. Although in the earlier study, other vocal factors such as the addition of creakiness to lax voice were found to enhance the affective colouring, such additional factors were not included in the present study.

2.1 The modal stimulus

The Klatt formant synthesiser KLSYN88a [3] was used to produce the stimuli, with the modified LF source model [4] for generation of the glottal source signal. The starting point for the present continuum was the modal stimulus used in [1], with some minor changes to the AH and AV parameters of KLSYN88a. Only a brief outline of how this stimulus was generated is given here, for full details see [1].

The modal stimulus was based on a copy synthesis of a natural Swedish utterance 'ja adjö' [ˈjɑː aˈjøː], spoken by a male speaker having a voice quality approximating the modal voice quality of the Laver system [5]. Source and filter data were obtained using interactive inverse filtering and model matching software for the analysis of all 106 pulses of the utterance.

The variation in the source and filter parameters for the 106 pulses was stylised so as to capture the essential dynamic variation of each parameter. Depending on the parameter, between 7 and 15 timepoints were used in the resynthesis.

2.2 Generating the tense – lax continuum

In order to generate a continuum from very tense to very lax voice quality, 11 different stimuli were generated by manipulating the following parameters of synthesiser: AV, amplitude of voicing; TL, spectral tilt; OQ, open quotient; SQ, speed quotient; AH, aspiration noise; and B1, first formant bandwidth.

ISBN 1-876346-48-5

To generate equidistant steps, a strategy was adopted, which is here illustrated for the OQ parameter (see Table 1 and Figure 1). Firstly, for each parameter in turn, extreme values at the tense and lax ends were established. In the case of OQ, a lax limit of 100% and a tense limit of 30% were chosen. These extreme values were decided upon on the basis of suggestions in the KLSYN88a manual, as well as from knowledge about voice production constraints.

Time (ms)	Limit Lax	*L5*	*L4*	*L3*	*L2*	*L1*	*M*	*T1*	*T2*	*T3*	*T4*	*T5*	Limit Tense	Step size
0	*100*	100	97	94	91	88	85	82	79	76	73	70	*30*	3.0
100	*100*	100	96	92	88	84	80	76	72	68	64	60	*30*	4.0
180	*100*	90	84	78	72	66	60	54	48	42	36	30	*30*	6.0
600	*100*	90	84	78	72	66	60	54	48	42	36	30	*30*	6.0
665	*100*	64	61	57	54	50	47	44	40	37	33	30	*30*	3.4
750	*100*	100	94	88	82	76	70	64	58	52	46	40	*30*	6.0
945	*100*	60	57	54	51	48	45	42	39	36	33	30	*30*	3.0
1025	*100*	100	94	88	82	76	70	64	58	52	46	40	*30*	6.0
1195	*100*	100	94	88	82	76	70	64	58	52	46	40	*30*	6.0

Table 1: OQ values, OQ extreme limits and the calculated interstimulus stepsize at each timepoint in the utterance (values are in %).

At timepoint 0 in Table 1, the OQ value for modal voice is 85%. Thus, OQ may vary by 15 percentage points in the tense direction or by 55 percentage points in the lax. Five equidistant steps would therefore yield steps of 3 percentage points in the tense direction or of 11 percentage points in the lax direction. In order to have equidistant steps between stimuli, the smaller stepsize was always chosen to prevent values from exceeding the limits. This stepsize was recalculated for each timepoint, as illustrated in Table 1. Values for TL and SQ were obtained in the same fashion.

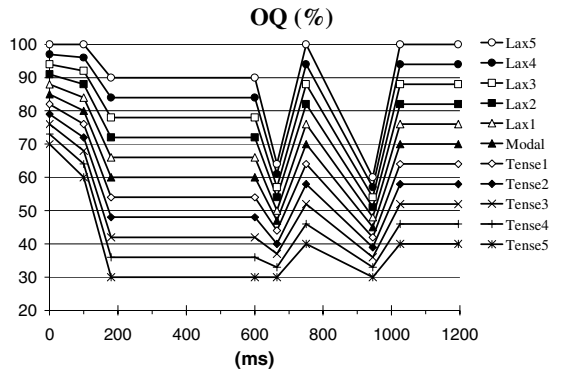

Figure 1: OQ variation in the utterance for the 11 stimuli.

Calculations of changes in the B1 parameter follow essentially the same procedure as for OQ. The difference is that the change does not involve a fixed stepsize, but rather a constant multiplication factor. This is motivated by the fact that a constant change in the amplitude level of a formant corresponds more closely to a relative change in the formant bandwidth.

For AH, the original modal stimulus was not taken as the starting point, given that AH was set to zero throughout. Rather, AH values were taken from the breathy voice stimulus of the experiment in [1] as the appropriate setting for the Lax3 stimulus. Using Lax3 as a staring point, these AH values were modified for the other stimuli in 2 dB steps, decreasing with increasing degree of tenseness and increasing with increasing degree of laxness.

In the case of the AV parameter, values varied only for vowel onsets and offsets, so as to provide for sharper onsets and offset for the tense stimuli as compared to more gradual onsets and offsets for the lax stimuli. Fundamental frequency variation in the modal stimulus was retained across the continuum.

Manipulations of the SQ parameter in KLSYN88a cause changes in the excitation amplitude (EE) of the glottal pulse generated by the modified LF model of KLSYN88a. Whereas a level difference across a tense-lax continuum of voice qualities would be expected, the AV values were assumed to map closely to variation in the excitation strength (i.e. to the EE values). The indirect changes in EE due to variation in SQ were deemed undesirable and were compensated for as follows.

The overall change to the amplitude level of the stimuli was first measured to be 16.0 dB. Then the changes in EE as a function of the range of SQ variation used here was estimated, by analysing a synthesised utterance with constant AV, but varying SQ. This variation was found to be 8.8 dB. Finally, the overall amplitude of the individual stimuli was adjusted to cancel the variation in EE as function of SQ. The level of the modal stimulus was taken as a reference and was not altered. For a given stimulus, e.g. Lax5, the level difference from modal was calculated. The proportion of this difference (8.8/16) was multiplied to the overall level difference to yield the compensation (in dB) that would be required to remove the influence of SQ on EE.

3 PERCEPTION EXPERIMENTS

3.1 Experiment I: Can the stimuli be discriminated?

Before proceeding to a direct test of affect mapping for these stimuli, a preliminary test was carried out to check whether stimuli (separated by one or two steps in the continuum) could be discriminated, as it would be pointless to have subjects rating affect differences for stimuli which were not in themselves judged to be auditorily different. A further aim was to reduce the number of stimuli to a more manageable number, in order to avoid an overlong test.

An AX design was used, whereby subjects were presented with pairs of stimuli in sequence, and judged simply whether they were the same or different. Each pair of stimuli differed by either one or two steps of the continuum. This amounted to a total of 19 pairs (10 one-step and 9 two-step differences). All pairs were presented in two orders, AX and XA, giving a total of 38 pairs in the test. The inter-pair interval was 0.75 s and the response time was 3 s. An earcon preceded each pair. The test was presented in ten trials, each of which contained a different randomisation of 76 dyads (38 different pairs x 2).

ISBN 1-876346-48-5

The test was administered to five subjects, and the results are shown in Figure 2. These show an overall poor discrimination of the one-step differences, whereas the two-step differences were generally detected as different. It was therefore decided that testing of affective differences would only be carried out on stimuli that differed by at least two steps.

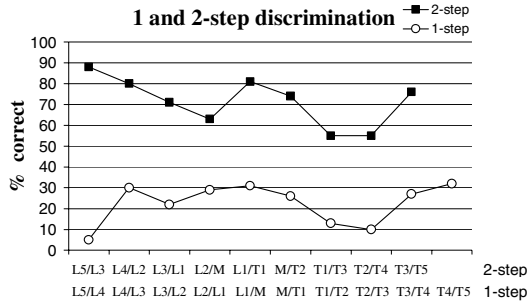

Figure 2: Percentage of stimulus pairs correctly judged to be different, involving one and two-step differences on the tense-lax continuum.

3.2 Experiment II: Mapping voice quality to affect

In this, the main experiment, subjects ratings on the possible affective colourings for a subset of stimuli were elicited. The stimuli for this test, chosen from the continuum described in Section 2 were: Lax5, Lax3, Modal, Tense3 and Tense5. The rationale for this selection was as follows. As we wished to include the extreme points as well as the midpoint of the continuum, the Lax5, Tense5 and Modal were initially selected. The decision to include Lax3 and Tense3 (rather than Lax2 and Tense2) was guided by the results of Experiment I. In Figure 1, one can see that the discrimination between Lax3 – Lax5 is considerably higher than that of Modal – Lax2. Thus, the auditory distinction between Modal and Lax2 is likely to be smaller, which would justify having the 'three-stimuli' step between Modal and Lax3, rather than between Lax5 and Lax2. On the tense side of the continuum, differences were very small, and the Tense3 stimulus was chosen primarily to mirror the selection made on the lax side of the continuum.

The perception test was administered as a series of eight mini-tests to 12 subjects, speakers of Irish English, following the procedure used in [1]. In each mini-test, 10 randomisations were presented, and responses were elicited for a pair of opposite affective attributes (e.g., *bored – interested*). Response sheets were arranged with the opposite terms placed on either side, with seven boxes in between, the central one of which was shaded in for visual prominence. Listeners were informed that they would hear a speaker repeat the same utterance in different ways and were asked to judge for each repetition whether the speaker sounded more *bored* or *interested*, etc. Subjects were instructed to chose the centre box if the utterance was deemed not to evoke either affect; ticking a box to the left or right of the central box should indicate the presence and strength to which a particular attribute was deemed present, with the most extreme ratings being furthest from the centre box. The full set of attribute pairs tested included *relaxed/stressed*, *content/angry*, *friendly/hostile*, *sad/happy*, *bored/interested*, *intimate/formal*, and *apologetic/indignant*.

4 RESULTS AND DISCUSSION

Results are presented in Figure 3 for each of the affect pairs that were tested. It is striking that for all attributes other than the *happy/sad* pair that the affect rating varies in a gradual fashion with the degree of *tenseness/laxness* of the stimulus. For the *happy/sad* pair, it is clear that there is little consistent association of affect with these stimuli.

These results provide support for the first hypothesis, that the mapping of tense/laxness to affect (when indeed there is an affect) is a gradual function rather than a categorical one. Thus, for example, ratings for *angry* were highest for the Tense5 stimulus, but even for the Tense3 stimulus there is a distinct trend in the *angry* direction. Insofar as one can tell from responses to the present set of stimuli, the detection of anger does not appear to depend on a certain threshold being achieved but maps in a gradual fashion to increasing degrees of anger.

One caveat here is that the stimuli are not all equidistant: in terms of source parameter settings, as the distance between Lax5 and Lax3 is lesser than the distance between Lax3 and Modal. Although it seems unlikely, there is a possibility that there could be a "hidden" discontinuity. A further test using the full continuum is currently underway.

The present results do not support the second hypothesis: moderate tenseness is not really any better for cueing happiness than is the higher degrees of tenseness used in the earlier experiments, and used here.

The striking "failure" of any of these stimuli to evoke a *happy* or *sad* affect reflects to some extent an uncertainty on the part of the subjects. There was a higher degree of intersubject variability, as can be seen by the higher standard deviation values for these mean rating values for *happy* and *sad*. This result does appear to indicate that tense or lax voice alone cannot evoke these emotions. It is interesting to note that the affect *happy* has been found to be particularly elusive, and less successfully captured in other production and perception studies [6,7].

Thus, on the face of it, Scherer's proposal linking tense voice and happiness is not supported here. Of course voice quality is only one of a constellation of features that are involved in signalling affect: other features such as pitch and amplitude dynamics, speaking tempo are also of major importance, see e.g., [2,6,7,8]. Furthermore in the case of *happy* speech some non-phonatory aspects of voice quality (lip spreading, with consequent acoustic changes such as F2 raising) are likely to be important, and visual cues (e.g., smiling) may be of great importance. Therefore, it may be premature to conclude that the tense/lax dimension is irrelevant to the detection of *happy/sad* emotions: we would suggest rather that if these voice qualities are employed, it is only in conjunction with other cues that they are likely to take effect.

ISBN 1-876346-48-5

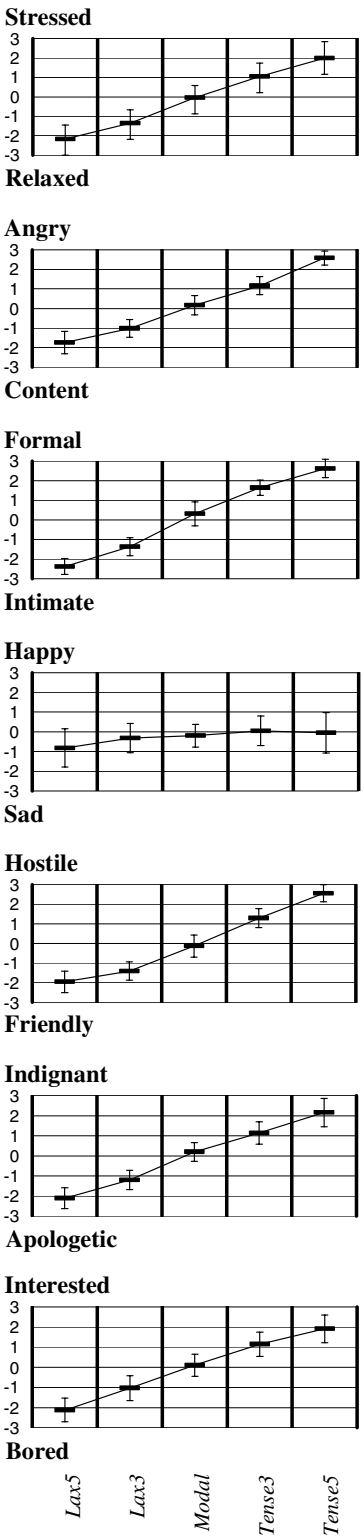

Figure 3: Mean ratings obtained for the subset of stimuli (vertical lines show one SD) for each pair of affective attributes (7-point scale, -3 to +3).

5 CONCLUSIONS

The results of the present experiment suggest that the mapping of affect to the tense to lax voice continuum is gradual rather than categorical. They also indicate that a moderate degree of vocal tension is not likely to be more effective (than extreme tension) in evoking a *happy* affect.

REFERENCES

[1] 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.

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

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

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

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

[6] S. Mozziconacci, "Speech variability and emotion: production and perception", Ph.D. thesis, Technische Universiteit Eindhoven, Eindhoven, 1998.

[7] R. Carlson, B. Granström and L. Nord, "Experiments with emotive speech, acted utterances and synthesized replicas", *Speech Communication*, vol. 2, pp. 347-355, 1992.

[8] I.R. Murray and J.L. Arnott, "Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *Journal of the Acoustical Society of America*, vol. 93, pp. 1097-1108, 1993.

ISBN 1-876346-48-5