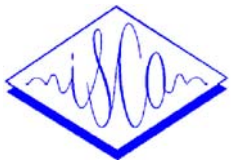

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Voice Quality and f_0 in Prosody: Towards a Holistic Account

Ailbhe Ní Chasaide & Christer Gobl

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

Abstract

This paper presents a discussion of the role of voice quality in prosody. Illustrations from past production and perception data by the authors indicate that source parameters other than f_0 are an inherent part of prosody, implicated in both its linguistic and paralinguistic functions. While prosodic (intonational) analyses of a language tend to be largely presented in terms of f_0 dynamics, the argument here is for an integrative approach, where f_0 and voice quality – two dimensions of the voice source – are treated together, and are related to the temporal/rhythmic structure of utterances. This should yield a fuller understanding of the nature of prosody and of the underlying production and perceptual correlates of prosodic elements such as pitch accent, declination, focus, phrase boundaries, etc. Such an approach may also serve to bring together the currently fragmented accounts of two core aspects of prosodic functioning: its role in signalling (i) linguistic, contrastive and discourse-related information and (ii) in communicating speaker affect, i.e. mood, emotional state and attitude. While the illustrations presented here provide initial hypotheses, a newly initiated project on Irish prosody will seek to incorporate such a holistic approach to prosodic analysis.

1. Introduction

This paper is concerned with voice quality as an inherent dimension of prosody. Descriptions of the prosody of languages are usually accounts of the intonation, focussing almost entirely on f_0 dynamics, with some coverage of temporal features. However important, the fundamental frequency is only part of a complex source signal, all of which bears the imprint of prosodic modulation. The quality of phonation, i.e. the shape of the glottal pulses, and consequently the relative amplitude levels among all the harmonics of the source spectrum, vary dynamically in the course of utterances in a way that is to a great extent governed by prosodic factors.

In this paper some analysis and perceptual data are discussed, which illustrate that modulations of source parameters other than f_0 are also part and parcel of the prosody of speech. We argue here for a holistic approach to prosodic analysis that integrates f_0 and other source parameters along with the temporal dimension.

While such an approach may be technically challenging, it is one that is worth pursuing for at least two reasons. Firstly, it is likely to provide a more insightful account of the phonetic correlates of traditionally used intonational terms such as *prominence*, *accent*, *focus*, *boundary*, *declination*, etc. It should thereby further our understanding of the underlying production and eventually the perception of these prosodic elements. Secondly, this approach may provide a key to tackling a fundamental problem of current prosodic analysis, i.e. that of providing a unified account of the linguistic (grammatical and discourse-marking) and paralinguistic (affect signalling) functions of prosody.

Although current intonational analysis tends to focus almost exclusively on the former, it is interesting to note that some earlier treatments, such as [30] saw the paralinguistic function (the expression of emotion, mood and attitude) as being of fundamental importance, providing affective glosses to individual intonational contours. More recently, it has been pointed out that there is no one-to-one mapping of pitch contour to affect, e.g. [32]. Furthermore, the currently dominant autosegmental metrical approach, using ToBI transcription, involves a considerable degree of abstraction and is more suited to probing the more contrastive, categorical aspects of prosody. Both factors have contributed to the lack of engagement with the important paralinguistic dimension.

Compensating this gap in linguistic coverage, the affective, expressive dimension of prosody has been mostly the preserve of psychologists like Scherer [35], researching within a rather different perspective the vocal cues to emotion. Here, the interest is directed at rather strong emotions such as anger, joy, fear, etc., providing coverage in principle of some paralinguistic phenomena. Over the years this research has provided a wealth of information on how global changes to f_0 dynamics, amplitude and tempo serve to differentiate utterances produced with different intended emotional overtones. Unfortunately, given the rather different methodological frameworks, it is difficult to relate these findings to current linguistic accounts of intonation and prosody (see for example the related discussion in [36]). From a perusal of the literature, one would be forgiven for thinking that two entirely different entities are involved, rather than two facets of a single phenomenon.

The need for a more integrated coverage of prosody is highlighted within speech synthesis, where, ironically, the now rather natural basic voice quality of the better systems serves to highlight the fact that they lack the expressive, affective character of normal human speech. A fresh impetus for research on the affective dimension of is evidenced by initiatives such as the JST/CREST *Expressive Speech Project* and the ISCA workshops *Speech and Emotion* and *VOQUAL'03*. And although the importance of voice quality to the paralinguistic aspect of spoken communication is amply acknowledged if little understood, the role of voice source modulation as a basic dimension of the prosody of non-emotive speech is little appreciated. It is our contention that its inclusion is likely to be crucial to understanding both aspects of prosody.

A multidimensional analysis that includes the voice source may in fact be the only way that will allow us to treat both aspects within a single framework, allowing us to model how the phonetic (f_0, voice quality and temporal) dimensions that comprise the linguistic constituents of utterances are modified in ways that impart affective colouring. This is our

long-term goal and we hope to explore it in a newly initiated project on the prosody of Irish dialects in a way that will provide a platform for a future more extended holistic treatment of prosody.

In Section 3, we illustrate in a qualitative way, how voice source modulation appears to be an inherent part of the basic intonation of utterances. Section 4 presents some perceptual, synthesis-based research which explores the role of voice quality (with f_0) in the paralinguistic signalling of affect. As a preliminary, Section 2 provides a brief sketch of some important source parameters.

2. Glottal source parameters

The glottal source varies constantly in the course of running speech, and this variation is governed by both the prosodic and the segmental context. In our past research on source variation, the main analysis technique has involved a two-step procedure: firstly, the speech pressure waveform is inverse filtered to obtain the differentiated glottal source signal; secondly, salient parameters of the source are quantified by matching a model of the glottal flow (the four-parameter LF model [9]) to the differentiated glottal flow pulses. Interactive software has been used which allows optimisation of the inverse filter settings and of the glottal model match in both the time and the frequency domain. A comprehensive account of this and other voice source analysis techniques is presented in [11] and [15].

In order to contextualise the source data presented in Figs. 2 and 3, some important parameters are illustrated in Fig. 1. Schematic representations on the left of the figure illustrate parameters in terms of true ($U_g(t)$) and differentiated ($U_g'(t)$) glottal airflow; on the right is shown the main effect that parameter changes would have on the source spectrum. For further details of these and other kinds of source measures see [11,28].

EE, the excitation strength is defined as the negative amplitude of the differentiated glottal flow at the main waveform discontinuity. An increase in EE boosts all source components equally, with the exception of the very lowest, particularly the first harmonic.

TA, RA and FA. TA is a measure of the effective duration of the return phase of the glottal pulse, and it relates to the sharpness of the glottal closure. RA is TA normalised to the fundamental period, i.e. TA/T_0. It is a major determinant of the slope of the source spectrum. In the frequency domain, the effect of the return phase can be modelled as first order low-pass filter with a cutoff frequency $FA = 1/(2\pi TA)$, the frequency above which there is an additional attenuation of the source spectrum. Thus, a high FA (or a low TA) value indicates a relative boosting of the higher harmonics of the source.

RK captures the degree of skewing of the glottal pulse: the larger the RK value, the more symmetrical the pulse shape.

RG is a measure of the glottal frequency normalised to f_0: it affects the relative amplitudes of the very lowest components in the source spectrum.

OQ, the open quotient is a measure of the proportion of the glottal cycle for which the glottis is open. In the frequency domain, there is a close correspondence between the OQ value and the amplitude of the first harmonic.

In the perception studies discussed in Section 4, stimuli with different voice qualities were synthesised using the LF source implementation in KLSYN88a [20]. For the synthesis, our standard parameters EE, RA, RG and RK were transformed into the corresponding KLSYN parameters AV, TL, SQ and OQ (for details see [17]). Two further KLSYN88a parameters were used: DI (diplophonia) for the generation of creakiness and AH for aspiration noise.

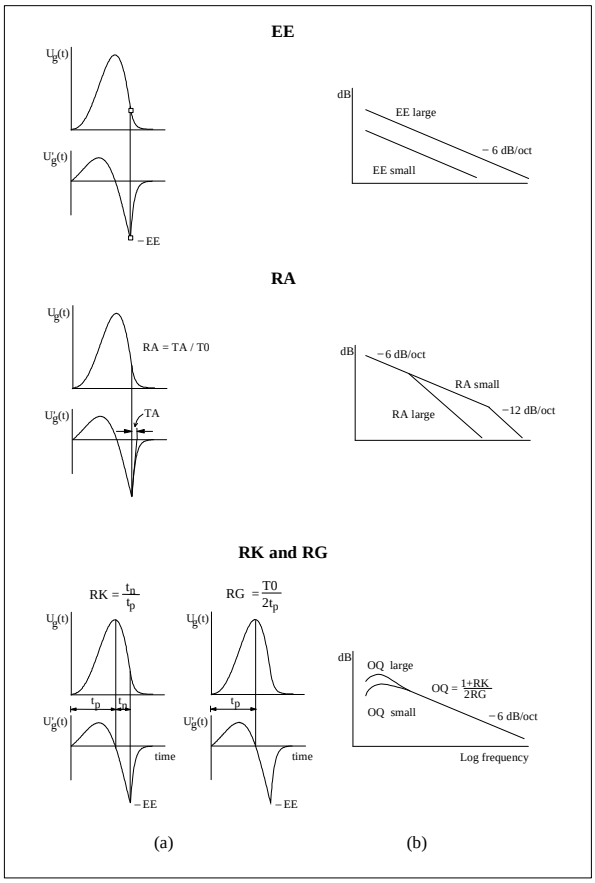

Figure 1: *Illustration of the source parameters EE, RA, RK and RG: (a) in terms of true and differentiated glottal flow and (b) showing the effects of a change in parameter values on the source spectrum.*

3. Voice quality as a dimension of intonation and tone

To illustrate how source parameters other than f_0 may be contributing to intonation, Fig. 2 shows the variation in a number of source parameters, along with f_0, in the course of the Swedish utterance *en alldeles utmärkt idé* (an altogether excellent idea) [ɛnˈaldɛsˣ ʉːtmæːrktɪˈdeː].

In this phrase, the word *utmärkt*, which carries the grave accent of Swedish, is focally accented: as the speaker has the East Swedish dialect, the focally accented word is characterised by a high falling pitch on the first syllable, which carries the word accent: the high peak on the following syllable is a manifestation of the focal accent (see, for example, [1]).

We would point out that this utterance (and those on which Fig. 3 are based) was spoken with what would be

described as modal voice. The source modulations that can be seen are *not* heard (consciously) by the listener as shifts in voice quality, but rather as part and parcel of the prosodic content of the utterance (or, as we also discuss below, the segmental structure).

The **declination** of the utterance is typically reflected in the EE parameter, which shows the reduction in the excitation strength over the course of the utterance. Because the focal word is strongly accented, the declining EE is perhaps less striking here than in many other examples, but see [10] for further illustrations. While there is typically a close correspondence between EE and intensity, it is worth noting that changes in EE have a relatively small effect on the very lowest end of the source spectrum, and therefore the first harmonic is likely to become increasingly dominant with the decline in EE. Changes to other parameters such as OQ indicate that the lower end of the spectrum is in fact boosted towards the end of the utterance, suggesting an increasingly lax phonation. Fant and Kruckenberg [7] have estimated that the declination in EE progresses at approximately 4 dB per second, with an increasing rate of decline towards the end of the utterance.

Phrase boundaries may also be marked by the voice source quality. A breathy-voiced utterance termination is very typical. The increasingly breathy mode of phonation at the end of the utterance can be observed in Fig. 2: the rising values for the OQ, RK and RA (the latter not shown) parameters, indicate a greater open quotient, increasing symmetry of the glottal pulse and increasing dynamic leakage. The rising OQ indicates a boosting of the low end of the source spectrum, while the corresponding drop in FA indicates a reduced level for the higher frequencies. Note that breathy voice (indicated by these same parameters) is also associated with phrase onsets, although in comparison to utterance final position, it tends to be of rather short duration.

Towards the end of the utterance (about timepoint 1515 ms) there is a brief episode of creaky voice, which shows up in the perturbations to the different source parameters. Such episodes are often associated with the very low f_0 towards the end of declarative utterances of Swedish and of some accents of English that we have looked at. A creaky voice termination has also been described by a number of other researchers. Epstein [5] has pointed out that the creaky phonation associated with the boundary in her American English data occurs only with the low boundary tone and not with the high. While this is not surprising, given the link between creaky voice and low pitch, she also points out that creaky voice is not otherwise associated with low pitch in her data.

The degree of **prominence** of syllables and the difference between **accented** and unaccented syllables appears to depend not only on f_0 and duration but also on specific characteristics of the glottal pulse, which tend to yield a tenser mode of phonation on the accented.

The focally accented word in Fig. 2 shows that the spectral tilt and the pulse excitation strength can contribute together or separately. The [ʉ] vowel of the first syllable, which carries the word accent, shows a momentary peak in FA (i.e. a relative boosting of higher harmonics) along with a rather high EE. The [æ] of the second syllable, shows a very high EE, but not the boosting of the higher harmonics as indicated by the relatively low FA. Both FA and EE can contribute to perceived loudness, but in different ways.

These differences suggest that different strategies may be involved to bring about increased prominence. In the vowel [eː] of the final accented syllable, FA is clearly contributing importantly to its prominence. Other studies [2,5,8,37] have also pointed to the boosting of higher harmonics as a correlate of accentuation.

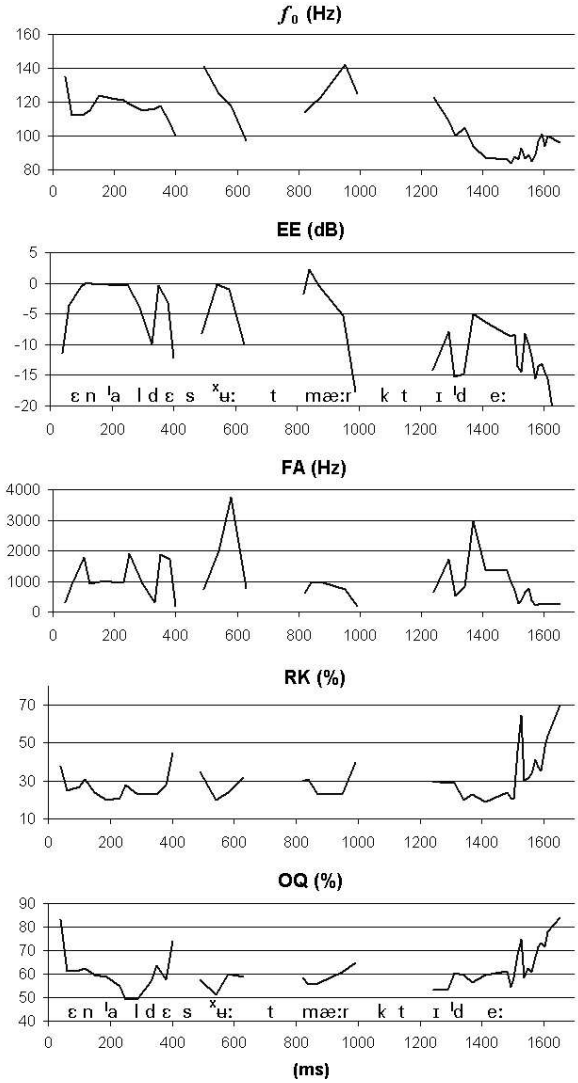

Figure 2: *Variation of f_0 and of the source parameters EE, FA, RK, and OQ in the course of an utterance of Swedish.*

Source characteristic are described in [10] for the word *behålla* in focal, prefocal and postfocal positions of the Swedish utterance *vi vill behålla honom* (we want keep him) [vɪvɪlbɛhɔlːahɔnːɔm]. In Fig. 3 values for f_0 and EE are shown for a single repetition of the three conditions. It is striking that the dynamic range of the glottal source excitation is considerably greater when the word is in focal position, being stronger for the vowels and weaker for the surrounding consonants. This is effectively an enhancement of the vowel-consonant distinction in the focally stressed word, rather than a simple boosting of the excitation pulse or a lessening of the spectral tilt.

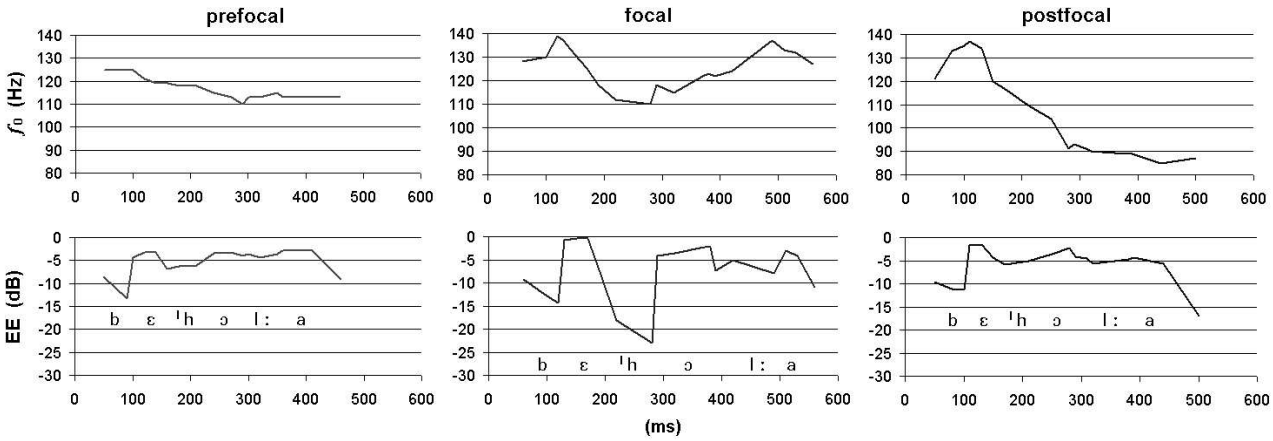

Figure 3: f_0 *and EE values for the word* [bɛˈhɔlːa] *in prefocal, focal and postfocal contexts.*

These data suggest that the glottal (and probably also the supralaryngeal) gestures are produced with greater effort. The corollary of this is, of course, that in deaccenting there is a relative loss of contrast between the vowel and consonant. For similar observations, see also Fant and Kruckenberg [6,7].

It is not clear whether are non-pitch source correlates to the distinction between **accented high** and **low tones**. Pierrehumbert [31] reported source correlates of pitch accents with high tones, which were not found otherwise with high pitch. Epstein's results [5] for American English are mixed and do not appear to yield distinct source characteristics for high vs. low pitch-accented tones.

As is also clear from Fig. 2, not all variations are prosodically determined. There are also segmental effects on source parameters, which are similar to the **micro-prosodic** effects that have been found for f_0. Note in Fig. 2, the sharp perturbations of the source parameters prior to the voiceless consonants in the utterance: sharp rises in OQ and RK with concomitant drops in FA and EE. These perturbations reflect the breathy-voiced termination of voicing in the vowel before the voiceless consonant, and they have described extensively in cross-language studies [16,27]. Although these source effects can be quite extensive, they are usually not detected as switches in voice quality per se, and can be treated as passive consequences of the early glottal abduction associated with voiceless consonants. Note that in the languages we have looked at, there are not thought to be segmental voice quality contrasts. There are also source changes that are associated with the altered phonatory conditions pertaining to consonant production, even when the consonants are voiced. These can be seen in Fig. 2 and are studied systematically in [13,29]. While these segment-related source effects are clearly not part of the prosody, it is nevertheless important to note them here, as one must take account of them when trying to get at the prosodic dimension of source variation.

We have not analysed **tonal** languages, but would propose that, in a similar way, variations of the voice source are likely to be equally relevant to the realisations of tonal contrasts. As with intonation, the modulation of the source may not typically be perceived as a change in voice quality. However, it is interesting to note that there is evidence in the literature that voice source changes appear to be sometimes salient enough to be detected as voice quality shifts associated with specific tones. For instance, Huffman [18] describes one of the seven tones in Hmong, a Sino-Tibetan language, as having a breathy voice quality. In the Wu dialect of Chinese, the yin and yang tones are also characterised by specific voice qualities [19]. The yang tones differ from the yin in that they employ breathy phonation and begin with a lower f_0. As f_0 differences tend to be associated with different voice qualities in any case, it is hardly surprising to find voice quality correlates of tonal contrasts and vice versa. Despite such correlations, many authors point out that f_0 and voice quality are separately controllable, and that variation in one does not allow prediction of variation in the other. The link between voice quality and f_0 may have historical implications, and the likelihood of tonal contrasts having evolved from earlier voice quality contrasts has been discussed [19,23,33]. Not surprisingly, there are cases where contrasts in specific languages are open to competing analysis as involving primarily voice quality or tone: see for example the debate concerning how the contrasts in Mon should be interpreted [4,22,38]. Discussions as to whether tone or voice quality should be considered the primary cue may be somewhat pointless, as clearly in these cases they are both salient and collaborating aspects of the distinction.

There is very little empirical data about the voice source contribution to basic intonational elements. Even less is known about its role in perception. Experience with rule-based synthetic speech indicates that we can produce highly intelligible utterances without including such source variation, but suggests that the naturalness of the speech output is compromised by the omission.

4. Voice quality, f_0 and the expression of affect

The expression of affect is a core function of prosody, but as discussed earlier is an area which is rather neglected within current linguistic approaches. Note that we use the term affect here as a cover term to encompass the speaker's state (emotion, mood, etc.) and those attitudes that are conveyed in interpersonal interactions (politeness, condescension, etc.). There is however a substantial body of research on the "strong" emotions, focussing particularly on f_0 dynamics and temporal variation (see for example [3,26,35]). As mentioned earlier, it is however difficult to relate this research to the linguistic body of research on intonation systems. So while we know, for example, that anger involves an increase in the average f_0 and in its dynamic range, we cannot tell how this is

distributed over the different linguistically relevant constituents of utterances.

People tend to be quite conscious of the voice quality shifts that occur for paralinguistic signalling, and phoneticians have traditionally linked specific voice qualities to particular affects (creaky voice to boredom, breathy voice to intimacy, whispery voice to confidentiality, etc.). Empirical work in on affect communication has been hampered by (at least) two obstacles. First of all, voice quality is difficult to measure, with the consequence that most researchers have tended to focus rather on the more measurable parameters of f_0, amplitude and tempo. Secondly, there is the problem of eliciting a suitable corpus. Most researchers rely on read sentences with acted emotions, and there has been considerable debate as to whether such data may be exaggerated and stereotypical samples, rather different from naturally occurring affective speech. Pioneering work which tackles both these problems is currently being carried out within the JST/CREST *Expressive Speech Project*, where sophisticated voice quality analysis is being carried out on entirely spontaneous corpora.

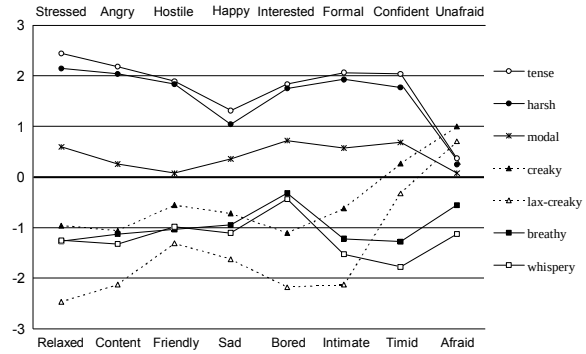

Figure 4: *Mean ratings of perceived affective strength of pairs of attributes for seven voice qualities. 0 = no affective content +/-3 = maximally perceived.*

Our own research in this area has not involved analytic studies. The approach adopted has been rather to probe the role of voice quality by conducting perception tests where listener's attributions of affect are elicited for stimuli synthesised with different voice qualities. In an experiment reported in [17] a short Swedish utterance *ja adjö* ['jɑ: a'jø:] was synthesised with seven different voice qualities: breathy voice, whispery voice, tense voice, harsh voice, creaky voice, lax-creaky voice, and modal voice.

The synthesis was guided by prior experience with the analysis of different voice qualities, e.g. [14], based on recordings of a male phonetician acquainted with the Laver classification system [21]. Full details on the design of the stimuli are presented in [17]. A series of perception tests elicited for a pair of opposite affective attributes (e.g., bored/interested) how each voice quality was rated, using a seven-point scale. The full set of attribute pairs tested included *relaxed*/*stressed*, *content*/*angry*, *friendly*/*hostile*, *sad*/*happy*, *bored*/*interested*, *intimate*/*formal*, *timid*/*confident*, and *afraid*/*unafraid*. For the listeners, Hiberno-English speakers, the utterance was semantically neutral.

The ratings for the different pairs of attributes are illustrated in Fig. 4. Although each pair of attributes was tested separately, the ratings for particular voice qualities are here joined across the individual tests to show more clearly the attributes associated with particular voice qualities. Note that 0 in this figure means that no affective colouring was detected, and that the distance from 0 in the positive or negative direction indicates the strength of attribute rating.

Results do demonstrate that changes to voice quality can on their own impart affective colouring to an otherwise neutral utterance. They also show that there is no neat one-to-one correspondence between voice quality and a particular affect, as traditional observations (e.g. boredom/creaky voice) would suggest. The tense – lax dimension of voice emerged as particularly important: tense voice appeared to be associated with high activation and high power states (*stressed, angry, confident, formal*) while the lax-creaky voice was associated with low activation states (*relaxed, bored, intimate*).

A further experiment, reported in [34] looked closer at the tense – lax dimension, to explore whether gradient changes in voice quality would yield similarly gradual alterations in listeners' affect ratings. It has often been claimed that linguistic aspects of prosody depend on categorical judgments, while the paralinguistic dimension involves more continuous changes in the phonetic dimensions as well as in perception. An alternative possibility we had considered was that at different points in a phonetic continuum (here the tense – lax continuum) discretely different affects might emerge (e.g., *happy* being conceivably associated with a moderately tense voice, while *angry* would be more likely to be associated with an extremely tense voice).

In the perception test, scalar ratings were elicited for similar pairs of affective attributes in a similar way to the previous experiment. There were five stimuli in this experiment, drawn from a synthesised tense – lax voice continuum, using the modal stimulus of the first experiment as the point of departure. The results are illustrated in Fig. 5 (note that Lax5 and Tense5 refer to the most lax and tense stimuli respectively). They indicate that affective ratings do vary in a continuous fashion with the voice quality continuum.

It was also striking in the first experiment (Fig. 4) that, with the clear exception of *angry*, strong emotions (such as *happy, sad, afraid*) yielded rather low ratings, whereas milder states of being (such as *bored, relaxed, formal*) tended towards higher ratings. A likely explanation for the low ratings of strong emotions is that these stimuli lacked the large f_0 excursions, which are known to be associated with the signalling of strong emotions [24,25,35].

To probe this question, a further experiment [12] explored the way in which voice quality and f_0 combine, by presenting listeners with two types of stimuli: 'f_0 + VQ' comprised a stimulus set where the most likely voice quality candidate for a particular emotion was given a more appropriate f_0 contour, described for that emotion; and 'f_0 only' stimuli, which contained the same set of f_0 contours, but had modal voice quality throughout. Note that what we are referring to here as different 'contours' involved in fact differences in f0 level and range, and all shared a similar basic, double-peaked contour.

The utterance used was as in the first experiment, and the voice qualities were largely similar. The f_0 contours were adapted from experimental data presented by Mozziconacci [24], based on analyses of utterances recorded with the following affects: indignation, fear, joy, anger, sadness, boredom and neutral. The basic f_0 contour for the neutral utterance in [24] was rather similar to that of our own modal utterance, and so the latter served as our neutral stimulus.

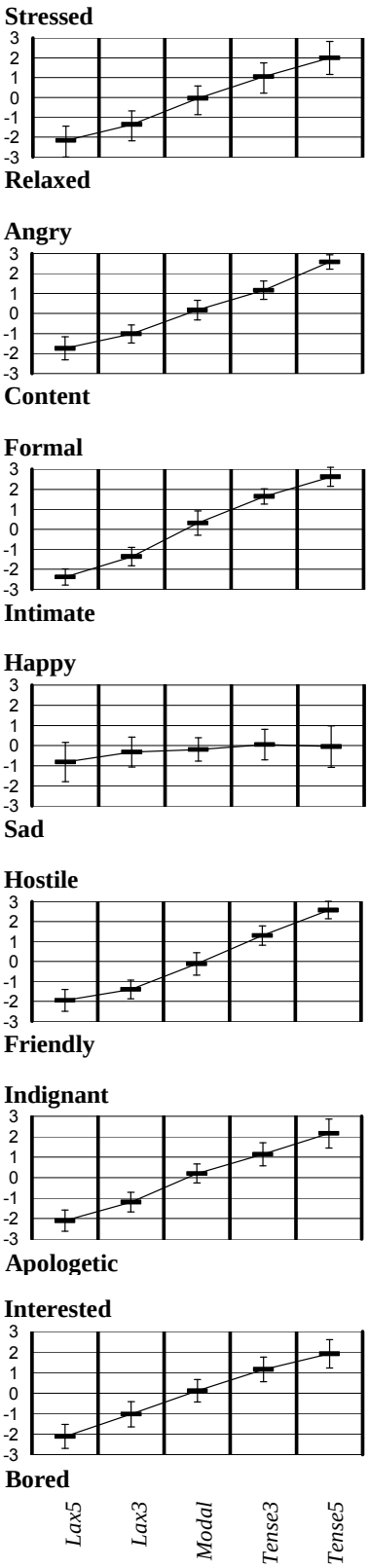

Figure 5: *Mean ratings for five stimuli spanning the tense – lax continuum (vertical lines show one SD) for seven pairs of affective attributes.*

The non-modal f_0 values were arrived at, by a linear scaling of the values in [24] to retain the relative differences with respect to the f_0 values of our neutral stimulus (for a fuller account, see [12]). Listeners' ratings were elicited using the same method as in the first experiment: the attribute-pairs included not only those affects for which Mozziconacci had described f_0 contours, but also some pairs such as intimate/formal.

The results are shown for both sets of stimuli in Fig. 6. It is striking that stimuli with the f_0 excursions but without additional voice quality adjustments (the 'f_0 only' stimuli) yield rather low affective ratings, considerably lower on the whole than the 'f_0 + VQ' stimuli. Although these results are not directly comparable with the results of the first experiment, we can infer from the relatively high ratings obtained here for *sad* and *afraid*, that an appropriate pitch contour may indeed be particularly important in cueing the strong emotions.

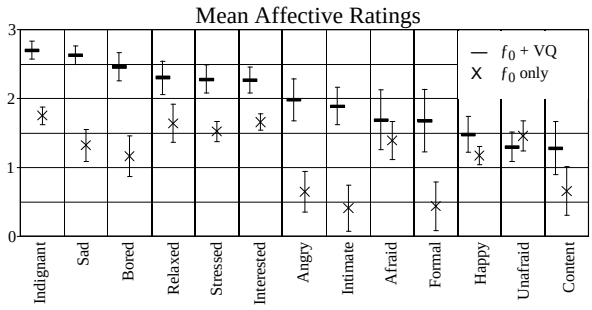

Figure 6: *Maximum mean rating and estimated standard error of the mean for each affect: stimuli 'f_0 + VQ' (–); 'f_0 only' (x). Affect ratings: 0 = none, 3 = max.*

These experiments indicate that the voice quality dimension must be included in any attempt to capture this aspect of prosody. The ineffectiveness of the 'f_0 only' stimuli in the last experiment is likely to reflect two different, though related causes. Firstly, and quite simply, voice quality is an essential cue to affect and was lacking in these stimuli. But furthermore, given that in human productions, large f_0 excursions are unlikely to occur without concomitant shifts in other voice source parameters, the 'f_0 only' stimuli are likely to sound relatively unnatural.

5. Integrating the linguistic and paralinguistic?

These experiments prompt further questions, e.g., concerning the extent to which the voice quality to affect mapping is language specific, or reflects universal tendencies. We plan to explore the area further using this experimental approach. Additionally, however, we feel that a research priority is to find a way to analyse prosody in a way that encompasses both the linguistic and paralinguistic: to illuminate the latter we need to understand how those features that constitute the linguistic elements are modified to express affect.

Within a newly initiated project on the prosody of Irish dialects, one of the goals will be to explore such a holistic approach. As we hold that even to specify linguistic aspects, the voice source dimension is likely to be important, we hope to include it in our analysis in so far as is feasible. Voice source data need however to be considered in conjunction

with f_0 and temporal information. The goal is not to seek simple, or invariant source correlates of the linguistic elements of prosody. Neither do we want to make exaggerated claims about source correlates, or go down the road of regarding cues as in some sense competing, with some more dominant. Rather the goal is to gain some understanding into how the different phonetic dimensions conspire to achieve the linguistic and discourse-related prosodic marking of utterances. Our past observations from limited amounts of analytic data, such as those illustrated in Section 3, serve to provide us with intuitions and hypotheses, which the Irish prosody project gives us an opportunity to explore and quantify. For example, in relation to Fig. 3 it was suggested that, even at the level of the voice source, different strategies may be contributing to achieve prominence: strengthening the excitation (EE) and/or decreasing the spectral tilt (raising FA). Changes in the dynamic range of EE in a way that enhances the vowel/consonant contrast would also appear to play a role. These source properties would enhance the contribution of the f_0 shifts and of increased duration.

The fact that perceived prominence can be achieved by different strategies means that we can expect languages, dialects, or even individuals to vary the way that they exploit the possibilities. Cross-speaker differences in acoustic correlates of focal accentuation of prominence are reported in [8]. From our perspective, this is not a problem: invariance of individual correlates should probably not be expected, and to seek it might blind us to some of the more interesting aspects of prosody. A richer, integrated coverage of all these phonetic dimensions should throw light on, or at least lead to new questions and hypotheses concerning the underlying control of prosody, e.g., concerning the respiratory vs. laryngeal contribution to the accenting of syllables, the different types of laryngeal gestures that may be involved, etc. Likewise, such data should in principle raise new questions concerning the perceptual integration of f_0, voice quality and duration.

A particular challenge will be to describe how affect is coded in these same phonetic dimensions, basing the description within a linguistic (but quantitative) account. So for example, rather than an account of which parameters are associated with a specific affect, we would aspire to a description which would detail how an indignant utterance can differ from an affectively neutral rendition, in terms of how the elements of the linguistic analysis (accented/deaccented syllables, declination, etc.) are transformed for these phonetic dimensions. While experiments such as the ones described in Section 4 do suggest that paralinguistic signalling involves gradient transformations of acoustic parameters, we feel it is unlikely that these transformations operate in a global, undifferentiated fashion across the prosodic constituents of utterances.

6. Conclusions

We suggest that voice quality is an inherently important dimension of prosody regardless of whether we are describing its more grammatical and discourse-related aspect or the more paralinguistic one of signalling affect. It is also imperative that the source be considered as a whole, and that voice source parameters be treated along with f_0 and indeed with temporal data. Given that f_0 is easier to hear, to analyse and to quantify, data on voice quality is inevitably going to lag behind. However, we would argue that in analysing prosody simply as f_0 dynamics, it should always be borne in mind that it is a partial account.

In order to understand the paralinguistic aspect of prosody, we suggest that an account will be required that is framed within the same terms as the description of the linguistic dimension. A major goal will be to understand how the phonetic parameters that combine to mark the linguistically relevant elements of prosody are transformed to communicate affect and attitude.

Clearly, these are long-term aspirations, but we believe that any progress in these directions will bring new insights into the nature of prosody.

7. Acknowledgments

This work has been financially supported by a Government of Ireland Senior Research Fellowship to the first author, funded by the Irish Research Council for Research in the Humanities and Social Sciences, and by the research project *Prosody of Irish Dialects: the use of intonation, rhythm and voice quality for linguistic and paralinguistic signalling*, which is also funded by the Irish Research Council for Research in the Humanities and Social Sciences.

8. References

[1] Bruce, G.; Gårding, E., 1978. A prosodic typology for the Swedish dialects. In *Nordic Prosody*, E. Gårding, G. Bruce and R. Bannert (eds.). Lund: Department of Linguistics, 219-228.

[2] Campbell, N.; Beckman, M., 1997. Stress, prominence, and spectral tilt. *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, Greece.

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

[4] Diffloth, G., 1985. The registers of Mon vs. the spectrographist's tones. *UCLA Working Papers in Phonetics*, Department of Linguistics, University of California, Los Angeles, 60, 55-58.

[5] Epstein, M., 2003. Voice quality and prosody in English. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2405-2408.

[6] Fant, G.; Kruckenberg, A., 1994. Voice source parameters in connected speech. A progress report. *Working Papers 43*, Department of Linguistics, Lund University, Lund, Sweden, 58-61.

[7] Fant, G.; Kruckenberg, A., 1995. The voice source in prosody. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Vol. 2, 622-625.

[8] Fant, G.; Kruckenberg, A.; Liljencrants, J; Botinis, A., 2002. Individual variations in prominence correlation. Some observations from "lab-speech". *TMH-QPSR*, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Vol. 44, 177-180.

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

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

[11] Gobl, C., 2003. The voice source in speech communication – production and perception experiments involving inverse filtering and synthesis. D.Sc. thesis, Royal Institute of Technology (KTH), Stockholm.

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

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

[14] Gobl, C.; Ní Chasaide, A., 1992. Acoustic characteristics of voice quality. *Speech Communication*, 11, 481-490.

[15] Gobl, C.; Ní Chasaide, A., 1999. Techniques for analysing the voice source. In *Coarticulation: Theory, Data and Techniques*, W.J. Hardcastle and N. Hewlett (eds.). Cambridge: Cambridge University Press, 300-320.

[16] Gobl, C. and Ní Chasaide, A., 1999. Voice source variation in the vowel as a function of consonantal context. In *Coarticulation: Theory, Data and Techniques*, W.J. Hardcastle and N. Hewlett (eds.). Cambridge: Cambridge University Press, 122-143.

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

[18] Huffman, M.K., 1987. Measures of phonation type in Hmong. *Journal of the Acoustical Society of America*, 81, 495-504.

[19] Jianfen, C.; Maddieson, I., 1989. An exploration of phonation types in Wu dialects of Chinese. *UCLA Working Papers in Phonetics*, Department of Linguistics, University of California, Los Angeles, 72, 139-160.

[20] 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.

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

[22] Lee, T., 1983. An acoustical study of the register distinction in Mon. *UCLA Working Papers in Phonetics*, Department of Linguistics, University of California, Los Angeles, 57, 79-96.

[23] Maddieson, I.; Hess, S.A., 1987. The effect on F0 of the linguistic use of phonation type. *UCLA Working Papers in Phonetics*, Department of Linguistics, University of California, Los Angeles, 67, 112-118.

[24] Mozziconacci, S., 1995. Pitch variations and emotions in speech. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, 1, 178-81.

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

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

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

[28] Ní Chasaide, A. and C. Gobl, C., 1997. Voice source variation. In *The Handbook of Phonetic Sciences*, W.J. Hardcastle and J. Laver (eds.). Oxford: Blackwell, 427-461.

[29] Ní Chasaide, A.; Gobl, C.; Monahan, P., 1993. Dynamic variation of the voice source in VCV sequences: intrinsic characteristics of selected consonants. *Proceedings of the Grenoble Workshop, Esprit/Basic Research Action no. 6975: SPEECH MAPS*, Vol. 2, Grenoble, Institut de la Communication Parlée, 44 pp.

[30] O'Connor, J.D.; Arnold, G.F., 1961. *The Intonation of Colloquial English*. London: Longman.

[31] Pierrehumbert, J.B., 1989. A preliminary study of the consequences of intonation for the voice source. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 4, 23-36.

[32] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*, P. Cohen, J. Morgan and M. Pollack (eds.). Cambridge, Massachusetts: MIT Press.

[33] Rose, P., 1989. Phonetics and phonology of Yang tone phonation types in Zhenhai. *Cahiers Linguistiques Asia Orientale*, 18, 229-245.

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

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

[36] Scherer, K.R.; Ladd, R.D.; Silverman, K.E.A, 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, 76, 1346-1356.

[37] Sluijter, A.; van Heuven, V., 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471-2485.

[38] Thongkum, T.L., 1987. Another look at the register distinction in Mon. *UCLA Working Papers in Phonetics*, Department of Linguistics, University of California, Los Angeles, 67, 132-165.