Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 189–212

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

The role of voice quality in communicating emotion, mood and attitude

Christer Gobl *, Ailbhe Ní Chasaide

Phonetics and Speech Science Lab., Centre for Language and Communication Studies, Trinity College, Dublin 2, Ireland

Abstract

This paper explores the role of voice quality in the communication of emotions, moods and attitudes. Listeners' reactions to an utterance synthesised with seven different voice qualities were elicited in terms of pairs of opposing affective attributes. The voice qualities included harsh voice, tense voice, modal voice, breathy voice, whispery voice, creaky voice and lax–creaky voice. These were synthesised using a formant synthesiser, and the voice source parameter settings were guided by prior analytic studies as well as auditory judgements. Results offer support for some past observations on the association of voice quality and affect, and suggest a number of refinements in some cases. Listeners' ratings further suggest that these qualities are considerably more effective in signalling milder affective states than the strong emotions. It is clear that there is no one-to-one mapping between voice quality and affect: rather a given quality tends to be associated with a cluster of affective attributes.

Keywords: Voice quality; Affect; Emotion; Mood; Attitude; Voice source; Inverse filtering; Fundamental frequency; Synthesis; Perception

1. Introduction

The present paper focuses on the role that voice quality plays in the signalling of speaker affect, broadly defined to include aspects of speaker attitude, mood, emotion, etc. The experiments described are very exploratory in nature and part of ongoing research on voice source variation in speech and on its function in communicating paralinguistic, linguistic and extralinguistic information. As part of this endeavour, we have been working towards the provision of acoustic descriptions of individual voice qualities (e.g., Gobl, 1989; Gobl and Ní Chasaide, 1992; Ní Chasaide and Gobl, 1995). Although the work has been mainly analytic, synthesis has been used to test and fine-tune our descriptions, and further to explore how individual source parameters or combinations of them may cue particular voice qualities (e.g., Gobl and Ní Chasaide, 1999a).

Growing out of this work, in the present study, listeners' responses were elicited for the affective colouring of synthetic stimuli differing in terms of voice quality. This allows us to demonstrate in the first instance some of the kinds of affective colouring that can be achieved through synthesis. Insofar as our synthetic stimuli approximate to the human qualities they were meant to capture, we hope ultimately to shed light on the role of different voice qualities in the human communication

* Corresponding author.
E-mail address: cegobl@tcd.ie (C. Gobl).

0167-6393/02/$ - see front matter
PII: S0167-6393(02)00082-1

of affect. By focussing on voice quality in this experiment, the aims were: firstly, to demonstrate whether and to what extent voice quality differences such as these can alone evoke distinct affective colourings, as has traditionally been assumed by phoneticians; secondly, to see to what extent results can lend support to past assumptions concerning the affective mapping of individual qualities and help clarify where rather contradictory claims have been made. A third objective is to provide a framework for subsequent exploration of how voice quality combines with f_0, and ultimately with the other known acoustic and temporal features that are involved in the expression of affect.

To date, research on the vocal expression of emotion has demonstrated that many features may be involved. Whereas there has tended to be an overwhelming focus on pitch variables (especially f_0 level, and range, but also the pitch contour and the amount of jitter) many studies have included the investigation of speech rate and intensity differences (Scherer, 1981, 1986, 1989; Mozziconacci, 1995, 1998; Stibbard, 2000; Williams and Stevens, 1972; Carlson et al., 1992). Other features may play a role, such as pausing structure (see, for example, Cahn, 1990a,b), segmental features, particularly those that relate to the precision of supraglottal articulation (Kienast et al., 1999; Laukkanen et al., 1996; Scherer, 1986; Carlson et al., 1992) or even rather fine grained durational effects such as the duration of accented and unaccented syllables (Mozziconacci, 1998). When present, extralinguistic interjections such as sighs, cries, inhalations (Scherer, 1994; Schröder, 2000) can provide powerful indications of the speaker's emotion. Comprehensive studies dealing particularly with f_0 variation, intensity, timing and spectral information have been carried out by Scherer and co-researchers over more than two decades. Useful overviews of empirical studies in this area can be found in (Scherer, 1986, 1989; Kappas et al., 1991; Frick, 1985; Murray and Arnott, 1993).

Although many researchers tend to stress its fundamental importance, relatively little is known about the role of voice quality in communicating affect. As pointed out by Scherer (1986) the tendency has been to concentrate on those parameters that are relatively easy to measure, such as f_0, intensity and timing, whereas voice quality has been neglected, relatively speaking, because of the methodological and conceptualisation difficulties involved (see Section 2). Scherer (1986) further asserts that "although fundamental frequency parameters (related to pitch) are undoubtedly important in the vocal expression of emotion, the key to the vocal differentiation of discrete emotions seems to be *voice quality*". Experimental support for the basic importance of voice quality can be found in experiments by Scherer et al. (1984), where different degradation or masking procedures were applied to spoken utterances as a way of masking features of intonation, voice quality and verbal content. Listener's evaluations of affect appeared to be primarily determined by voice quality cues, relatively independent of distortions in f_0 cues or presence/absence of identifiable verbal content.

Although there have been source analyses of different voice qualities in the literature (see, for example, Alku and Vilkman, 1996; Childers and Lee, 1991; Gobl, 1989; Gobl and Ní Chasaide, 1992; Lee and Childers, 1991; Price, 1989), very few empirical studies have focussed on the voice source correlates of affective speech. Laukkanen et al. (1996) studied variations in source parameters, sound pressure level (SPL) and intraoral pressure related to stress and emotional state. Their source data were obtained using the IAIF iterative technique of inverse filtering (Alku, 1992), and they found significant variation in the glottal wave, independent of f_0 and SPL, for different emotional states. Angry speech was included in the study by Cummings and Clements (1995) on styles of speech, which employed an inverse filtering technique based on that of Wong et al. (1979). Some further source data for different emotions, obtained by inverse filtering based on closed-phase covariance LPC, is reported by Klasmeyer and Sendlmeier (1995). Johnstone and Scherer (1999) present electroglottographic data on glottal parameters, including irregularities in fundamental period (jitter), for seven emotional states. Alter et al. (1999) present examples of estimates of the noise component, in terms of measures of the harmonic-to-noise ratio, for different

emotional states. In spite of these contributions, no clear picture emerges and our understanding of the voice source correlates of affect remains limited.

Much of what we know about the mapping of voice quality to affect has come in the form of received wisdom, based on impressionistic phonetic observations. Some of these are summarised by Laver (1980): breathy voice has been associated with *intimacy*, whispery voice with *confidentiality*, harsh voice with *anger* and creaky voice with *boredom*, for speakers of English at any rate. From the way such traditional observations are put, one would infer that a given voice quality is associated with a particular affect. On the basis of predictions from hypothesised physiological correlates of specific emotions, and of observations in a wide range of studies (mostly based on the relative strength of high versus low frequency energy in the spectrum) Scherer (1986) suggests that tense voice is associated with anger, joy and fear; and that lax voice (at the phonatory level essentially the same as breathy voice) is associated with sadness. In a similar vein, Laukkanen et al. (1996) have reported that in their data anger was characterised by low open quotient values of the glottal flow (suggesting a rather tense setting) and that sadness, surprise and enthusiasm tended to have high open quotient values, and low glottal skew values, which would indicate a more breathy setting. Not all researchers agree however on the mapping between affect and voice quality. On the basis of a wide review of literature sources, Murray and Arnott (1993) suggest very different associations: in their Table 1, breathy voice is associated with both anger and happiness; sadness is associated with a 'resonant' voice quality, which we would here interpret as a quality somewhere along the modal to tense voice continuum. It is hoped that the present study would shed some light on the nature of these associations, providing possible support for traditional assumptions, or clarifying where there is clear disagreement in the literature.

However scant the production literature, there is even less information on perception aspects. In experiments by Laukkanen et al. (1995, 1997), the role of glottal parameters on the perception of emotion were studied by manipulations to a vocalic interval (recorded with different emotions) so as to neutralise the effects of f_0, SPL and duration. They concluded that the glottal source contributes to the perception of valence as well as vocal effort. They note, however, that the type of f_0 manipulations used in their experiments may lead to certain artefacts, and suggest that synthesis would be a useful tool for further research in this area.

Synthesis offers in principle an ideal tool for examining how individual features of the signal contribute to the perception of affect, as demonstrated by experiments on f_0 and temporal parameters (e.g., Carlson et al., 1992; Mozziconacci, 1998). The lack of empirical voice source information presents a problem in the case of voice quality. Nonetheless, there have been a number of attempts to generate emotive synthetic speech, through manipulation of a large number of parameters, including voice quality. The work by Cahn (1990a,b) and by Murray and Arnott (1995) utilised the capabilities of DECtalk: in these cases problems arise from the inherent limitations of the synthesis system, in that it did not always provide adequate control of the desired parameters.

The GLOVE system, described by Carlson et al. (1991), offers a potentially high level of control, and in a study by Meurlinger (1997), the source parameters of this system were exploited in an attempt at generating synthetic speech with emotional overtones. Burkhardt and Sendlmeier (2000) describe a synthesis system for the generation of emotional speech, which uses the KLSYN88 synthesiser, which also allows direct control of many voice source parameters. They report experiments involving manipulations of f_0, tempo, voice quality and segmental features. As regards the voice quality aspects of this work, they found that falsetto voice yielded a very good response for *fear*, tense voice is associated with *anger*, falsetto and breathy voice were weakly associated with *sadness*. Results for *boredom* appeared uncertain: one experiment indicated some association with creaky or with breathy voice, but a second experiment concluded that these voice qualities reduced rather than enhanced the percept. Unfortunately, details

are not included concerning the source parameters used, nor how they were controlled to generate different qualities.

Attempts to generate emotive speech have also been made using concatenative synthesis, e.g., by Murray et al. (2000). As these systems use pre-recorded speech units, and permit very limited control of source parameters other than f_0, they are less relevant to this study. Note however the approach adopted by Iida et al. (2000), whereby recording multiple corpora with different emotional colourings provides an expanded database from which the concatenative units are drawn.

Most past research carried out in the field has tended to focus on a small set of rather strong emotions, such as anger, joy, sadness and fear. Voice quality contributes greatly to the expressiveness of human speech, and signals to the listener not only information about such strong emotions, but also about milder states, which we might characterise as feelings, moods, and general states of being. Furthermore, in an obviously related way, voice quality signals information concerning the speaker's attitude to the interlocutor, the subject matter and the situation. In this study we have tried to allow for as broad as possible a set of affective states, and therefore, the range of possible affective attributes for which responses were elicited, included not only emotions (e.g., *afraid*, *happy*, *angry*, *sad*) but also attributes that relate to speaker state and mood (e.g., *relaxed*, *stressed*, *bored*) or speaker attitude (e.g., *formal*, *interested*, *friendly*).

It is worth noting that a broad approach is also potentially more useful for downstream technology applications. A major area of application of this type of research is the provision of expressive voice in speech synthesis. If one wants to aspire to a synthesis that approximates how humans employ their capacity to vary the tone-of-voice, it makes little sense to begin by excluding much of the subject of interest. The voice modifications most frequently sought in specific synthesis applications tend to be ones pertaining to state, mood and attitude (e.g., relaxed, friendly, polite, etc.), rather than to the 'strong' emotions.

2. Voice quality and emotion: conceptual and methodological problems

This area of research presents numerous difficulties. Some of these are general ones, and pertain also to any research on the vocal features of affect communication. A fundamental problem is the lack of a widely accepted system for categorising affective states, and the potential inadequacy of English language terms, such as *angry* to represent emotional states (see discussion in Scherer, 1986). Another major difficulty has been that of obtaining emotionally coloured speech data. These aspects have been widely aired in the literature, and will not be discussed further here.

The paucity of information on the role of voice quality in communicating affect reflects the very specific additional difficulties that arise both at a conceptual level in terms of defining voice qualities, and at the methodological and technical level in obtaining reliable measures of the voice source. Firstly, most work on voice quality depends on the use of impressionistic auditory labels such as breathy, harsh, etc., which are rarely defined. The problem with impressionistic labels such as 'harsh voice' is that they can mean different things to different researchers. Thus, a given label may refer to different phenomena while different labels may be used to describe very similar phenomena, depending simply on the users' understanding of the term. The potential uncertainty can be illustrated in terms of the discussion above on voice quality correlates of emotion: where different researchers attribute very different voice qualities to an emotion (e.g., anger is associated with tense voice in Scherer, 1986 and with breathy voice in Murray and Arnott, 1993) or the same voice quality to very different emotions, it begs the question as to whether the implied differences/similarities actually relate to voice quality phenomena or arise spuriously out of a different understanding of the descriptive terms. And whereas one might expect some degree of cross-researcher consensus on how "breathy voice" or "tense voice" might be interpreted, this is unlikely for many other terms (e.g., "blaring" and "grumbled" in Murray and Arnott, 1993, Table 1).

This is a problem that besets all research in the area of voice quality, whether in normal or pathological speech (see for example the discussion in Hammarberg, 1986), or whether it is based on simple auditory/impressionistic or empirical methods. Measurements of voice source parameters in different emotions (as presented in some of the studies mentioned below) can be very difficult to interpret meaningfully if they cannot be related to the auditory impression as well as to the underlying production correlates and their spectral consequences. Laver (1980) has proposed a classification system, which is backed by physiological and acoustic data where available, which provides, in the words of Scherer (1986) "a coherent conceptual system" for voice quality research. In our earlier analyses of voice quality as in the present perceptual study, we have attempted to situate our descriptions within Laver's frame of reference, pointing out where we deviate from, or extend Laver's usage (see descriptions in Section 3).

The other major problem in this area of research is a methodological one, pertaining to the difficulty of obtaining appropriate measures of the glottal source. Direct, high fidelity recordings of the source signal would be very desirable. A technique involving miniature pressure transducers (Cranen and Boves, 1985; Kitzing and Löfqvist, 1975) inserted between the vocal folds could in principle provide this. However, the procedure involved is not only highly invasive, requiring a local anaesthetic, but may also encounter problems in transducer stability as well as possibly also interfering with the vocal production. Given the practical difficulties involved, it is not surprising that very little source data have been obtained with this technique.

Inverse filtering of the oral airflow or of the speech pressure waveform offers a non-invasive alternative. Speech production may be modelled as the convolution of the source signal and the vocal tract filter response. Inverse filtering the speech signal separates source and filter by cancelling the effects of the vocal tract, and the resulting signal is an estimate of the source. However, inverse filtering of the speech signal in order to separate source and filter is inherently difficult, as it is fundamentally an ill-posed problem. In decomposing the speech signal there are three basic elements: source, filter and speech signal. As only one of these is known (the speech signal), determining the other two is in principle not possible. Only by exploiting knowledge about the characteristics and constraints of the source and of the filter in particular is it possible to identify the likely contribution of each to the speech signal, and thus to separate the two.

Numerous fully automatic inverse filtering algorithms have been developed, most of which are based on some form of linear predictive analysis (e.g., Alku, 1992; Alku and Vilkman, 1994; Chan and Brookes, 1989; Ding et al., 1994; Fröhlich et al., 2001; Kasuya et al., 1999; Lee and Childers, 1991; Ljungqvist and Fujisaki, 1985; McKenna and Isard, 1999; Strik et al., 1992; Talkin and Rowley, 1990; Wong et al., 1979). These techniques have provided some useful information on source behaviour (e.g., Alku and Vilkman, 1996; Cummings and Clements, 1995; Laukkanen et al., 1996, 1997; Olivera, 1997; Palmer and House, 1992; Strik and Boves, 1992). However, automatic techniques tend to perform least well when there is no true closed phase to the glottal cycle and where automatic estimation of formant peaks is least reliable, as is the case for many non-modal voice qualities.

A further problem concerns how to effectively measure parameters from the glottal signal. There is no single set of clearly defined source parameters that have been generally adopted, which makes comparisons difficult. Furthermore, estimating values for salient parameters from the inverse filtered signal typically involves some level of compromise, as critical timing and amplitude events of the glottal pulses are not always clear-cut. How to get optimal measures from the inverse filtered signal is therefore often not self-evident.

In some techniques source and filter parameters are estimated simultaneously (e.g., Fröhlich et al., 2001; Kasuya et al., 1999; Ljungqvist and Fujisaki, 1985), but often the parameters are measured from the estimated source signal. This can be done directly from the waveform, thus using only time domain information, but more common is perhaps the technique of adjusting a parametric source model in order to capture the characteristics

of the glottal pulses obtained from the inverse filtering.

The model matching technique has the advantage of allowing for both time and frequency domain optimisation of the parameters, as well as providing suitable data for synthesis. However, parameterising data in this way will to some extent depend on the model used. Numerous source models have been proposed in the literature (e.g., Ananthapadmanabha, 1984; Fant, 1979a,b, 1982; Fant et al., 1985; Fujisaki and Ljungqvist, 1986; Hedelin, 1984; Klatt and Klatt, 1990; Price, 1989; Qi and Bi, 1994; Rosenberg, 1971; Rothenberg et al., 1975; Schoentgen, 1993; Veldhuis, 1998). However, the four-parameter LF model of differentiated glottal flow (Fant et al., 1985) seems to be emerging as the main model employed in analytic studies. This model also benefits from being incorporated within available synthesisers, such as the KLSYN88 (Klatt and Klatt, 1990). It is clearly an advantage if the same source model can be used in both analysis and synthesis. In the present study, this is the model used and it is also the model we have hitherto used in our analyses of voice source variation.

Several automatic procedures for model matching exist. Some of them optimise the fit in the time domain (e.g., Jansen, 1990; Strik et al., 1993; Strik and Boves, 1994) and others employ frequency domain optimisation (e.g., Olivera, 1993). Some of the techniques have been evaluated on synthesised speech, where they seem to perform reasonably well. Nevertheless, obtaining robust and fully reliable source estimates from natural speech still seems to be a problem (Fröhlich et al., 2001). As with the automatic inverse filtering techniques, the problems are likely to be worse again when dealing with non-modal voice qualities, particularly those with glottal pulse shapes substantially different from what can be generated by the source model. Given the potential for producing large amounts of data, the problems of robustness may, at least in part, be an explanation for the surprisingly small body of source data on different voice qualities reported in the literature using these automatic techniques.

Interactive manual techniques for inverse filtering and parameterisation offer a way of overcoming the problem of robustness, but have their own limitations (Carlson et al., 1991; Hunt et al., 1978; Ní Chasaide et al., 1992; Gobl and Ní Chasaide, 1999b). Given that subjective judgements are involved, it requires considerable expertise and knowledge on the part of the experimenter if results are not to be spurious. Across highly experienced experimenters, it seems that a high degree of consistency can be achieved (Scully, 1994). Similar findings have also been reported by Hunt (1987). The main limitation, however, of this technique is that it is extremely time-consuming, and is thus only suitable for the analysis of limited amounts of data. Notwithstanding, micro-studies involving such manual techniques have afforded useful insights into inter- and intra-speaker voice source variation (e.g., Fant, 1995; Gobl, 1988, 1989; Gobl et al., 1995; Gobl and Ní Chasaide, 1992; Hertegård and Gauffin, 1991; Kane and Ní Chasaide, 1992; Karlsson, 1990, 1992; Karlsson and Liljencrants, 1996; Ní Chasaide and Gobl, 1993; Pierrehumbert, 1989; Scully et al., 1995).

Indirect techniques such as electro-glottography (EGG) have been also used by, e.g., Johnstone and Scherer (1999) and Laukkanen et al. (1996) and can offer many useful insights. But insofar as the technique registers contact across the vocal folds, data are difficult to interpret when the vocal folds do not meet or have reduced contact during the 'closed' phase (see Laukkanen et al., 1996, for a discussion on the difficulties with EGG in analysing source parameters, and for a comparison with inverse filtering).

Measures from the speech output spectrum can provide useful insights into aspects of voice quality. For instance, the comparison of the amplitude levels of H1 and F1 or of H1 and H2, have been frequently used in the phonetics and linguistics literature to make inferences on source behaviour. Johnstone and Scherer (1999) have used these types of measures specifically for the analysis of voice quality and emotion. Note however that the levels of the output spectrum reflect filter as well as source characteristics, and thus measures are potentially problematic (for further discussion on this, see Ní Chasaide and Gobl, 1997). The relative balance of higher versus lower frequencies measured in the long term average spectrum can also

be useful, particularly for differentiating voice quality variation in the tense–lax dimension (see observations of Scherer, 1986, also discussed above). Although these measures are in themselves useful, they provide only a gross indication of what is a multifaceted phenomenon. Furthermore, with regard to the synthesis of voice quality variation, they are not likely to be readily incorporated into current synthesis systems.

3. Experimental procedure

As mentioned in Section 1, the purpose of the experiment was to explore the role of voice quality in the communication of emotions, moods and attitudes, by testing listeners' reactions to an utterance synthesised with different voice qualities. The basic procedure involved the recording of a natural utterance, which was analysed and parameterised in order to facilitate the resynthesis of an utterance with modal voice quality. Parameter settings for this synthetic stimulus were modified to generate the six non-modal voice quality stimuli. The seven stimuli were then used in a set of perception tests to elicit listeners' responses to the affective content of the stimuli.

3.1. Voice qualities

In this pilot experiment on the perceived affective correlates of a selection of stimuli synthesised with different voice qualities, we tried as far as possible to capture the characteristics of particular targeted voice qualities. These included five qualities for which earlier analyses had been carried out—modal (neutral) voice, tense voice, breathy voice, whispery voice and creaky voice—and two additional qualities—harsh voice and lax–creaky voice.

The physiological correlates of voice quality are described by Laver (1980) in terms of three parameters of muscular tension: *adductive tension* (the action of the interarytenoid muscles adducting the arytenoids), *medial compression* (the adductive force on the vocal processes adducting the ligamental glottis) and *longitudinal tension* (the tension of the vocal folds themselves).

In Laver's system, modal voice is characterised as having overall moderate laryngeal tension. Vocal fold vibration is efficient and the ligamental and the cartilaginous parts of the glottis are vibrating as a single unit. Tense voice is described as having a higher degree of tension in the entire vocal tract as compared to a neutral setting. At the laryngeal level, adductive tension and medial compression are thought to be particularly implicated. Breathy voice involves minimal laryngeal tension. Vocal fold vibration is inefficient and the folds do not come fully together, resulting in audible frication noise. Whispery voice is characterised by low tension in the interarytenoid muscles, but a fairly high medial compression, resulting in a triangular opening of the cartilaginous glottis. Laryngeal vibration is very inefficient and is accompanied by a high degree of audible frication noise.

Harsh voice involves very high tension settings. To this extent it is essentially a variety of tense voice, but may have more extreme settings. A defining characteristic is that harsh voice tends to have additional aperiodicity due to the very high glottal tension. In the present experiment, as we were interested to focus on the specific role of the aperiodic component, we have only manipulated this parameter, and retained the remaining source parameter settings of tense voice.

Creaky voice is described as having high medial compression and adductive tension, but low longitudinal tension. Because of the high adductive tension, only the ligamental part of the glottis is vibrating. The quality which is here termed 'lax–creaky' voice is not included in the system presented by Laver (1980), where creaky voice is described as having rather high glottal tension (medial compression and adductive tension). In our descriptive work referred to earlier, it was indeed found that creaky voice has source parameter values tending towards the tense. It was also our auditory impression that creaky voice, as produced by the informant in question, did have a rather tense quality. Yet, we are aware that creaky voice can often sound quite lax in auditory impressionistic terms. It is for this reason that a lax–creaky quality was included, which is essentially based on breathy voice source settings but with

reduced aspiration noise and with added creakiness. Although this lax–creaky voice quality to some extent runs counter to the general thrust of Laver's description for creaky voice, it is worth noting that some of the sources he cites imply a rather lax glottal setting (e.g., Monsen and Engebretson, 1977). Clearly more descriptive work on creaky voice is required both at the physiological and acoustic levels.

3.2. Speech material

The starting point for generating the synthetic voice quality stimuli was a high quality recording of a Swedish utterance, "ja adjö" [ˈjɑː aˈjøː], where f_0 peaks were located on the two stressed vowels. This utterance should be semantically neutral to our subjects, native speakers of Irish English who do not speak Swedish. The male speaker's voice was judged by the authors to be in reasonable conformity with modal voice as described by Laver (1980).

The recording was carried out in an anehoic chamber, using a Brüel & Kjær condenser microphone at a distance of approximately 30 cm from the speaker. The utterance was recorded on a SONY F1 digital tape recorder, and no filters were employed so as to avoid introducing phase distortion. The recording was subsequently transferred to computer, digitised at 16 kHz sampling frequency and 16 bit sample resolution. At this point, the recording was high-pass filtered in order to remove any DC offset of the zero-pressure line, due to the inevitable intrusion of some inaudible low frequency pressure fluctuations into the anechoic chamber. The filter used was a third order digital Butterworth filter with a cutoff frequency of 20 Hz, and to ensure phase linearity, the speech signal was passed through this filter twice, the second pass being time-reversed (i.e. starting with the last sample, finishing with the first).

3.3. Analysis

The analysis technique involved source filter decomposition and source model matching using the software system described in (Ní Chasaide et al., 1992). This system incorporates automatic or semi-automatic inverse filtering based on closed-phase covariance LPC. Further, optional, manual interactive analysis can subsequently be carried out if deemed necessary. As the amount of data here was limited to one short utterance, all the 106 pulses of the utterance were inverse filtered using the interactive technique.

For this speaker there were 9 formants present in the output, within the 8 kHz frequency range determined by the sampling rate. Thus 9 anti-resonances were used in the inverse filter to cancel the filtering effect of the vocal tract.

The output of the inverse filter yields an estimate of the differentiated glottal flow. From this signal, data on salient source parameters were obtained by matching a parametric voice source model to the differentiated glottal flow signal. As mentioned in Section 2, the model we use is the four-parameter LF model of differentiated glottal flow (Fant et al., 1985).

For similar reasons as for the inverse filtering, the fitting of the LF model to the 106 glottal pulses was done manually, using an interactive technique which facilitates parameter optimisation in terms of both time and frequency domain aspects of the glottal pulse. As the objective here was to generate good copy synthesis of the utterance, the disadvantages of the manual technique were of minor importance.

On the basis of the modelled waveform the principle parameters measured were EE, RA, RG and RK, which are briefly glossed here (for a fuller description see, e.g., Fant and Lin, 1991; Ní Chasaide and Gobl, 1997). EE is the excitation strength, measured as the amplitude of the differentiated glottal flow at the main discontinuity of the pulse. The RA value is a measure that corresponds to the amount of residual airflow after the main excitation, prior to maximum glottal closure. RG is a measure of the 'glottal frequency', as determined by the opening branch of the glottal pulse, normalised to the fundamental frequency. RK is a measure of glottal pulse skew, defined by the relative durations of the opening and closing branches of the glottal pulse.

3.4. Synthesis of the modal voice stimulus

The KLSYN88a synthesiser (Sensimetrics Corporation, Boston, MA, see also Klatt and Klatt, 1990) was chosen for the generation of the voice quality stimuli. This is a well established formant synthesiser which allows for direct control of both source and filter parameters, and it has been shown to have the capability of producing high quality copy synthesis (Klatt and Klatt, 1990). As mentioned earlier, it also incorporates the LF voice source model (as an option), albeit in a somewhat modified implementation.

To generate the modal stimulus, copy synthesis of the natural utterance was carried out using the data from the analysis. In the synthesiser, the modified LF model was selected for the voice source generation. In order to carry out the synthesis, the LF parameters of the analyses were transformed into the corresponding source parameters of KLSYN88a. It should be noted that care has to be taken when transforming parameters derived from the LF model (in this case EE, RA, RG and RK) into the corresponding parameters for the modified LF model of KLSYN88a: AV (amplitude of voicing, derived from EE), TL (spectral tilt, derived from RA and f_0), OQ (open quotient, derived from RG and RK), SQ (speed quotient, derived from RK). See Mahshie and Gobl (1999) for details on the differences between the LF model and the version of the model in KLSYN88a.

As there was no practical way of entering the data for all 106 pulses into the synthesiser, the input data were reduced by selecting values at specific timepoints for each parameter (the number of values ranging between 7 and 15, depending on the parameter). The timepoints were chosen so that the linear interpolation generated by the synthesiser between selected points would capture the natural dynamics as closely as possible. The stylisation is somewhat similar to that carried out by Carlson et al. (1991), who used the GLOVE synthesiser for the copy synthesis of a female utterance. However, they did not extract the data from a pulse-by-pulse analysis, but rather used data from a small number of analysed pulses, selected on the basis of the segmental structure.

Initial attempts to synthesise at a sampling rate of 16 kHz were unsuccessful, due to unpredictable behaviour of the synthesiser. Thus, the synthesiser's default sampling rate of 10 kHz was opted for, which seemed to ensure a reliable output. The default setting of 5 ms for the update interval of parameter values was also used. In the natural utterance, there were 6 formants present in the output spectrum below 5 kHz, and thus 6 formant resonators were used in the synthesis.

14 synthesis parameters were varied dynamically. The vocal tract parameters varied included the first five formant frequencies (F1, F2, F3, F4, F5) and the first and second formant bandwidths (B1, B2). Seven source parameters were varied: fundamental frequency, AV, TL, OQ, SQ, AH (aspiration noise) and DI ('diplophonia'—used for the generation of creakiness).

The AH parameter controls the level of the aspiration noise source. This aperiodic source is produced by a pseudo-random number generator, with an even amplitude distribution within the range of 16 bit amplitude representation. The amplitude spectrum (when combined with the filter modelling the radiation characteristics at the lips) is essentially flat above 1 kHz. Below 1 kHz the amplitudes gradually drop off so that the level is approximately 12 dB lower at 100 Hz relative to the level above 1 kHz. When AV is non-zero (i.e. when there is voicing and aspiration simultaneously) the amplitude of the aspiration noise is modulated: for the second half of the period from one glottal opening to the next, the amplitudes of all noise samples are reduced by 50%. This modulation is always the same regardless of the particular glottal pulse shape, but the result is generally that stronger aspiration is produced in the open portion of the glottal cycle relative to the closed portion (Klatt, 1980; Klatt, unpublished chapter; Klatt and Klatt, 1990).

The DI parameter alters every second pulse by shifting the pulse towards the preceding pulse and at the same time reducing the amplitude. The shift as well as the amount of amplitude reduction is determined by the DI value. Thus, the fundamental period with respect to the preceding pulse is reduced, which results in an equivalent increase

in the fundamental period with respect to the following pulse (Klatt and Klatt, 1990).

The resulting synthesis of the natural utterance is a very close replica of the original, but it is of course not indistinguishable from it, given the data reduction procedure that was carried out. More importantly, however, the voice quality of the original was retained, and thus this synthesised utterance was used as our modal voice stimulus.

3.5. Synthesis of non-modal stimuli

On the basis of the modal voice stimulus, six further stimuli were generated with non-modal voice qualities by manipulating eight parameters: the seven source parameters mentioned above and the first formant bandwidth, B1.

The transforms from modal to a non-modal quality were typically not constant for any given parameter, but allowed for dynamic variation partly prompted by earlier analytic studies, e.g., allowing for differences that relate to stress variation and voice onset/offset effects (Gobl, 1988; Ní Chasaide and Gobl, 1993). Parameter values for the different voice qualities were guided by prior analytic studies (e.g., Gobl, 1989; Gobl and Ní Chasaide, 1992; Ní Chasaide and Gobl, 1995). However, as the auditory quality was the main goal here, settings were ultimately determined by auditory judgement of the effect. This was particularly the case for the settings of parameters AH and DI, for which quantitative data were not available. Fundamental frequency was varied only to the extent deemed required as part of the intrinsic, voice quality determined characteristics. The main changes carried out to the control parameters for the different stimuli are summarised below, whereas full details on the parameter dynamics can be found in Fig. 1.

Compared to modal voice, tense voice involved lower OQ, higher SQ, lower TL, narrower B1 and slightly higher f_0 values (5 Hz). Breathy voice, again relative to modal, involved lower AV, higher OQ, lower SQ, higher TL, and wider B1 settings. The level of AH was set on the basis of auditory judgement. Creaky voice was based on modal voice, with a basic f_0 lowering of 30 Hz, but for the first f_0 peak this lowering was gradually reduced to 20 Hz. The baseline value for the DI parameter was set to 25%, changing gradually to 5% to coincide with the f_0 peaks of the stressed vowels.

The lax–creaky voice quality involved modifications to the source settings for the breathy voice stimuli. As mentioned above, this quality departs from the definitions presented in (Laver, 1980). However, to maintain some link with the physiological adjustments he proposes for creaky voice, the source settings for lax–creaky voice were modified from the breathy voice ones, by changing the OQ values to those of creaky voice. Further changes involved lowering f_0 by 30 Hz and reducing AH by 20 dB. The baseline value for the DI parameter was set to 25%, changing gradually to 15% to coincide with the f_0 peaks of the stressed vowels. The resulting stimulus was judged auditorily by the authors as a realistic reproduction of the type of lax–creaky voice discussed above.

To synthesise harsh voice, the same basic source settings as tense voice were adopted. Aperiodicity was added by using the DI parameter, although it is not clear whether this form of aperiodicity is optimal for synthesising harsh voice. However, using a baseline value of 10% gradually changing to 20% to coincide with the f_0 peaks of the stressed vowels, seemed to result in a reasonably convincing harsh voice quality.

Whispery voice turned out to be the most problematic quality to synthesise. The first attempt was based on breathy voice settings, modified so that AV was relatively lowered, AH increased, OQ slightly lowered and SQ slightly increased. Although these transformations are in keeping with analytic data, they resulted in a very unconvincing voice quality, where the aspiration noise was unnaturally high-pitched with a "whistling" quality. Widening higher formant bandwidths only marginally improved the quality. In order to achieve an acceptable whispery voice quality, it was necessary to reduce the number of formants from six to five. By thus reducing the amplitude of the aspiration noise in the higher end of the spectrum, the whistling quality was avoided. The DI parameter was set to 5% throughout.

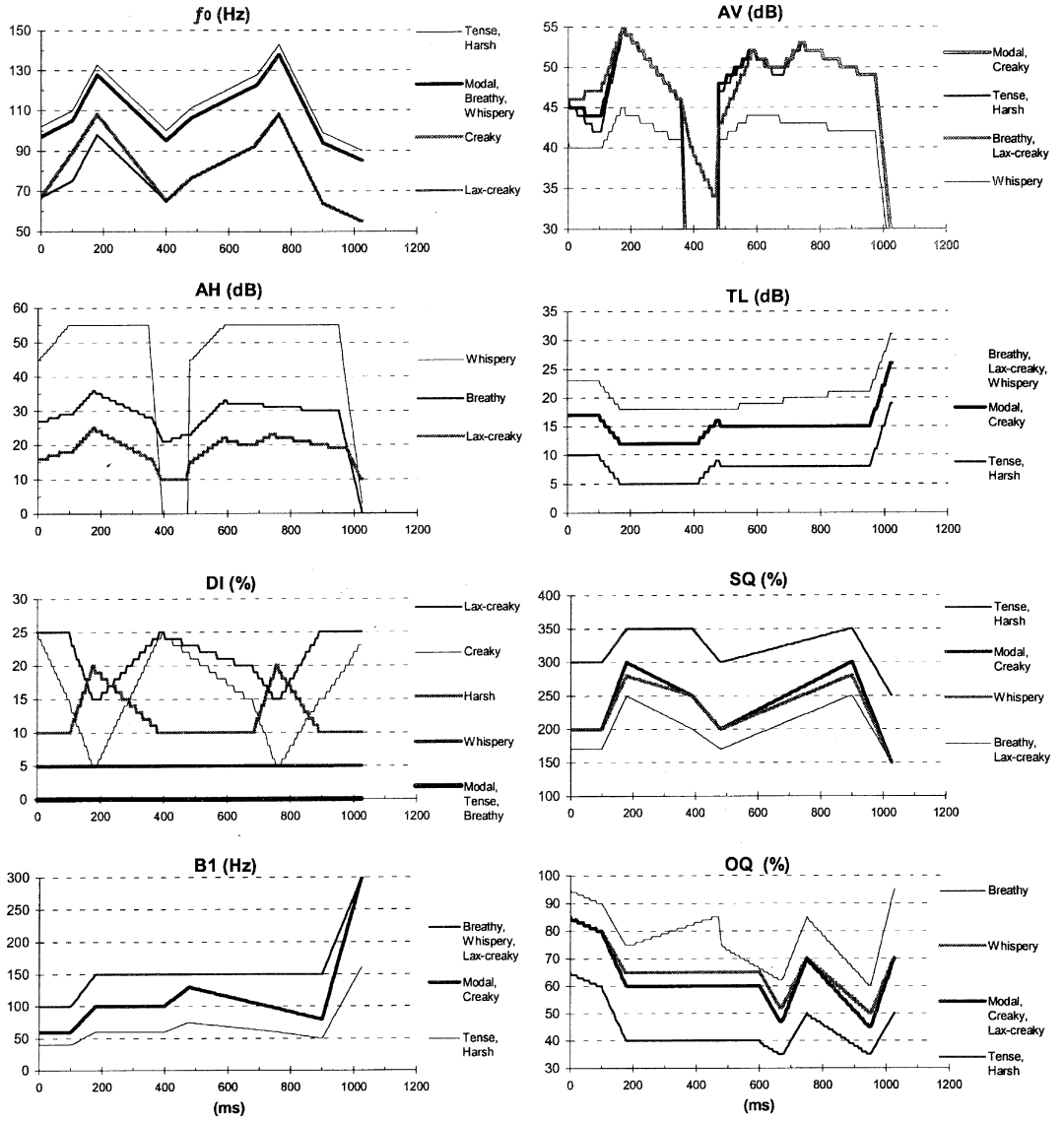

Fig. 1. Parameter variation for the synthetic stimuli. Note that for the modal, tense, harsh and creaky stimuli, there was no aspiration noise (AH).

3.6. Perception test

The perception experiment consisted of 8 short sub-tests. For each sub-test, 10 randomisations were presented of the seven stimuli (modal, tense, breathy, whispery, creaky, harsh and lax–creaky voice). The interval between each set of stimuli was 7 s, and the onset of each group was signalled by a specific earcon. Within each set of stimuli, the interstimulus interval was 4 s and a short tone was presented 1 s before each stimulus, to ensure the listener was in a state of readiness. For each individual sub-test, responses were elicited only for one particular pair of opposite affective attributes (such as *bored/interested*) in a way that was loosely modelled on Uldall (1964). 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 instructed 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. In the case where an utterance was not considered to be marked for either of the pair of attributes, they were instructed to choose the centre box. 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*, *timid/confident* and *afraid/unafraid*.

The test was administered to 12 subjects, 6 male and 6 female. All were speakers of Southern Irish English, living in Dublin and their ages ranged from early 20s to late 40s. Most of the subjects were university staff or students, and the remainder were professional people. Whereas a few subjects had a knowledge of phonetics, none had previously been involved in a perception experiment involving voice quality. The test was presented in a soundproofed studio, over high-quality studio loudspeakers which were set at a level deemed to be comfortable listening level. A short break was given between each sub-test.

4. Results

A 2-way ANOVA was carried out on the listener's scores for each of the 8 sub-tests, where voice quality and subject were the factors. Results show that the voice quality and subject variable were statistically highly significant and that there was a voice quality/subject interaction. For the majority of attribute pairs tested, the differences between the individual voice qualities were statistically significant, and the significance levels for each pairwise comparison for each sub-test are shown in Table 1. The multiple comparison technique used was Tukey's Honestly Significant Difference; this was implemented in MINITAB (Minitab, 2001). The overall mean ratings obtained for the different affective attributes with each of the stimulus types is shown in Fig. 2, along with median values. To provide an indication of the cross-subject variability, the interquartile range of subjects means and extreme values are also plotted. To make for easier broad comparisons across voice qualities and across affective attributes, the mean scores only are shown in Fig. 3. In both Figs. 2 and 3 the distance from 0 (no affective content) indicates the strength with which any attribute was perceived. The use of positive and negative values in the y-axis of the figure is not in itself important: results have simply been arranged in these figures so that the positive (or negative) sign groups together somewhat related attributes. Although there is no necessary connection between individual affective attributes, rating values across attributes are joined by lines in Fig. 3 for each of the voice qualities, to make it easier to relate individual voice qualities to their affective correlates. A subset of this information is shown in a slightly different format in Fig. 4, where the maximum strength (the highest mean score) with which each of the attributes was detected across all voice qualities is shown as deviations from 0 ($=$no perceived affect) to 3 (i.e. $\pm 3 =$ maximally perceived). The estimated standard error of the mean is also shown.

Clearly, not all affective attributes are equally well signalled by these stimuli. From Fig. 4 we see that the most readily perceived ones are *relaxed* and *stressed*, and high ratings are found for *angry*, *bored*, *intimate*, *content*, *formal* and *confident*. The least readily perceived are *unafraid*, *afraid*, *friendly*, *happy* and *sad*. By and large, those affective attributes which got high scores in this test are more aptly described as states, moods or attitudes (the exception being *angry*), whereas those least well detected tend to be emotions.

As can be seen in Figs. 2 and 3, the individual stimuli are not associated with a single affective attribute: rather they are associated with a constellation of attributes. Thus, tense/harsh voice gets high ratings not only for *stressed*, but also for *angry*, *formal*, *confident* and *hostile*. The broad picture to emerge is of two groups of voice qualities, which signal polar opposite clusters of attributes (see Fig. 3). The stimuli for tense/harsh voice

Table 1
Significance level of the difference in ratings for each pair of stimuli, shown for each of the eight sub-tests

	Tense	Breathy	Whispery	Harsh	Creaky	Lax–creaky
Modal						
Relaxed–stressed	***	***	***	***	***	***
Content–angry	***	***	***	***	***	***
Friendly–hostile	***	***	***	***	***	***
Sad–happy	***	***	***	***	***	***
Bored–interested	***	***	***	***	***	***
Intimate–formal	***	***	***	***	***	***
Timid–confident	***	***	***	***	***	***
Afraid–unafraid	0.12	***	***	0.73	***	***
Tense						
Relaxed–stressed		***	***	0.25	***	***
Content–angry		***	***	0.59	***	***
Friendly–hostile		***	***	1.00	***	***
Sad–happy		***	***	0.33	***	***
Bored–interested		***	***	0.99	***	***
Intimate–formal		***	***	0.98	***	***
Timid–confident		***	***	0.11	***	***
Afraid–unafraid		***	***	0.91	***	0.05
Breathy						
Relaxed–stressed			1.00	***	0.20	***
Content–angry			0.19	***	0.98	***
Friendly–hostile			1.00	***	***	0.18
Sad–happy			0.88	***	0.57	***
Bored–interested			0.99	***	***	***
Intimate–formal			0.41	***	**	***
Timid–confident			***	***	***	***
Afraid–unafraid			***	***	***	***
Whispery						
Relaxed–stressed				***	0.25	***
Content–angry				***	*	***
Friendly–hostile				***	**	0.06
Sad–happy				***	0.05	**
Bored–interested				***	***	***
Intimate–formal				***	***	**
Timid–confident				***	***	***
Afraid–unafraid				***	***	***
Harsh						
Relaxed–stressed					***	***
Content–angry					***	***
Friendly–hostile					***	***
Sad–happy					***	***
Bored–interested					***	***
Intimate–formal					***	***
Timid–confident					***	***
Afraid–unafraid					***	**
Creaky						
Relaxed–stressed						***
Content–angry						***
Friendly–hostile						***
Sad–happy						***

(*continued on next page*)

Table 1 (*continued*)

	Tense	Breathy	Whispery	Harsh	Creaky	Lax–creaky
Bored–interested						***
Intimate–formal						***
Timid–confident						***
Afraid–unafraid						0.12

* $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$.

are associated with the cluster of features just mentioned, which we might broadly characterise as involving high activation/arousal and/or high control. On the other hand, the stimuli for breathy voice, whispery voice, creaky voice and lax–creaky voice are by and large associated with opposite, low activation characteristics, shown with negative values in Fig. 3.

The modal stimulus, used as the starting point for the other qualities does not turn out to be fully neutral: as can be observed in Figs. 2 and 3, responses veer somewhat in the direction of tense voice for a number of attributes, namely *confident*, *formal* and *stressed*, although not to any great degree.

Distinct responses were not obtained for all synthesised qualities. Results for the tense and harsh stimuli are very similar, with the tense eliciting in all cases slightly more extreme ratings. The difference is very small, and not significant for any of the attribute rating sub-tests (see Table 1). Furthermore, what difference there is runs counter to initial expectations, which were that the addition of aperiodicity to tense voice should heighten the effects of tense voice rather than attenuate them. Caution is needed however in interpreting this result for harsh voice, as it may be more a reflection on the synthetic stimulus than a reliable indication of how listeners judge harsh voice per se (see further discussion on this below).

The breathy and whispery stimuli also yield very similar response patterns, and the difference between them is only significant for the attributes *afraid* and *timid* (Table 1), where whispery voice achieves stronger ratings (Fig. 3). In the case of whispery voice, results also need to be interpreted with some caution for reasons mentioned earlier, concerning the difficulty of synthesising this quality. Furthermore, it may be that whispery voice needs to be more distinctly different from breathy voice than was achieved in the present stimulus.

Ratings for the creaky voice stimulus tend to be on the whole close to those of the breathy and whispery stimuli, although the differences are generally significant (see Table 1). The most striking divergence is found for the attributes *afraid* and *timid* (Fig. 3). Responses to lax–creaky voice follow the same trends as creaky voice, but are more extreme: as can be observed in Fig. 3, the trend of responses is very similar but is shifted towards the non-aroused, low activation end of the scale. The differences between responses for the creaky and lax–creaky stimuli are highly significant (Table 1) for all attributes except *afraid–unafraid*, where neither yields a strong response. Broadly speaking, it would appear that the addition of more lax settings to the creaky voice stimulus results in a considerable enhancement of its intrinsic colouring.

It is rather striking in this experiment that the highest ratings for most of the affective attributes tested were obtained by just two of the range of stimuli presented. The tense stimulus accounted for the highest ratings for attributes with high arousal/activation and high power/control, whereas the lax–creaky stimulus obtained generally highest ratings for attributes with low arousal/activation. A third stimulus, whispery voice, produced the highest ratings for the attributes *timid* and *afraid*, but note that responses for *afraid* in particular are not very high, and show considerable cross-subject variability. It furthermore appears to be the case that as one moves from the high activation to the low-activation group of stimuli, there is an increase in cross-subject variability (Fig. 2).

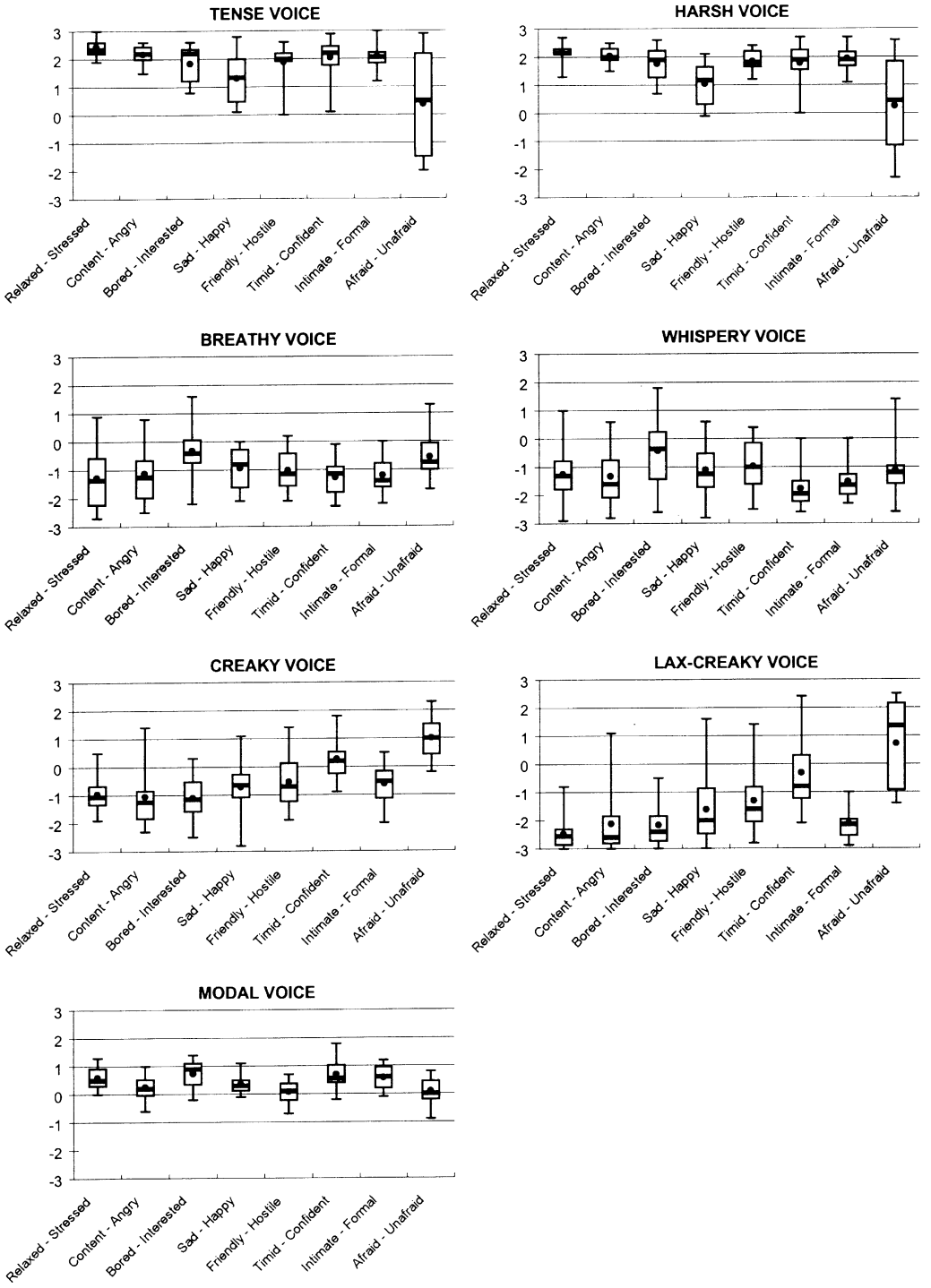

Fig. 2. Subjects' mean responses for each voice quality stimulus, in each of the eight sub-tests, showing interquartile range (box); mean (filled circle); median (horizontal line in box) and extreme values (whiskers).

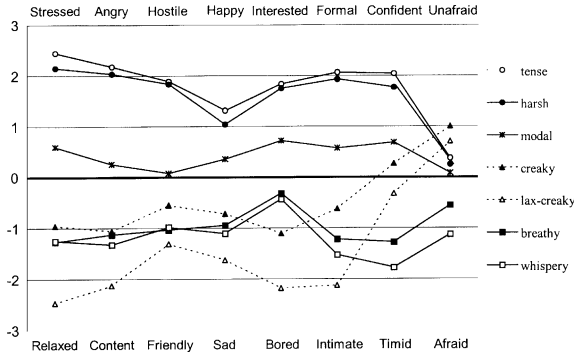

Fig. 3. Mean ratings for 12 listeners of the perceived strength of pairs of attributes for seven voice qualities. 0 = no affective content and ±3 = maximally perceived.

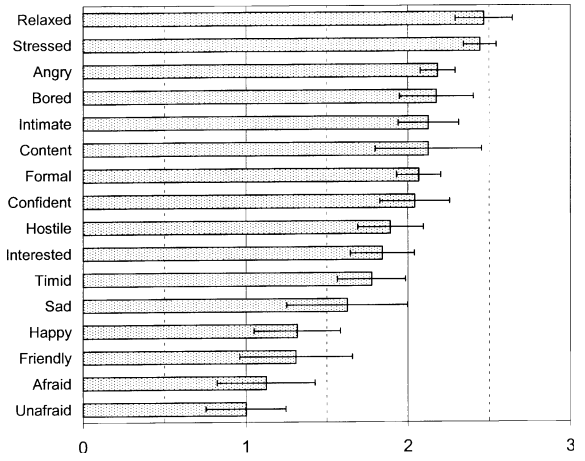

Fig. 4. Maximum mean ratings for 12 listeners of the perceived strength of each affective attribute, shown by the bars as deviations from 0 (no affective content) to 3 (maximally perceived). The lines through the bars indicate ± the estimated standard error of the mean.

5. Discussion

The results demonstrate that voice quality changes alone can evoke differences in speaker affect. They also show that unlike the one-to-one mapping often implied by traditional impressionistic observations, a specific voice quality is multicoloured in terms of affect, being associated with a cluster of mostly, though not necessarily, related attributes. It has been suggested (see, for example, Laukkanen et al., 1996, 1997) that voice quality may serve more to communicate the valence of an emotion rather than its activation, which would depend rather on pitch, loudness and duration. In the case of the qualities represented by the present stimuli, the differentiation appears not to be in terms of valence but rather activation, and to a lesser extent, power. The attributes associated with the tense/harsh stimuli have high activation and/or high power, but include affects with positive (*confident*, *interested*, *happy*) and negative (*angry*, *stressed*) valence. The other, non-modal group of stimuli—the breathy, whispery, creaky and especially the lax–creaky voiced stimuli—are associated with attributes which have low activation but both positive (*relaxed*, *content*, *intimate*, *friendly*) and negative (*sad*, *bored*) valence.

As a preface to the following discussion we would stress certain limitations of this study. Firstly, the reader should bear in mind that results tell us about voice quality in the human communication of affect only insofar as the synthesised stimuli are good approximations of the intended voice qualities. Secondly, voice qualities vary in a continuous, not a categorical fashion: there can be differing degrees of say breathy voice or tense voice: by choosing single points in these continua, we are only exploring to a limited extent what the role of a particular quality such as tense voice may be in affect signalling. A question for future research will be to look at how more gradient changes in source parameters relate to the associated affect. For example, if parameters associated with tense voice are varied in a more continuous fashion, will this yield correspondingly different degrees of anger? Alternatively, it is not inconceivable that one might find different affective correlates, such as *happy* and *angry* for different parts of the continuum. Finally, we would point out that the qualities investigated here are only a partial sampling of the voice quality types that speakers may use in affect signalling. In all these senses, this study must be viewed as an exploratory exercise.

We look now at whether the associations of voice quality and affective states traditionally assumed, or mentioned in the literature, are supported by the results for the range of synthesised stimuli in this study. Breathy voice has tradition-

ally been thought to have connotations of intimacy (Laver, 1980). The present results suggest that although the breathy stimulus did have some such colouring, the percept was much more effectively signalled by the lax–creaky stimulus.

In his review of earlier studies, Scherer (1986) has suggested that lax voice (i.e., breathy voice at the phonatory level) would be associated with sadness. The results of Laukkanen et al. (1996) also point to such an association. Whereas the breathy stimuli did achieve a somewhat *sad* response in this study, the effect was not very strong, and ratings for this attribute were also considerably higher for the lax–creaky stimulus. Note however in Fig. 2, that there is more cross-subject variability in ratings for the latter quality. The large difference in total range of responses for the lax–creaky case reflects the fact that one of the twelve subjects responded very differently from the others, and perceived this stimulus as moderately *happy*.

Very different suggestions linking breathy voice with anger and happiness are presented in the literature summary by Murray and Arnott (1993, Table 1). These associations are not supported by present results: in both cases, listeners rated the breathy stimulus as being associated rather with the opposite attributes.

To sum up on results for the breathy voiced stimulus in this experiment: there is some support for past suggestions linking breathy voice with intimacy and sadness, none for a link with anger or happiness. Even in the case of intimacy and sadness, the response rates obtained here were not particularly high, and not at all as high as for the lax–creaky stimulus. Furthermore, for both these stimuli, response rates for *sad* or *intimate* were at about the same levels as for other attributes, such as *content* and *relaxed*.

In his review of the literature, Scherer (1986) associates tense voice with anger, and also with joy and with fear. Laukkanen et al. (1996) also found an association between anger and source parameter values that would indicate tense voice, and a similar association would also be indicated by Burkhardt and Sendlmeier (2000). The association of tense voice with anger is strongly supported in the present results. As can be seen in Fig. 2, responses are high and show little variability across subjects.

The association of tense voice with joy finds some support in that there is a moderate colouring towards *happy* in responses for the tense stimulus, which is nonetheless significant, as a comparison of tense versus modal stimuli in Table 1 and Fig. 2 indicates. A comparison of the *happy* and *angry* responses for the tense stimulus in Fig. 2 shows not only that mean ratings for the former are lower, but also that they vary more across subjects. Nevertheless, the tense stimulus was the only one of the present set that yields a *happy* connotation (except for harsh voice, which is not here differentiated from tense).

The association of tense voice with *fear* as suggested by Scherer (1986) is not supported here, as mean responses for the tense stimulus are close to zero for this attribute. Furthermore, there was very high cross-listener variability in *fear* ratings for the tense stimulus: compare, for example with responses for the modal stimulus in Fig. 2, where the mean is also close to zero but there is more agreement across subjects. The whispery voice stimulus provides the strongest responses for fear, but note that *fear* is nevertheless one of the least well signalled attributes in this experiment. Furthermore, as can be observed in Fig. 2, not all listeners necessarily associate whispery voice with *fear*. Burkhardt and Sendlmeier (2000) report that falsetto voice (not included in this study) is a successful voice quality for portraying fear. One might conjecture that some type of whispery falsetto voice with appropriate aperiodicity would be an effective quality for portrayals of *fear*.

Laver (1980) has suggested that harsh voice (a variety of tense voice) is associated with anger. As mentioned earlier, the fact that listeners did not differentiate between the harsh and tense stimuli in this test probably reflects the similarity between them, the only difference being the addition of aperiodicity (as controlled by the DI parameter in KLSYN88a) in the former. In order to produce a well-differentiated harsh voice, it may be the case that a greater degree of aperiodicity would be required and/or a different type of aperiodicity. Furthermore, harsh voice may also require more

extreme settings of those parameters that reflect glottal tension. Although in this test, more extreme tension settings were not adopted, this is something we would hope to include in further tests.

In Murray and Arnott's summary, sadness is associated with a resonant voice quality (Murray and Arnott, 1993, Table 1). We would interpret the term resonant to be a quality somewhere on the modal–tense continuum. As can be seen in Fig. 2, neither the tense nor the modal stimuli elicited a *sad* response: as mentioned above, the shift from modal to tense enhanced the *happy* rather than the *sad* overtones.

To sum up on tense voice: the present study provides strong support for the association of anger with tense voice. The linkage between tense voice and anger is hardly surprising, being intuitively to be expected, and probably the most widely suggested association of voice quality–affect one finds in the literature. There is also some support for some degree of association of tense voice with joy, as suggested by Scherer (1986). Other previously suggested associations of tense voice with fear or with sadness are not supported here. Note, however, that in the present study, a number of further strong associations with tense voice are suggested. The fact that these have not been previously reported may simply relate to the fact that the overwhelming focus of past studies in this field has been on the 'strong' emotions—*anger*, *joy*, *fear*, *sadness*. The tense stimulus in this experiment yielded very high ratings for *stressed*, *formal*, *confidant*, *hostile* and *interested*. And whereas the attributes *stressed* and *hostile* are clearly very related to *angry*, others such as *confident*, *formal* and *interested* appear to be rather different in terms of valence, power and even degree of activation.

It tends to be taken as axiomatic that creaky voice signals boredom for speakers of English (see Laver, 1980). In the present experiment, high response rates were achieved by the lax–creaky stimulus, which combines features of creaky and breathy voice. It is worth noting (Figs. 2 and 3) that this stimulus is considerably more potent in signalling boredom than the creaky voice stimulus which was modelled on Laver's (1980) specification of creaky voice, and on our own earlier analyses of creaky voice (e.g., Gobl, 1989; Gobl and Ní Chasaide, 1992). And as pointed out for other qualities, there is not a one-to-one mapping to affective attributes: lax–creaky voice also gets high ratings for *relaxed*, *intimate* and *content*, and moderately high ratings (the highest in this test) for *sad* and *friendly*. In the responses for *intimate*, and particularly for *sad* and *friendly*, there would appear to be greater cross-subject variability (Fig. 2). Note however, that the very extended total range of values here results from atypical responses of a single subject in each case.

In contrast to the rather high ratings for boredom obtained with the lax–creaky stimulus here, Burkhardt and Sendlmeier (2000) report that this association is not clearly indicated, and may even be counter-indicated. Two factors might be responsible for these differences in results. Firstly, the stimuli presented to subjects may have been very different, but as there is little detail in that study on source parameter settings for the generation of the different voice qualities, a direct comparison is not possible here. As our results indicate that not all types of creaky voice are necessarily highly rated for boredom, a difference in the stimuli could be highly relevant. A further factor may be cross-language differences. Creaky voice is often mentioned as related to the expression of boredom for speakers of English, and this is not necessarily assumed to be universal. Burkhardt and Sendlmeier's subjects were German, and differences in results could be influenced by this difference in subjects.

When assessing the strength of ratings achieved for individual attributes by the present stimuli, it must be borne in mind that, however important, voice quality is only one of a number of vocal features that speakers may exploit to communicate affect. When we find a strong and consistent association of affect with a particular stimulus (e.g., the tense stimulus and *angry*) in this experiment, we can be fairly confident that this type of quality can alone evoke the affect, even though in real discourse features other than voice quality may further enhance its perception. In cases where we find a moderate association between a stimulus and a particular affect (e.g., the tense stimulus and *happy*) it is less obvious what this might be telling

us. It could mean that the quality approximated by the tense stimulus is not quite appropriate for the communication of happiness. Or it might indicate that although appropriate, the voice quality is not a dominant cue to this affect, and that some other critical features (such as tempo or specific f_0 variations) are lacking without which happiness is not effectively conveyed.

It is striking that for the range of voice qualities that were synthesised for this experiment, milder states were better signalled than the strong emotions (the exception being anger). It may well be the case, that to communicate strong emotions one would need, at the very least, to incorporate those large f_0 dynamic changes described in the literature on the vocal expression of emotion, e.g., by Scherer (1986) or Mozziconacci (1995, 1998). In the present stimuli, only relatively small f_0 differences were included such as were deemed intrinsic correlates of individual voice qualities. A possible hypothesis at this stage is that voice quality and pitch variables may have at least partially different functions in affect signalling, with voice quality playing a critical role in the general communication of milder affective distinctions (general speaker states, moods and attitudes), especially those that have no necessary physiological component, and pitch variables, such as major differences in f_0 level and range being more critical in the signalling of strong emotions where physiologically determined glottal changes are more likely. This type of hypothesis would we feel be compatible with arguments and findings of other researchers, e.g., Scherer (1986) who has suggested that whereas large f_0 shifts signal gross changes in activation/arousal levels, voice quality variables may be required to differentiate between subtle differences in affect. Support for this viewpoint can be construed from the results of the experiments of Scherer et al. (1984), which are unusual in that the typically studied strong emotions are excluded. Voice quality emerged in that study as the overwhelmingly important variable that correlated with listener's judgements of affect. The possibility of voice quality and f_0 serving different and potentially independent functions in affect signalling have also been raised by Laukkanen et al. (1997), Murray and Arnott (1993), Scherer et al. (1984) and Ladd et al. (1985).

An alternative hypothesis that should also be borne in mind is that voice quality differences, but of a much more extreme nature than those simulated in the present study would be required for the signalling of strong emotions. This would imply that both voice quality and f_0 variables function in a similar and essentially gradient fashion in the signalling of strong emotions. This would not necessarily entail that the communication of mild affective states might not rely more heavily on voice quality differentiation.

The way in which voice quality variables combine with pitch variables is the focus of some of our current ongoing research. To test the first hypothesis mentioned above, we are looking at whether large pitch excursions, as described by Mozziconacci (1995), with and without voice quality manipulations would achieve a better signalling of the strong emotions. Some preliminary results are included by Ní Chasaide and Gobl (2002). We also plan to test the extent to which the relatively smaller f_0 differences included in the present stimuli (deemed intrinsic to these voice qualities) may have contributed to the perception of these affects.

Of course, f_0 itself is a source parameter and an intrinsic part of voice quality. The fact that these have to date been studied as separate entities is at least partially a reflection on the methodological constraints that pertain to voice source analysis. The broader linguistic phonetic literature, dealing with languages which have register (voice quality) and tonal contrasts highlights two things. Firstly, f_0 and voice quality can operate in a largely independent way, and secondly, there are broad tendencies for them to covary, so that for a number of register contrasts there are salient pitch correlates, whereas for a number of tonal contrasts there may be striking voice quality correlates (see discussion of this point in Ní Chasaide and Gobl, 1997). Even within modal voice in the mid-pitch range, there are some interactions between f_0 and other source parameters. To the extent that these have been studied, results appear to be sometimes contradictory and suggest that they may depend on rather complex factors (see, for instance, Fant, 1997;

Koreman, 1995; Pierrehumbert, 1989; Strik and Boves, 1992; Swerts and Veldhuis, 2001). For the very large differences in pitch level and range, described in the literature on the vocal expression of (strong) emotions, it seems very unlikely that these would occur without major adjustments to voice quality. If this is the case, the absence of the voice quality domain in analyses is a serious deficit, and likely to lead to unsatisfactory results in synthesis. This could provide one explanation as to why perception tests of large f_0 excursions to cue emotions can sometimes yield disappointing results (see, for example, Mozziconacci, 1995, 1998).

6. Conclusions

In this study we have focussed on voice quality, which is of course one of a variety of features used in the communication of affect. Results illustrate that differences in voice quality alone can evoke quite different colourings in an otherwise neutral utterance. They further suggest that there is no one-to-one mapping between voice quality and affect: individual qualities appear rather to be associated with a constellation of affective states, sometimes related, sometimes less obviously related.

Some previously reported associations between voice quality and affect (e.g., anger and tense voice) are supported by the present results, whereas others are clearly not (e.g., tense voice and fear). In certain cases (e.g., the association of creaky voice with boredom, or breathy voice with sadness or intimacy) refinements would be suggested. For these affects, the lax–creaky stimulus (which combined features of breathy and creaky voice) yielded considerably higher responses. Furthermore, the 'broad palette' approach adopted here, whereby listeners rated a rather wide variety of affective states, rather than the smaller selection of strong emotions more typically included, allowed other strong associations to emerge, such as the *formal*, *confident* and *interested* colouring of tense voice. Results also permit us to see at a glance which of the synthetic utterances presented here were rated as the most *friendly*, *stressed*, *relaxed*, etc.

The voice qualities presented in this experiment were considerably more effective in signalling the relatively milder affective states and generally ineffective in signalling strong emotions (excepting anger). This raises the question as to whether the role of pitch dynamics and voice quality may be somewhat different in the communication of affect: voice quality may be critical in the differentiation of subtle variations in affective states, whereas large pitch excursions, such as described in the emotion literature may be critical to the signalling of strong emotions.

The findings are based on synthetic stimuli and tell us about how voice quality in human speech communication only insofar as the targeted voice qualities were successfully synthesised. Specific difficulties were encountered in the synthesis of whispery voice and the similarity in responses to the whispery and breathy stimuli suggests that further work would be required at the level of generating a better simulation of the former quality in particular. Similarly, in the case of harsh voice, results also indicate that this stimulus may not have been optimal. For both the whispery and harsh stimuli, caution is required in interpreting results. This highlights the need for further work on both the production and perception correlates of these two qualities in particular, but also more generally, on all qualities. One other aspect that we would hope to explore concerns how more gradual changes in source parameters along a given voice quality continuum relate to changes in the perception of associated affect(s).

While results demonstrate that voice quality differences alone can impart very different affective overtones to a message, this does not imply that speakers use this feature in isolation. As discussed in Section 1, there are other source features (pitch dynamics), vocal tract features (segmental differences) and temporal features which speakers can and do exploit for such paralinguistic communication. As a step towards understanding how these may combine, we are currently extending the present study to look at how voice quality combines with f_0 variables in signalling emotions. It is hoped that these efforts will contribute to the bigger picture, which concerns not only how voice quality combines with the other known vocal correlates of

affective speech, but also how the precise meaning of an utterance results from an interaction of these vocal cues with its verbal content.

Acknowledgements

We are grateful to Elizabeth Heron of the Department of Statistics, TCD, for assistance with the statistical analysis.

References

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11, 109–118.

Alku, P., Vilkman, E., 1994. Estimation of the glottal pulseform based on discrete all-pole modeling. In: Proceedings of the International Conference on Spoken Language Processing, Yokohama, pp. 1619–1622.

Alku, P., Vilkman, E., 1996. A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. Folia Phoniatrica et Logopaedica 48, 240–254.

Alter, K., Rank, E., Kotz, S.A., Pfeifer, E., Besson, M., Friederici, A.D., Matiasek, J., 1999. On the relations of semantic and acoustic properties of emotions. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 2121–2124.

Ananthapadmanabha, T.V., 1984. Acoustic analysis of voice source dynamics. STL-QPSR 2–3, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 1–24.

Burkhardt, F., Sendlmeier, W.F., 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. In: Cowie, R., Douglas-Cowie, E., Schröder, M. (Eds.), Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Queen's University, Belfast, pp. 151–156.

Cahn, J., 1990a. The generation of affect in synthesized speech. Journal of the American Voice I/O Society 8, 1–19.

Cahn, J., 1990b. Generating expression in synthesized speech. Technical report, MIT Media Laboratory, Boston.

Carlson, R., Granström, B., Karlsson, I., 1991. Experiments with voice modelling in speech synthesis. Speech Communication 10, 481–489.

Carlson, R., Granström, B., Nord, L., 1992. Experiments with emotive speech, acted utterances and synthesized replicas. Speech Communication 2, 347–355.

Chan, D.S.F., Brookes, D.M., 1989. Variability of excitation parameters derived from robust closed phase glottal inverse filtering. In: Proceedings of Eurospeech '89, Paris, paper 33.1.

Childers, D.G., Lee, C.K, 1991. Vocal quality factors: Analysis, synthesis, and perception. Journal of the Acoustical Society of America 90, 2394–2410.

Cranen, B., Boves, L., 1985. Pressure measurements during speech production using semiconductor miniature pressure transducers: impact on models for speech production. Journal of the Acoustical Society of America 77, 1543–1551.

Cummings, K.E., Clements, M.A., 1995. Analysis of the glottal excitation of emotionally styled and stressed speech. Journal of the Acoustical Society of America 98, 88–98.

Ding, W., Kasuya, H., Adachi, S., 1994. Simultaneous estimation of vocal tract and voice source parameters with application to speech synthesis. In: Proceedings of the International Conference on Spoken Language Processing, Yokohama, pp. 159–162.

Fant, G., 1979a. Glottal source and excitation analysis. STL-QPSR 1, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 85–107.

Fant, G., 1979b. Vocal source analysis – a progress report. STL-QPSR 3–4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 31–54.

Fant, G., 1982. The voice source – acoustic modeling. STL-QPSR 4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 28–48.

Fant, G., 1995. The LF-model revisited. Transformations and frequency domain analysis. STL-QPSR 2–3, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 119–156.

Fant, G., 1997. The voice source in connected speech. Speech Communication 22, 125–139.

Fant, G., Lin, Q., 1991. Comments on glottal flow modelling and analysis. In: Gauffin, J., Hammarberg, B. (Eds.), Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms. Singular Publishing Group, San Diego, pp. 47–56.

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

Frick, R.W., 1985. Communicating emotion: the role of prosodic features. Psychological Bulletin 97, 412–429.

Fröhlich, M., Michaelis, D., Strube, H.W., 2001. SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. Journal of the Acoustical Society of America 110, 479–488.

Fujisaki, H., Ljungqvist, M., 1986. Proposal and evaluation of models for the glottal source waveform. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, pp. 31.2.1–31.2.4.

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

Gobl, C., 1989. A preliminary study of acoustic voice quality correlates. STL-QPSR 4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 9–21.

Gobl, C., Ní Chasaide, A., 1992. Acoustic characteristics of voice quality. Speech Communication 11, 481–490.

Gobl, C., Ní Chasaide, A., 1999a. Perceptual correlates of source parameters in breathy voice. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 2437–2440.

Gobl, C., Ní Chasaide, A., 1999b. Techniques for analysing the voice source. In: Hardcastle, W.J., Hewlett, N. (Eds.), Coarticulation: Theory, Data and Techniques. Cambridge University Press, Cambridge, pp. 300–321.

Gobl, C., Monahan, P., Ní Chasaide, A., 1995. Intrinsic voice source characteristics of selected consonants. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol. 1, pp. 74–77.

Hammarberg, B., 1986. Perceptual and acoustic analysis of dysphonia. Studies in Logopedics and Phoniatrics 1, Huddinge University Hospital, Stockholm, Sweden.

Hedelin, P., 1984. A glottal LPC-vocoder. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, pp. 1.6.1–1.6.4.

Hertegård, S., Gauffin, J., 1991. Insufficient vocal fold closure as studied by inverse filtering. In: Gauffin, J., Hammarberg, B. (Eds.), Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms. Singular Publishing Group, San Diego, pp. 243–250.

Hunt, M.J., 1987. Studies of glottal excitation using inverse filtering and an electroglottograph. In: Proceedings of the XIth International Congress of Phonetic Sciences, Stockholm, Tallinn, Vol. 3, pp. 23–26.

Hunt, M.J., Bridle, J.S., Holmes, J.N., 1978. Interactive digital inverse filtering and its relation to linear prediction methods. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, pp. 15–18.

Iida, A., Campbell, N., Iga, S., Higuchi, H., Yasumura, M., 2000. A speech synthesis system with emotion for assisting communication. In: Cowie, R., Douglas-Cowie, E., Schröde, M. (Eds.), Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Queen's University, Belfast, pp. 167–172.

Jansen, J., 1990. Automatische extractie van parameters voor het stembron-model van Liljencrants & Fant. Unpublished master thesis, Nijmegen University.

Johnstone, T., Scherer, K.R., 1999. The effects of emotions on voice quality. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 2029–2032.

Kane, P., Ní Chasaide, A., 1992. A comparison of the dysphonic and normal voice source. Journal of Clinical Speech and Language Studies, Dublin 1, 17–29.

Kappas, A., Hess, U., Scherer, K.R., 1991. Voice and emotion. In: Feldman, R.S., Rimé, B. (Eds.), Fundamentals of Nonverbal Behavior. Cambridge University Press, Cambridge, pp. 200–238.

Karlsson, I., 1990. Voice source dynamics for female speakers. In: Proceedings of the International Conference on Spoken Language Processing, Kobe, Japan, pp. 225–231.

Karlsson, I., 1992. Modelling voice source variations in female speech. Speech Communication 11, 1–5.

Karlsson, I., Liljencrants, J., 1996. Diverse voice qualities: models and data. SMH-QPSR 2, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 143–146.

Kasuya, H., Maekawa, K., Kiritani, S., 1999. Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 2505–2512.

Kienast, M., Paeschke, A., Sendlmeier, W.F., 1999. Articulatory reduction in emotional speech. In: Proceedings of Eurospeech '99, Budapest, pp. 117–120.

Kitzing, P., Löfqvist, A., 1975. Subglottal and oral pressure during phonation—preliminary investigation using a miniature transducer system. Medical and Biological Engineering 13, 644–648.

Klasmeyer, G., Sendlmeier, W.F., 1995. Objective voice parameters to characterize the emotional content in speech. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol. 1, pp. 182–185.

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America 67, 971–995.

Klatt, D.H., unpublished chapter. Description of the cascade/parallel formant synthesiser. Sensimetrics Corporation, Cambridge, MA, Chapter 3, 79 pp.

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.

Koreman, J., 1995. The effects of stress and F_0 on the voice source. Phonus 1, University of Saarland, pp. 105–120.

Ladd, D.R., Silverman, K.E.A., Tolkmitt, F., Bergman, G., Scherer, K.R., 1985. Evidence for the independent function of intonation contour type, voice quality and F_0 range in signaling speaker affect. Journal of the Acoustical Society of America 78, 435–444.

Laukkanen, A.-M., Vilkman, E., Alku, P., Oksanen, H., 1995. On the perception of emotional content in speech. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol. 1, pp. 246–249.

Laukkanen, A.-M., Vilkman, E., Alku, P., Oksanen, H., 1996. Physical variation related to stress and emotionally state: a preliminary study. Journal of Phonetics 24, 313–335.

Laukkanen, A.-M., Vilkman, E., Alku, P., Oksanen, H., 1997. On the perception of emotions in speech: the role of voice quality. Scandinavian Journal of Logopedics, Phoniatrics and Vocology 22, 157–168.

Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge.

Lee, C.K., Childers, D.G., 1991. Some acoustical, perceptual, and physiological aspects of vocal quality. In: Gauffin, J., Hammarberg, B. (Eds.), Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms. Singular Publishing Group, San Diego, pp. 233–242.

Ljungqvist, M., Fujisaki, H., 1985. A method for simultaneous estimation of voice source and vocal tract parameters based on linear predictive analysis. Transactions of the Committee on Speech Research, Acoustical Society of Japan S85-21, 153–160.

Mahshie, J., Gobl, C., 1999. Effects of varying LF parameters on KLSYN88 synthesis. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 1009–1012.

McKenna, J., Isard, S., 1999. Tailoring Kalman filtering toward speaker characterisation. In: Proceedings of Eurospeech '99, Budapest, pp. 2793–2796.

Meurlinger, C., 1997. Emotioner i syntetiskt tal. M.Sc dissertation, Speech, Music and Hearing, Royal Institute of Technology, Stockholm.

Minitab, 2001. Minitab Inc. MINITAB Statistical Software, Release 13, Minitab, State College PA, 2001.

Monsen, R.B., Engebretson, A.M., 1977. Study of variations in the male and female glottal wave. Journal of the Acoustical Society of America 62, 981–993.

Mozziconacci, S., 1995. Pitch variations and emotions in speech. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol. 1, pp. 178–181.

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

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.

Murray, I.R., Arnott, J.L., 1995. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication 20, 85–91.

Murray, I.R., Edgington, M.D., Campion, D., Lynn, J., 2000. In: Cowie, R., Douglas-Cowie, E., Schröder, M. (Eds.), Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Queen's University, Belfast, pp. 173–177.

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.

Ní Chasaide, A., Gobl, C., 1995. Towards acoustic profiles of phonatory qualities. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol. 4, pp. 6–13.

Ní Chasaide, A., Gobl, C., 1997. Voice source variation. In: Hardcastle, W.J., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Oxford, pp. 427–461.

Ní Chasaide, A., Gobl, C., 2002. Voice quality and the synthesis of affect. In: Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (Eds.), Improvements in Speech Synthesis. Wiley and Sons, New York, pp. 252–263.

Ní Chasaide, A., Gobl, C., Monahan, P., 1992. A technique for analysing voice quality in pathological and normal speech. Journal of Clinical Speech and Language Studies, Dublin 1, 1–16.

Olivera, L.C., 1993. Estimation of source parameters by frequency analysis. In: Proceedings of Eurospeech '93, Berlin, pp. 99–102.

Olivera, L.C., 1997. Text-to-speech synthesis with dynamic control of source parameters. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springar-Verlag, New York, pp. 27–39.

Palmer, S.K., House, J., 1992. Dynamic voice source changes in natural and synthetic speech. In: Proceedings of the International Conference on Spoken Language Processing, Banff, pp. 129–132.

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

Price, P.J., 1989. Male and female voice source characteristics: inverse filtering results. Speech Communication 8, 261–277.

Qi, Y.Y., Bi, N., 1994. Simplified approximation of the 4-parameter LF model of voice source. Journal of the Acoustical Society of America 96, 1182–1185.

Rosenberg, A.E., 1971. Effect of glottal pulse shape on the quality of natural vowels. Journal of the Acoustical Society of America 49, 583–598.

Rothenberg, M., Carlson, R., Granström, B., Lindqvist-Gauffin, J., 1975. A three-parameter voice source for speech synthesis. In: Fant, G. (Ed.), Proceedings of the Speech Communication Seminar, Stockholm, 1974, Vol. 2. Almqvist and Wiksell, Stockholm, pp. 235–243.

Scherer, K.R., 1981. Speech and emotional states. In: Darby, J. (Ed.), The Evaluation of Speech in Psychiatry and Medicine. Grune and Stratton, New York, pp. 189–220.

Scherer, K.R., 1986. Vocal affect expression: A review and a model for future research. Psychological Bulletin 99, 143–165.

Scherer, K.R., 1989. Vocal measurement of emotion. In: Plutchik, R., Kellerman, H. (Eds.), Emotion: Theory, Research, and Experience, Vol. 4. Academic Press, San Diego, pp. 233–259.

Scherer, K.R., 1994. Affect bursts. In: van Goozen, S.H.M., van de Poll, N.E., Sergeant, J.A. (Eds.), Emotions. Lawrence Erlbaum, Hillsdale, NJ, pp. 161–193.

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.

Schoentgen, J., 1993. Modelling the glottal pulse with a self-excited threshold autoregressive model. In: Proceedings of Eurospeech '93, Berlin, pp. 107–110.

Schröder, M., 2000. Experimental study of affect bursts. In: Cowie, R., Douglas-Cowie, E., Schröder, M. (Eds.), Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Queen's University, Belfast, pp. 132–137.

Scully, C., 1994. Data and methods for the recovery of sources. Deliverable 15 in the Report for the Speech Maps Workshop,

Esprit/Basic Research Action no. 6975, Vol. 2, Institut de la Communication Parlée, Grenoble.

Scully, C., Stromberg, K., Horton, D., Monahan, P., Ní Chasaide, A., Gobl, C., 1995. Analysis and articulatory synthesis of different voicing types. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Vol. 2, pp. 482–485.

Stibbard, R., 2000. Automated extraction of ToBI annotation data from the Reading/Leeds emotional speech corpus. In: Cowie, R., Douglas-Cowie, E., Schröder, M. (Eds.), Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Queen's University, Belfast, pp. 60–65.

Strik, H., Boves, L., 1992. On the relation between voice source parameters and prosodic features in connected speech. Speech Communication 11, 167–174.

Strik, H., Boves, L., 1994. Automatic estimation of voice source parameters. In: Proceedings of the International Conference on Spoken Language Processing, Yokohama, pp. 155–158.

Strik, H., Jansen, J., Boves, L., 1992. Comparing methods for automatic extraction of voice source parameters from continuous speech. In: Proceedings of the International Conference on Spoken Language Processing, Banff, Vol. 1, pp. 121–124.

Strik, H., Cranen, B., Boves, L., 1993. Fitting a LF-model to inverse filter signals. In: Proceedings of Eurospeech '93, Berlin, pp. 103–106.

Swerts, M., Veldhuis, R., 2001. The effect of speech melody on voice quality. Speech Communication 33, 297–303.

Talkin, D., Rowley, J., 1990. Pitch-synchronous analysis and synthesis for TTS systems. In: Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, France, pp. 55–58.

Uldall, E., 1964. Dimensions of meaning in intonation. In: Abercrombie, D., Fry, D.B., MacCarthy, P.A.D., Scott, N.C., Trim, J.L.M. (Eds.), In Honour of Daniel Jones. Longman, London, pp. 271–279.

Veldhuis, R., 1998. A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation. Journal of the Acoustical Society of America 103, 566–571.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. Journal of the Acoustical Society of America 52, 1238–1250.

Wong, D., Markel, J., Gray, A.H., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transaction on Acoustics, Speech and Signal Processing 24 (4), 350–355.