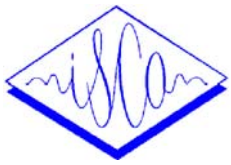
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Decomposing linguistic and affective components of phonatory quality

Ailbhe Ní Chasaide & Christer Gobl

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

Abstract

This paper is concerned with the role of phonatory quality in signalling affect. An overview of perception experiments is presented, which used synthetic stimuli with different phonatory qualities and f_0 contours in order to explore the mapping of voice quality to affect as well as the way in which voice quality combines with f_0. Results highlight the need for these phonetic parameters to be considered together. To identify the phonatory correlates of affect, we also need to understand the substrate of voice source variation, due to the linguistic content of utterances (prosodic and segmental) as well as to speaker specific characteristics. Illustrations of the former type of variation are presented, based on source parameterisation of inverse filtered data. A holistic analytic approach is advocated, which incorporates the main phonetic dimensions (voice quality, f_0 and temporal parameters) and which integrates the affective dimension with the more linguistic dimension of prosody.

1. Introduction

This paper focuses on the role of voice quality in affect communication. Although always acknowledged as being of crucial importance, there is little quantitative information available in comparison to other phonetic aspects, such as f_0 and amplitude dynamics, or temporal features, all of which have been extensively studied, e.g., [1].

There are two issues of fundamental importance in the modelling of the affective role of voice quality. Firstly, phonatory quality carries different kinds of information. In a single utterance, it serves not only to communicate affect, but also tells us about speaker characteristics and imparts crucial linguistic information. Section 3 focuses on the latter, illustrating how vital it may be to take account of it in attempting to capture the affective dimension. Secondly, voice quality and pitch are two dimensions of the voice source, which act together in conveying affective (as well as linguistic) information. To gain a true understanding of affect communication, we need to appreciate how these two phonetic dimensions interact. Section 2 presents an overview of perception experiments relevant to this question.

2. Voice quality and f_0 in affect perception

The experiments reviewed here explore the mapping of phonatory quality to affect, as well as the question of how phonatory quality and f_0 combine. The perception tests involved synthetic stimuli varying in voice quality, and the elicitation of listeners' ratings on whether and to what degree different affects were perceived. The synthesis was enabled by prior analyses of voice qualities, e.g., [2], following Laver's [3] descriptive classification. Inverse filtering was used to obtain the differentiated glottal flow, and the LF glottal flow model [4] was fitted to our data to obtain measures of source parameters. To produce stimuli with differing voice qualities, the LF implementation of the KLSYN88a synthesiser was used. For fuller details, see [5].

A first experiment described in [5] provides an initial exploration of whether phonatory quality shifts can alter the affective colouring of an utterance and whether such shifts coincide with traditional claims by phoneticians (e.g., creaky voice signalling boredom for speakers of English). The Swedish utterance "ja adjö" ['jɑː a'jøː] was synthesised with seven voice qualities: modal, breathy, whispery, tense, harsh, creaky and lax-creaky voice. While the first six of these conform to the Laver descriptions, lax-creaky voice is an extension to the system. Source manipulations within the utterance included dynamic variations aimed at capturing linguistic (prosodic and segmental) effects discussed in Section 3 below.

In this, as in the subsequent experiments discussed here, listeners rated the stimuli in a series of tests for pairs of opposite affective attributes, e.g., bored/interested. A seven-point scale was used: the midpoint was taken to indicate "no affect", and the three steps either side showed the degree to which the speaker was deemed to sound bored or interested.

Fig. 1 presents the ratings obtained for each attribute pair. Results show that these kinds of phonatory changes do alter the affective colouring. By and large, findings were compatible with the traditional observations about voice quality, but suggest some refinements. For example, creaky voice yielded lower ratings for boredom than the lax-creaky quality. And although in the past a one-to-one mapping of voice quality to affect has been implied, these results suggest rather that a particular quality is associated with a number of (often related) affective attributes, e.g., lax-creaky voice is highly rated for the attributes *relaxed*, *content*, *bored* and *intimate*.

The tense-lax dimension of voice quality emerged as an important dimension. Tense voice yielded high ratings for states involving high activation, while lax qualities (particularly the lax-creaky) were associated with low activation.

With the exception of *angry*, strong emotions (*happy*, *sad*, *afraid*) yielded relatively low ratings compared to the milder states of being (*confident*, *bored*, *intimate*, etc.).

Given that voice quality varies in a continuous fashion, these results beg the question as to whether the affective attribute(s) associated with an individual phonatory quality varies in a gradient way: e.g., does the rating for anger increase with increasing phonatory tenseness? Conceivably, the relationship between strength of phonatory quality and degree of affect is not continuous. For example, the detection of anger could depend on a certain threshold of tenseness being passed. It is also possible that, in a tense-lax continuum, discreetly different affects might emerge, with say, a moderate degree of tenseness signalling happiness while more extreme tenseness might signal anger.

These questions were explored in a second experiment where ratings for similar groups of affective attributes were obtained for five stimuli ranging from tense to lax. Details of the experiment and of the construction of the stimuli are provided in [6].

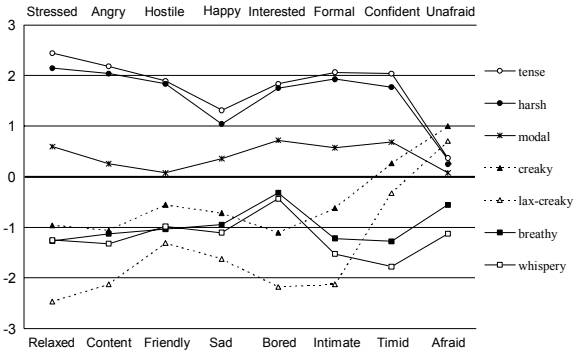


Figure 1: *Mean ratings of perceived affective strength of pairs of attributes for seven voice qualities. 0 = no affective content +/-3 = maximally perceived.*

Fig. 2 illustrates results for two of the affect pairs tested. For the pair *angry*/*content* (upper panel) the strength of the rating varies in a continuous fashion across the five stimuli. This was the pattern that emerged for all the other affect-pairs tested, with the exception of the *happy*/*sad* pair (lower panel), which yielded no consistent differentiation for these stimuli.

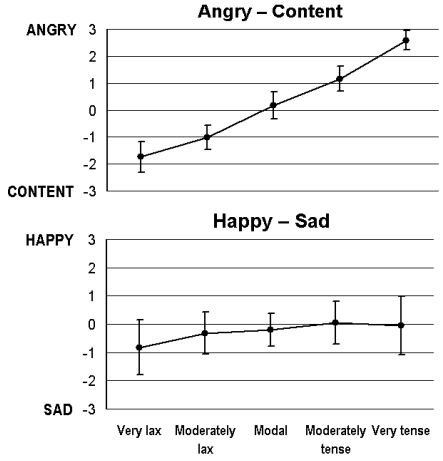


Figure 2: *Mean ratings (vertical lines show one SD) of perceived affective strength for different degrees of tenseness/laxness.*

Further experiments have explored how phonatory quality and f_0 combine to signal affect. It was felt that the generally low ratings obtained for strong emotions in the first experiment was likely to reflect the fact that these stimuli lacked the large f_0 excursions, described in the literature for these emotions, e.g., [1,7].

To examine this, in a third experiment, two sets of stimuli were constructed, one including voice quality differences and large f_0 excursions (the 'f_0 + VQ' stimuli), and a second set which differed similarly in terms of f_0 variation, but retained the phonatory settings for modal voice (the 'f_0 only' stimuli). The f_0 values were adapted from data presented by Mozziconacci [7], based on measurements of utterances produced with the following types of affect: joy, sadness, anger, indignation, boredom, fear and neutral. For the 'f_0 + VQ' stimuli, the matching of a phonatory quality to a particular f_0 contour was guided by the results of the first experiment where relevant, and otherwise by suggestions in the literature (for details, see [8]). The pairs of affective attributes tested included not only those for which we had specific f_0 contours, but also further affect-pairs, used in the first experiment.

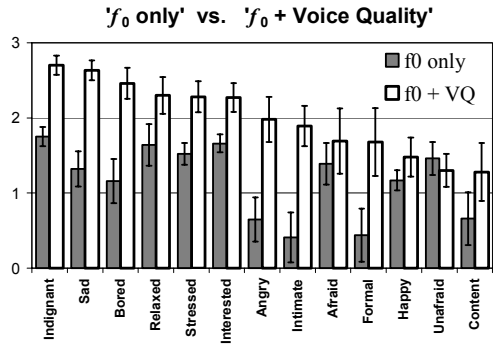


Figure 3: *Maximum mean ratings and estimated standard error of the mean: 'f_0 only' (grey); 'f_0 + VQ' (black). Affect ratings: 0 = none, 3 = max.*

Results do show relatively higher ratings for 'strong' emotions when the large pitch variations are incorporated. In Fig. 3 is shown the maximum mean rating obtained for each affect across the stimuli of either set. Overall, it is striking that the 'f_0 only' stimuli (grey) yield rather low ratings and are considerably less effective than the 'f_0 + VQ' stimuli (white) in signalling these affects.

The highest ratings were not always found for the expected stimulus and a closer inspection can be instructive. The highest rating for *sad* was obtained for the stimulus with the f_0 contour for *boredom* + lax-creaky voice, rather than by the stimulus with the f_0 contour for *sadness* + breathy voice. In the case of *happy* it is interesting to note that the highest rating was for a stimulus with f_0 for *indignation* + harsh voice, the same stimulus as gave the highest rating for *angry*.

This experimental approach provides a tool that can be exploited to look at cross-language differences in the paralinguistic coding of affect. The above experiments all involved subjects who were speakers of Irish English. A similar perception experiment involving Greek subjects has been carried out by Cherouvis [9], and will be extended to other language groups. This latter experiment evolved from the last one described, and includes three groups of stimuli: 'f_0 only', 'VQ only' and 'f_0 + VQ'. Although results cannot be directly compared with those of the third experiment above, we can tentatively point to some likely differences between the two language groups, as well as similarities. As for the Irish English subjects, the tense-lax dimension was of major importance. The most striking difference concerns the lax-creaky voice quality. Whereas the addition of creakiness to breathy voice (as in the lax-creaky stimuli) appears to be a potent indicator of boredom for the Irish English subjects, this is not the case for the Greek subjects.

Results broadly confirm the findings of the third experiment in that the highest ratings across the board were obtained for the 'f_0 + VQ' stimuli. The 'VQ only' stimuli

yielded considerably higher ratings generally than the 'f_0 only', which typically obtained low ratings.

3. Non-affective functions of voice quality

These experiments highlight the fact that the two aspects (f_0 and voice quality) of the voice source need to be taken together in trying to get at the affective role of phonation. Furthermore, there is yet another sense in which an integrated approach is required. In any act of spoken communication, phonatory quality carries three different types of information: speaker specific, linguistic, and affect-related paralinguistic information, and we would argue that the latter cannot be treated in isolation.

There are of course potentially large differences in speakers' long-term average phonatory settings. This to a large extent reflects differences in underlying physiology, but it can also be at least partially dependent on the language and social group the speaker belongs to. It has nothing to do with affect communication, but it presents a measurement problem, greatly complicating any quest for invariant voice quality mappings of affect.

While cross-speaker variations are self-evident, perhaps less widely appreciated is the fact that there are extensive modulations of the source that are connected to the linguistic content of utterances. To exemplify, in Fig. 4 are shown the dynamic variation of voice source parameters for the Swedish utterance "Inte i detta århundrade" [ɪntɪˣdɛtːaoːrhəndradə], taken from [10]. The prosodic modulation of this utterance is reflected not only in the variation of f_0, but also in the other source parameters. For example, declination shows up clearly in the EE parameter. The prominence of accented syllables is a consequence of different factors, not only their f_0 prominence, but also by source indicators of increased phonatory tenseness. These are here the level of EE (excitation strength of the glottal pulse) and, sometimes, FA (showing the relative strengthening of the higher components of the source spectrum). Phrase boundaries are typically characterised by increasingly breathy voice as can be inferred in this example form the increasingly symmetrical glottal pulse (rising RK towards phrase end) and the falling FA. For explanations of these source parameters, see [5]. Source data are complex: an episode of creakiness towards the end of the utterance brings about major perturbations to these parameters, but is also fairly typically associated with the phrase boundaries of declarative sentences of Swedish.

The fragment in Fig. 5 illustrates the final two words in an intonational phrase (from a Swedish utterance). The nuclear accent is on "sal" [sɑːl]. Note that prominence on the word "sal", associated with the low tone in early part of the vowel and is clearly marked by a tenser phonatory quality. Again a combination of parameters point to this: the high excitation strength (EE), the relative strengthening of the higher source components (high FA) and the change in the lower end of the source spectrum, indicated by the high "glottal frequency" (high RG). For a somewhat more extended discussion of prosody related source variation, see [11].

Although a large part of (non-affective) voice source variation is a manifestation of the prosody, some variation is not, but reflects segmental influences. There are intrinsic differences among different classes of vowels and consonants, which are the consequence of the varying supraglottal conditions pertaining to their articulation (see, e.g., [12]). There are also potentially large effects on the source at the transition to voiceless consonants, and generally smaller effects at voice onset. Some such effects can be observed in Fig. 4. The offset effects can be very pervasive, and in certain languages/dialects may influence a large part of a preceding vowel [13].

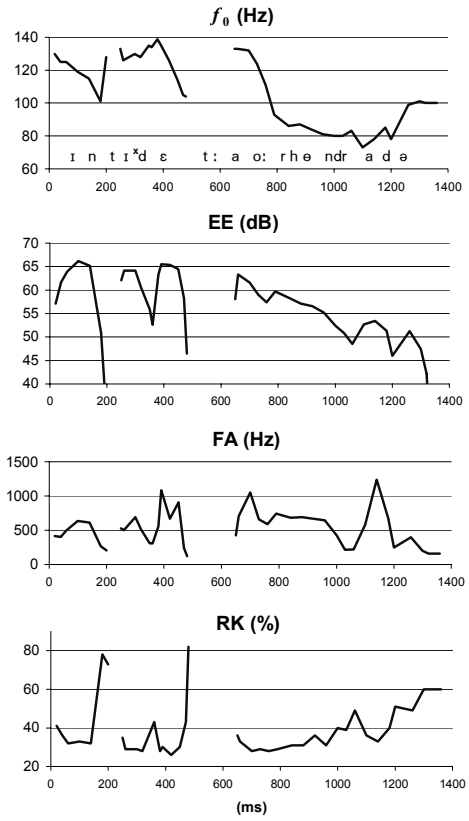


Figure 4: *Voice source parameter variation for an utterance of Swedish.*

4. Discussion and conclusions

The results of the experiments reported in Section 2 indicate that to understand the role of phonatory quality in the expression of affect, it needs to be looked at in conjunction with f_0 variation, as these two aspects of the source work together. (Temporal aspects, which we have not dealt with here, are of course also important.) These results also suggest that attempting to synthesise affective speech by incorporating appropriate f_0 without the appropriate voice quality is a strategy unlikely to yield the desired results. This may explain the frequently poor results reported for perception experiments of affect judgements, based on production measures of f_0 and timing parameters, but not voice quality.

The fragments of data presented in Section 3 reinforce this argument. Furthermore, it is clear that in order to understand the phonatory correlates of affect, we will need to take a broader perspective, which takes account of the speaker specific as well as the linguistic (prosodic and segmental) contributions to the dynamic variation of the source – on two accounts. At a practical level it is necessary in order to deal

with the "measurement problem" – how to obtain measurements for phonatory correlates of affect that are not confounded by these other factors which govern voice quality. But more fundamentally, we believe that the broader approach is required to gain a proper understanding of how affect is coded in the voice. Our current hypothesis is that when it comes to the source, different affects are signalled, not by simple, global transforms affecting entire utterances, but rather by changes to the utterance-internal prosodic constituents and to the relationships among them.

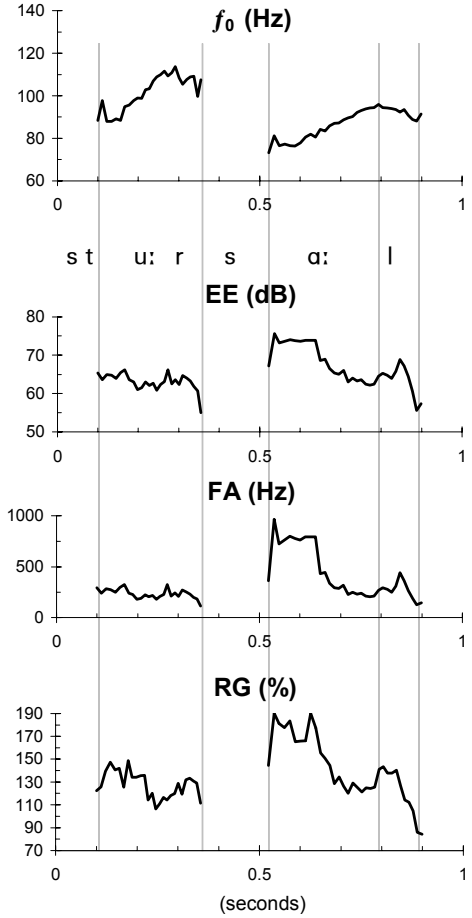


Figure 5: *Variation in voice source parameters for the Swedish utterance fragment* [stu:r sɑ:l].

These considerations provide the rationale of a recently initiated project, *Prosody of Irish Dialects*. As part of an account of Irish prosody, this project will attempt to provide some parallel coverage of phonatory quality (and temporal) correlates of the linguistic constituents, which are normally defined only in terms of f_0. Furthermore, the project will attempt to illustrate how certain aspects of the paralinguistic code of affect expression influences the realisation of the linguistically defined constituents. So for example, rather than seek an account of which phonetic parameters are associated with a specific affect, e.g., *indignant*, we would aim to illustrate how an indignant utterance can differ from an affectively neutral rendition, in terms of how the elements of the linguistic analysis (focally accented/deaccented syllables, declination, etc.) are transformed for these phonetic dimensions. Ultimately, we argue for a holistic approach where the paralinguistic signalling of affect is treated as an inherent, fundamental dimension of prosody and reintegrated into prosodic analysis [11].

5. Acknowledgments

This work is supported by a Govt. of Ireland Senior Research Fellowship to the first author, and by the project *Prosody of Irish Dialects*, both of which are funded by the Irish Research Council for Research in the Humanities and Social Sciences.

6. References

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

[2] Gobl, C. and Ní Chasaide, A., "Acoustic characteristics of voice quality", *Speech Communication*, Vol. 11, 481-490, 1992.

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

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

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

[6] Ryan, C., Ní Chasaide, A. and Gobl, C., "Voice quality variation and the perception of affect: continuous or categorical?", *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2409-2412, 2003.

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

[8] 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.

[9] Cherouvis, S., "Vocal Expression of Affective States in Modern Greek", unpublished M.Phil. dissertation, University of Dublin, Trinity College, Ireland, 2001.

[10] Gobl, C., "Voice source dynamics in connected speech", *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Vol. 1, 123-159, 1988.

[11] Ní Chasaide, A. and Gobl, C, "Voice quality and f_0 in prosody: towards a holistic account", *Proceedings of the 2nd International Conference on Speech Prosody*, Nara, Japan, 2004.

[12] Gobl, C., Monahan, P. and Ní Chasaide, A., "Intrinsic voice source characteristics of selected consonants", *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Vol. 1, 74-77, 1995.

[13] Ní Chasaide, A. and Gobl, C., "Contextual variation of the vowel voice source as a function of adjacent consonants", *Language and Speech*, Vol. 36, 303-330, 1993.