Publication I

Matti Airas and Paavo Alku, “Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient.” *Phonetica*, 63(1), pp. 26–46, March 2006.

Original Paper

Phonetica 2006;63:26–46
DOI: 10.1159/000091405

Received: September 20, 2004
Accepted: August 9, 2005

Emotions in Vowel Segments of Continuous Speech: Analysis of the Glottal Flow Using the Normalised Amplitude Quotient

Matti Airas Paavo Alku

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology (HUT), Espoo, Finland

Abstract

Emotions in short vowel segments of continuous speech were analysed using inverse filtering and a recently developed glottal flow parameter, the normalised amplitude quotient (NAQ). Simulated emotion portrayals were produced by 9 professional stage actors. Separated /a:/ vowel segments were inverse filtered and parameterised using NAQ. Statistical analyses showed significant differences among most of the emotions studied. Results also demonstrated clear gender differences. Inverse filtering, together with NAQ, was shown to be a promising method for the analysis of emotional content in continuous speech.

1 Introduction

According to the component process theory, emotions are processes of events affecting several psychophysical components of an organism, namely physiological arousal, motor expression, and subjective feeling [e.g. Scherer, 2000]. While the subjective feeling is a condition internal to an organism, physiological arousal and motor expression are components that can be externally observed from behaviour, facial expression and body postures, as well as from speech. Given the observable changes in physiological arousal and motor expression, emotions may be considered to possess an inherently communicative role. However, in contrast to language, which is widely considered to be a purely learned and cultural trait, emotional expression also incorporates a strong innate component. Several studies have indicated that facial and vocal expression of emotion can be correctly interpreted despite cultural differences or even across species [e.g. Darwin, 1872; Ekman, 1992; Elfenbein and Ambady, 2002; Frank and Stennett, 2001; Leinonen et al., 1991; Scherer et al., 2001].

Murray and Arnott [1993] and Scherer [2003] present reviews of the development of modern research in the vocal expression of emotion. According to Scherer [2003], the empirical research of the vocal expression of emotion began in the early 20th century. The invention of the telephone and the radio increased scientific interest in the

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2006 S. Karger AG, Basel
0031–8388/06/0631–0026
$23.50/0
Accessible online at:
www.karger.com/journals/pho

Matti Airas
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
PO Box 3000, FI–02015 HUT (Finland)
Tel. +358 9 451 2109, Fax. +358 9 460 224
E-Mail Matti.Airas@hut.fi

communication of speaker attributes and states, via vocal cues in speech, but systematic research programmes on the topic only began in the 1960s. Since then, research has become multidisciplinary, involving psychology, linguistics, phonetics, and electrical engineering as well as computer science. Within the last few years, research on the vocal expression of emotion has advanced rapidly, due to basic psychological as well as phonetic and speech processing research, and more applied research of speech recognition and synthesis. However, despite the progress in the study of vocal expression of emotion, knowledge of the topic is still far from conclusive. The open issues, some of which are discussed below, have been both methodological and conceptual.

Emotion elicitation methods control the ways emotions and emotional expression are induced. These methods can be classified into three categories: natural vocal emotion expression, induced emotional expression, and simulated emotional expression [Scherer, 2003]. Natural vocal expression has been used in some studies of vocal expression. The material has been recorded during naturally occurring emotional states, such as flight in dangerous situations and journalists reporting dramatic events [e.g. Williams and Stevens, 1969, 1972]. While tempting, natural expressions of emotion are difficult to obtain, the recording conditions are difficult to control and the precise nature of the underlying emotion is often uncertain. Emotion induction procedures use controlled, external means of inducing emotional states in the test subjects [Scherer, 2003]. These induction methods include imagination, repetition of self-reflective statements (Velten induction method), watching films, listening to stories or music and social interaction [Westermann et al., 1996]. While these methods offer a great degree of control, the affect produced is often relatively weak. Furthermore, the induction procedures do not necessarily produce similar emotional states in all individuals. Simulated vocal expressions have been the preferred way of obtaining emotional voice samples [Scherer, 2003]. Actors are asked to produce vocal expressions of emotion. The task may be based on emotion labels or acted scenarios. The speech task often utilises standard verbal content. The simulated emotion portrayals are usually much more intense and prototypical than natural emotions, and the possibility that actors tend to overemphasise obvious emotional cues and discard more subtle ones must be admitted. Simulated emotion expression may reflect sociocultural norms or expectations more so than the psychophysiological changes occurring under natural expression. However, since acted portrayals are reliably recognised as such by listeners, it can be assumed that the portrayals reflect, at least in part, natural expression patterns.

The absence of a consensual definition of emotion and of qualitatively different types of emotion has been a deterrent to progress in this area [Scherer, 2003]. Scherer [2000] has therefore proposed a design feature approach to distinguish the different affect states. In his article, he provided an outline of the different approaches to distinction of different emotions. Cowie and Cornelius [2003] also outlined different methods for describing emotions expressed in speech. Their systems include lists of key emotion categories, biological representations, abstract dimensions, such as activation-evaluation space, as well as structural models, such as the componential models [Scherer, 2003]. To date, no current model fits all emotion description requirements perfectly, and no such model can be expected to appear in the near future. Therefore it seems reasonable to choose, in any given study, a description model suitable for that study, and then use it systematically.

In speech, emotion is communicated by a combination of features at all three principal levels of speech abstraction: suprasegmental, segmental, and intrasegmental

[Murray and Arnott, 1993]. All of these three levels can be considered to consist of two components, namely verbal (words) and vocal (intonation, voice quality, and intensity). The vocal information at any of these levels can render the verbal information redundant, qualify it further, or directly contradict it. At suprasegmental and segmental levels, this information includes the fundamental frequency (F_0), sound pressure level (SPL), and duration patterns, as well as variations in formant configurations. While suprasegmental patterns are undoubtedly important for vocal expression of emotion, Pollack et al. [1960] noted that emotion can be recognised in segments of speech as short as 60 ms. On this intrasegmental level, vocal expression of emotion is performed with various voice quality adjustments generated by manipulation of the voice source (the glottal volume velocity waveform generated by the vibrating vocal folds).

In the context of voice research, voice quality does not refer to any measured excellence of the sound, but instead voice quality is defined as the characteristic auditory colouring of an individual speaker's voice, to which both laryngeal and supralaryngeal features contribute [Laver, 1980]. Some examples of voice qualities include modal, breathy, pressed, creaky, and falsetto. Laver and Hanson [1981] outlined a framework to describe voice qualities based on the concept of 'settings' of the speech organs [Laver, 1980]. This framework has since become commonly used in describing voice quality. Although voice quality is an important factor in determining the emotional content of speech, the two terms are not synonymous, as voice quality can also vary independently of the emotional content. For example, increased intensity usually leads to a more pressed voice quality. While Laver [1980] and Laver and Hanson [1981] include supralaryngeal features in voice quality, a narrower interpretation includes only features deriving solely from laryngeal activity. Multiple studies on acoustic parameters of emotional speech have been conducted in both suprasegmental, segmental, as well as intrasegmental level. The earliest of these were done in the 1930s, after the invention of phonophotographic oscillography [Metfessel, 1926]. Skinner [1935] investigated F_0 and intensity patterns in the expression of induced happiness and sadness. Cowan [1936] studied F_0 and intensity patterns of acted expressions. Fairbanks and Pronovost [1938, 1939], Fairbanks [1940] and Fairbanks and Hoaglin [1941] explored F_0, intensity and durational patterns of a speech passage elicited in five different dramatised expressions. Later research generally confirms the results of their studies. After the early forties, there was little research on the vocal expression of emotions until the 1960s, when renewed interest began to produce new results in the field. For example, Kaiser [1962] studied how single vowels could express affects. Six acted emotions were analysed using oscillograms, spectrograms, and F_0 and intensity patterns. The F_0 and timbre were found to have the strongest effect on the perception of emotions. More spectrographic studies on emotion in speech were presented by Williams and Stevens [1969, 1972]. Suprasegmental sound features, mainly F_0 contours, were analysed for real emotional recordings acquired from radio traffic between a civilian pilot and a control tower operator [Williams and Stevens, 1969]. The same features were later analysed for simulated emotions, yielding results that correlated with the previous ones [Williams and Stevens, 1972]. The F_0 ranges and contours were found to vary according to expressed emotion, with a definite increase of F_0 range in distress. During the 1970s, the use of computers became prevalent in the study of acoustical correlates of emotion in speech, and several studies utilising digital analysis of emotional speech were performed. For example, Levin and Lord [1975] studied simulated emotions using cepstrograms and F_0 ranges. They performed statistical analyses on the results and concluded that an analysis of the F_0 range is sufficient to provide an indication of emotional change. Scherer [1981] suggested that F_0, loudness and temporal characteristics correlate with the activation dimension of emotion, while the discrete emotions themselves cannot be modelled effectively using those features. Indeed, in the 1980s it became apparent that the previous studies concentrating on pitch, loudness, and temporal characteristics of speech could not adequately describe the vocal expression of emotion, but that the key to the vocal differentiation of discrete emotions is the voice quality [e.g. Scherer, 1986]. He cited conceptual and methodological difficulties as the reason for the neglect of the voice quality in empirical studies of the vocal expression of emotion.

Banse and Scherer [1996] studied emotion portrayals by professional actors in 14 emotions varying in intensity and valence. Multiple data including different F_0, energy, speech rate, and spectral parameters were compiled and analysed. Rather good performance in statistical emotion classification was reported. Although some information on the voice quality was encoded in the spectral parameters, an explicit analysis of the voice quality or voice source parameters was not attempted. To acquire direct information on voice quality, analysis of the laryngeal and supralaryngeal features affecting voice quality is required. If supralaryngeal features are ignored, voice quality is determined purely by the laryngeal features, which are all embedded in the airflow through the glottis. As direct measurement of this flow is not possible, inverse filtering is used to measure the glottal flow waveform from the extraoral sound pressure or oral volume velocity waveform. Miller [1959] introduced the basic principles for inverse filtering of a voice pressure signal. He used manual placement of analogue anti-resonators to cancel the formants and perform the inverse filtering, and to acquire the glottal flow waveform. Manual inverse filtering is still in common use [e.g. Gobl and Ní Chasaide, 2003a]. The negative aspects of manual inverse filtering are the amount of work required and reliance on the experimenter's subjective preferences in the shape of the glottal flow waveform. Rothenberg [1973] introduced an inverse filtering method utilising a pneumotachograph. This method facilitated measurement of absolute amplitude values, including the DC airflow, and was less susceptible to low-frequency noise than the previous methods. Although used commonly in research on voice source function, it is cumbersome in the study of voice quality, as it arguably makes the expression of different voice types and intensities unnecessarily difficult for inexperienced speakers. Automatic inverse filtering utilising linear prediction and digital filtering was first suggested by Allen and Curtis [1973]. The closed phase covariance method was presented shortly thereafter [e.g. Strube, 1974; Wong et al., 1979]. However, it gives reliable results only when the glottal source has a sufficiently long closed phase. This limits the usefulness of the method in the analysis of a vocal expression of emotion. More robust methods for automatic inverse filtering were then developed by Mataušek and Batalov [1980], Milenkovic [1986], Javkin et al. [1987], and Alku [1992]. Contemporary inverse filtering methods allow for inverse filtering of voices traditionally considered difficult, such as very breathy voices and female voices.

Inverse filtering by itself only results in an estimate of the glottal flow volume waveform, which needs to be parameterised to get quantitative measures of the voice source function. Different voice source parameterisation methods include time-based parameters, amplitude-based parameters, function fitting, and frequency domain parameters. Time-domain parameterisation, originating from work by Timcke et al. [1958], is performed by segmenting the inverse filtered signal to discrete events and measuring the relative lengths of different segments, such as the open quotient and closing quotient (ClQ). Accurate computation of time-based parameters is problematic due to difficulty in the interpretation of the exact opening time of the vocal folds as well as to formant ripple and noise often present in the glottal waveforms [Dromey et al., 1992; Holmberg et al., 1988]. Amplitude domain parameters can be used if inverse filtering is performed using airflow signals [e.g. Hertegård et al., 1992; Holmberg et al., 1988]. Closely related to these are the parameters acquired from the derivative of the glottal flow waveform, for example, the maximum flow declination ratio [Fant, 1993; Holmberg et al., 1988]. Airflow parameterisation may also be performed in the frequency domain by means of parameters such as the harmonic richness factor [Childers and Lee, 1991] and the spectral slope [Titze and Sundberg, 1992]. Parabolic spectral parameter is a frequency domain glottal waveform parameter, which expresses the spectral decay of the voice source with respect to its maximal theoretical decay [Alku et al., 1997]. It is thought to be useful in quantifying the glottal source in cases when the phonation type is changed. Some spectral slope parameters may also be approximated directly from the speech signal [e.g. van Bezooyen, 1984]. A widely used approach is to fit certain mathematical functions to the waveform obtained by inverse filtering. One of the most widely used voice source models is the Liljencrants-Fant model (LF model), which is based on the quantification of the first derivative of the glottal flow with four parameters [Fant et al., 1985].

As mentioned, the voice quality differences produced by laryngeal features are one of the fundamental factors in differentiating vocal emotional expression. A straightforward way to inspect these differences is to inverse filter the speech signal and parameterise it suitably. One of the first studies combining inverse filtering in voice quality research was performed by Gobl [1989]. He parameterised the glottal waveform of different voice qualities by LF model fitting and spectral analysis. Cummings and Clements [1990, 1995] analysed glottal waveforms across eleven different stress styles. The chosen speech styles included both emotions (anger) and voice qualities (softness), as well as suprasegmental speech styles such as slow, fast, and questioning. Statistical analyses on different time parameters were performed, and each of the eleven stress styles was found to have a unique glottal waveform, which does not depend on the vowel spoken or on the speaker. Childers and Lee [1991] studied inverse filtered glottal flow and electroglottographic (EGG) signals in four different voice qualities. They found the major distinguishing features to be the pulse skewing of the glottal waveform and the richness of the harmonic spectrum.

Laukkanen et al. [1996] conducted one of the first studies actually coupling variations of the glottal flow signal and emotions. Nonsense utterances were produced in five simulated emotional states by 3 subjects. In addition to the regular F_0 and SPL measures, glottal flow was estimated and parameterised using the speed quotient and the quasi-open quotient. The emotions were found to differ from one another in the F_0 and SPL results as well as in the glottal waveform. Johnstone and Scherer [1999] performed a feasibility study of EGG analysis to the study of emotional voice production. Eight speakers were asked to portray seven imagined mild emotions. An analysis of some acoustic parameters together with EGG parameters was performed. It was noted that for high arousal emotions, the glottis closes faster, indicated by small ClQ values. Gobl and Ní Chasaide [2003b] also studied the role of voice quality in the communication of emotions. They inverse filtered a semantically neutral utterance spoken in a modal voice. The LF model was matched to the glottal pulses. The acquired LF model parameters were used to synthesise modal voice stimulus. The parameters were then further manipulated to generate non-modal stimuli. Perception test results showed statistically significant differences among the different synthetic voice types. However, the focus of their work was on the perception of emotional expression, not on the parameterisation of the emotional vocalisation.

The groundwork for studying voice quality changes in the vocal expression of emotion has been laid in previous research, including emotion induction, description and classification, as well as robust inverse filtering and glottal flow parameterisation methods. However, previous analyses have either used a very limited number of test subjects, or they have analysed voice qualities or stress styles instead of well-defined emotions. In contrast, this study analyses voice quality changes in the vocal expression of emotion. The material, recorded in a non-invasive manner using an ordinary free-field microphone, consists of excerpts of dramatised continuous speech. The recordings are inverse filtered and parameterised using a robust voice source parameter, the normalised amplitude quotient (NAQ). Since the current algorithms do not yet allow for inverse filtering of long recordings, it can be assumed that predefined /a:/ vowels sufficiently represent the emotional content of the whole expression. This study is restricted to an analysis of these segments.

The first goal of the study is to extract quantitative data of the voice source in the expression of emotions, as reflected by the NAQ parameter. The second goal is to develop and evaluate inverse filtering algorithms and methodologies for an analysis of spoken language, with the emphasis on continuous speech.

2 Materials and Methods

2.1 Speech Samples

Acted emotion portrayals were used in the study. Nine professional stage actors employed at the City Theatre of Oulu, Finland, served as speakers. The 5 male and 4 female subjects, all native speakers of Finnish, ranged between 26 and 45 years of age (median: 42). The subjects were paid for the recording.

Five different emotions were used: neutral, sadness, joy, anger, and tenderness. These emotions were chosen so that they could be separated clearly in the activation-evaluation space. Consequently, the acoustic and perceptual differentiation of the emotion portrayals can be assumed to be maximal. Sadness and tenderness have low activation, while joy and anger have high activation. Both sadness and anger have a negative valence, while the valence of joy and tenderness is positive. The term ‘neutral’ actually stands for weak emotional expression and was assumed to have average valence and activation.

The speech material of the study was a text passage of 83 words of Finnish prose. The contents of the passage could be expressed easily in different emotions. The recitations took approximately 1 min each. No explicit scenarios were given for different emotions; the actors were simply asked to recite the passage in a given emotion. The emotions were referred to by their Finnish names only: ‘neutraali’, ‘suru’, ‘ilo’, ‘viha’, and ‘hellyys’. The actors were free to use any method they deemed suitable to express the emotion.

The five emotions were repeated by the nine actors ten times each, giving a total of 50 recitations per actor (referred to as the recitation number). Thus, the total number of spoken passages was 450. The recitations were chosen in random order, but no emotion was presented twice in a row.

The speech samples were recorded over a course of 3 days in a radio anechoic chamber at the University of Oulu. The subject stood in the chamber with an instructor who supervised the recording session. A microphone was placed at a distance of 50 cm from the subject’s mouth. The distance was controlled before each recitation using a string tied to the microphone stand.

Before recording each subject, the microphone signal level was measured using a calibrator (Brüel & Kjær 4231). The recording microphone was Brüel & Kjær 4188, which has a frequency range from 8 to

Table 1. Contingency table for listening test grade medians for each sound sample

	Grade				
	1	2	3	4	5
Amount	5	43	110	250	132
Proportion, %	0.9	8.0	20.3	46.3	24.4

12,500 Hz (±2 dB). The microphone was connected through a preamplifier (Brüel & Kjær 2138 Mediator) to a Sony DTC-690 DAT recorder. The DAT recorder used a standard sampling rate of 48 kHz. To prevent signal degradation, the recorded signals were digitally transferred from DAT tapes to a computer.

To reduce the segmentation and inverse filtering effort of the ten portrayals of each emotion by each actor, only the first four were chosen for subsequent analyses. Thus, the total amount of analysed recitations was $4 \times 5 \times 9 = 180$, or 40% of the total amount of recitations. Signal segmentation was performed manually. Three /a:/ vowels, surrounded by an unvoiced plosive or fricative, were cut from predefined positions at the beginning, in the middle and at the end of each of the 180 analysed recitations (words [t**aa**shan], [t**aa**kkahan], [muistak**aa**pa]).

The /a:/ vowels were downsampled to 22.05 kHz and 40-ms sound segments were cut from them. The location of the segment was selected so that the segment was at a stationary part of the vowel, and, if possible, in the middle of the vowel. If the voiced, stationary region was somewhat shorter than 40 ms, it was nevertheless included in the analysis, but non-stationary segments were omitted from further analysis.

2.2 Speech Sample Validation

To assess the quality of the emotion portrayals, a listening test was arranged. A waveform of 3 s in the context of /a:/ vowels of each speaker and emotion was selected as speech material for this validation test. The waveform length was chosen so that the sample would include approximately one entire sentence around the /a:/ vowels.

Thirteen listeners (4 female) served as subjects in the listening test. The listeners had normal hearing and were between 22 and 32 years of age (median 26). They were shown the label of the expressed emotion, after which the sound sample was played back. Then, the following question was asked: 'How well did the sound sample convey the given emotion (1 = not at all, 5 = very well)?' Using a simple computer user interface, the listeners graded the sample on the given discrete scale. Every listener graded every sound sample, resulting in 540 graded samples.

Once the test had been conducted, grade medians were calculated for each sample. These median values were then used as a basis for subsequent analyses. The contingency table of the grade medians is given in table 1. Inter-rater agreement was estimated using weighted Light's kappa, which equals the mean of the weighted kappas obtained from each pair of listeners [Cohen, 1968; Conger, 1980]. The weighted kappa value was 0.44, indicating a moderate agreement amongst the listeners.

As can be seen in table 1, only 8.9% of the samples were graded 2 or less. Because the validation test yielded such a good result, it was decided to include all the samples in further processing.

The effect of the vowel position in the recitation to the grade was studied by calculating a contingency table of vowel positions and grades. The results are given in table 2. The contingency table was further analysed using the chi-squared test. The test indicated that the vowel position – or the semantic context around the vowel – had no effect on the grade ($\chi^2 = 2.9804$, $p = 0.9356$).

2.3 Estimation of the Glottal Flow with Inverse Filtering

The inverse filtering method used in this study was iterative adaptive inverse filtering (IAIF) [Alku, 1992]. The block diagram of the IAIF method is shown in figure 1. The only input required for

Table 2. Contingency table of vowel positions within the recitation and the grade given by the listeners

	Grade				
	1	2	3	4	5
Beginning	2	18	36	81	43
Middle	1	12	34	89	44
End	2	13	40	80	45

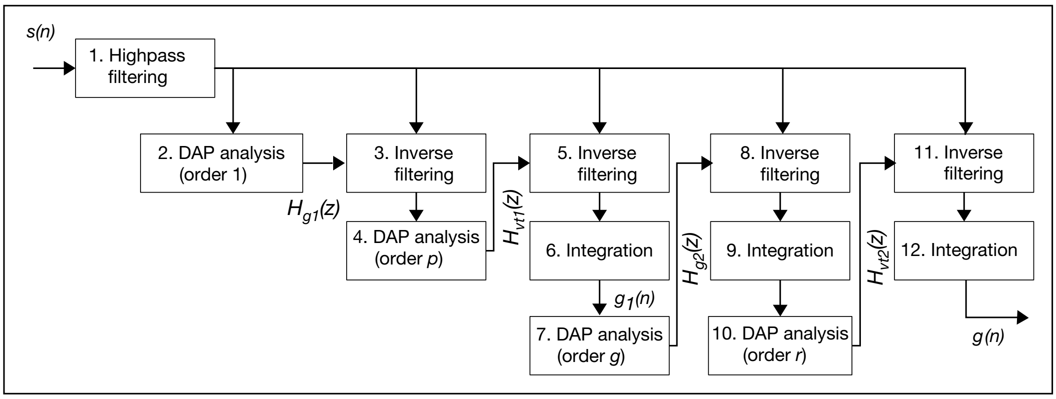

Fig. 1. The block diagram of the IAIF method for estimation of the glottal excitation $g(n)$ from the speech signal $s(n)$.

estimation of the glottal flow with IAIF is the speech pressure waveform $s(n)$ captured in free-field with a microphone. The output $g(n)$ is the estimated glottal flow. Figure 2 shows examples of the input and output signals. The method can be completely automated, although some supervision and interactive optimisation of the parameters often improves the quality of the inverse filtering greatly. IAIF can be implemented using either linear predictive coding (LPC) or discrete all-pole modelling (DAP) as an all-pole modelling technique. In this work DAP was used, because it is able to estimate the vocal tract filter more accurately than LPC, especially for high F_0 [Alku and Vilkman, 1994].

IAIF works by calculating a first estimate of the glottal waveform by inverse filtering, and then repeating the process by using the enhanced vocal tract model acquired in the first repetition as a basis for the next repetition. For a detailed description of IAIF, see also Alku et al. [1999].

The IAIF algorithm was implemented in MATLAB. A graphical user interface was devised for convenient inverse filtering of the signals. In the process, the number of formants and lip radiation parameters was adjusted to obtain an optimal glottal flow waveform. The number of formants varied typically from 8 to 14 (median 11), while the lip radiation value varied from 0.97 to 1.0 (median 0.9925). The quality of the waveform was graded subjectively by the experimenter from 0 (rejected) to 3 (excellent).

Of the 540 recordings, 76 (14%) could not be inverse filtered. Of the discarded samples, 15, 8, 8, 32, and 13 belonged to emotions anger, joy, neutral, sadness, and tenderness, respectively. In the case of anger, the predominant reason for failed inverse filtering was accidental clipping of the signal in the recording phase due to actor's shouting. In the case of sadness and tenderness, the failures were mostly due to insufficient voicing caused by whispering. It was suspected that the discarded recordings were among the most emotional ones. To investigate this, contingency tables of listening test grades and emotions were calculated for both the whole data set and the discarded samples. For each emotion, the distribution of the whole data set and the discarded samples were compared using the χ^2 test. In no case could the null hypothesis of the distributions being equal be rejected (p values ranged from 0.241

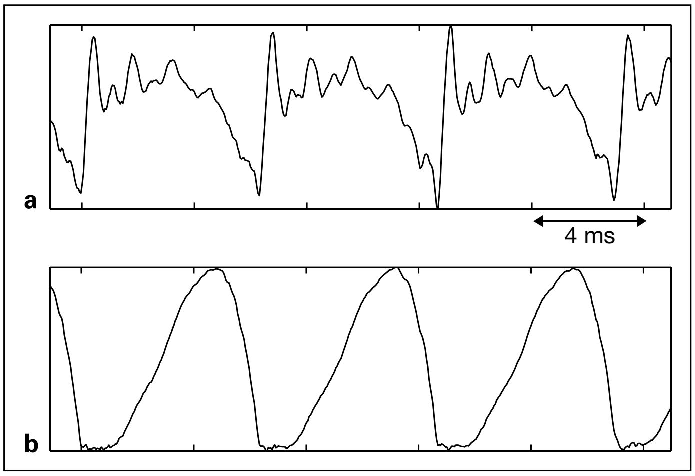

Fig. 2. An example of a short segment of a speech pressure waveform *s*(*n*) of a female speaker in 'tenderness' (**a**), and the corresponding glottal flow waveform *g(n)* estimated by IAIF (**b**).

to 0.265). From the successfully inverse filtered glottal waveforms, as many consecutive well-formed glottal pulses as possible were selected by visual inspection.

2.4 Parameterisation of the Glottal Waveform

In inverse filtering studies, the ClQ has been widely used for parameterisation of the glottal flow [e.g. Alku and Vilkman, 1996a; Holmberg et al., 1988; Sundberg et al., 1999]. The ClQ is defined as the ratio between the duration of the glottal closing phase and the fundamental period. The ClQ is one of the most important time-domain parameters, as it is affected by the changes of the glottal pulse during its closing phase in which the main excitation of the vocal tract occurs [Fant, 1993]. The value of ClQ reflects changes that occur in the glottal source when, for example, vocal intensity or phonation type is changed [Alku and Vilkman, 1996b; Holmberg et al., 1988].

Amplitude quotient (AQ) is a voice source parameter that represents the ratio of the glottal flow cycle amplitude to maximum negative peak of its first derivative [Alku and Vilkman, 1996a]. It is defined as follows:

$$\mathrm{AQ} = \frac{f_{\mathrm{ac}}}{d_{\mathrm{peak}}} \qquad (1)$$

where f_{ac} is the peak-to-peak flow (ac flow) of the glottal pulse, and d_{peak} is the amplitude of the negative peak of the first derivative of the flow waveform. This ratio was shown by Fant et al. [1994] to yield 'a measure of effective decay time of the glottal flow pulse'.

According to Fant [1997], there is considerable evidence that AQ is one of the most effective parameters for quantifying the characteristics of the voice source waveform with a single numerical value. AQ is simple to acquire since no subjective measurement of opening or closing instants of the glottal flow is needed. However, the AQ values are dependent on the F_0 of the signal. This dependency has been removed in the NAQ. NAQ is a time-domain voice source parameter derived from AQ and closely related to the ClQ [Alku et al., 2002]. It is calculated as follows:

$$\mathrm{NAQ} = \frac{\mathrm{AQ}}{T} = \frac{f_{\mathrm{ac}}}{d_{\mathrm{peak}}T} \qquad (2)$$

where f_{ac} is the peak-to-peak flow (ac flow) of the glottal pulse, d_{peak} is the amplitude of the negative peak of the first derivative of the flow waveform, and T is the period length. A visual representation of the different values is given in figure 3. The maximum range of NAQ values is from 0 to 1,

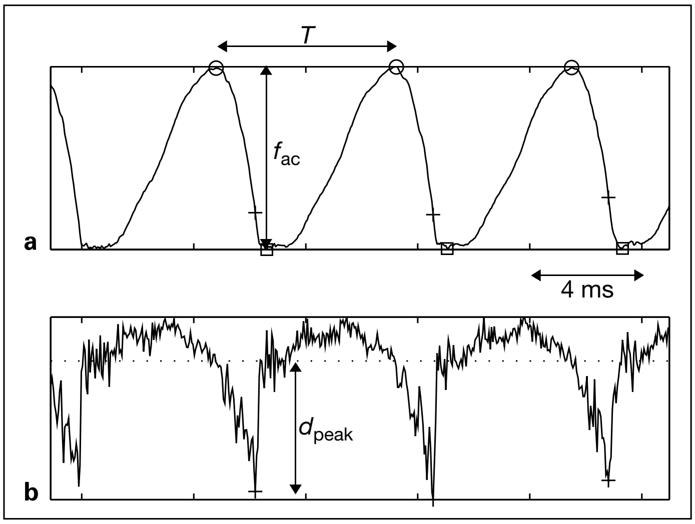

Fig. 3. NAQ values calculated from an inverse filtered glottal flow waveform (**a**) and its first derivative using Eq. (2) (**b**).

with practical values occurring between 0.05 and 0.30 for very pressed and very breathy voices, respectively.

Fant et al. [1994] and Fant [1995, 1997] previously introduced the R_d parameter which, apart from a scaling factor, is identical to NAQ. While this parameter was introduced as a method to reduce the number of the LF parameters in the modelling of the glottal source, it was also noted that it can be acquired directly from an inverse-filtered signal [Fant et al., 1994]. It was soon related to voice spectral properties, intensity, voice typology, gender differences and to articulatory variations [Fant, 1995, 1997]. The relation between the parameters is R_d = 1,000/110 NAQ. According to Fant [1995], the coefficient 1,000/110 'is dimensionless and was chosen so as to provide numerical equivalence between R_d and T_d [another parameter derived from the LF model] data of vowels and consonants derived from Gobl [1988], where F_0 averaged 110 Hz.'

NAQ can be used in the same manner as ClQ in analysing the behaviour of the glottal source in intensity or phonation changes or when the voice is loaded [Bäckström et al., 2002]. It has been shown to be more robust and to have a smaller variance than ClQ, thus providing results with a higher level of accuracy. For these reasons, NAQ parameterisation was used in this study.

The NAQ parameter calculation was implemented in MATLAB. NAQ analysis was included in the IAIF user interface, so that NAQ values were automatically acquired from the selected glottal pulses.

2.5 Auditory Discrimination Test

An auditory discrimination test was constructed to examine whether human listeners are able to discriminate the emotions in the analysed short vowel sections. The test was taken by 10 raters, of which 5 were females. The raters were between 22 and 49 years of age (median 29). All of the raters had normal hearing. The raters were asked to classify the vowel clips according to given emotions. They were presented with a total of 540 vowel clips, of which 54 (10%) were duplicated so that intrarater reliability could be monitored. The raters were asked to categorise the samples into one of the five emotion categories given. The samples were presented in random order, and each sample was repeated with an interval of approximately 250 ms, until a selection was made.

The loudness levels of the samples were equalised by normalising the sample energy levels. The equalisation was performed so that auditory discrimination would be performed using only cues induced by the voice source (F_0 and the voice quality). While it was acknowledged that this method may not be optimal in equalising the auditory perception of loudness, it was chosen for its simplicity and because equalising using auditory models might introduce unexpected effects in the samples. In informal testing, the energy level equalisation appeared to work well.

2.6 Statistical Analyses

Analysis of variance (ANOVA) was used to investigate the sources of variability in the NAQ values. Emotion, subject gender, vowel position within the passage and the ordinal number of recitation were used as the factors, while the NAQ value was the dependent variable. Eta squared (η^2) effect sizes of the factors were also calculated. Tukey's honestly significant difference (Tukey's HSD) was then used as a post hoc analysis to determine significant differences among the emotions. The test performs pairwise comparisons of different emotions and creates confidence intervals on the differences between the NAQ means of the different emotions. The intervals are based on the studentised range statistic.

The hit rate, the relation of the correct answers to the total number of stimuli in a given category, is a commonly used measure in the evaluation of the performance of category judgement studies. Wagner [1993] has noted that the hit rate is not an adequate measure of performance in categorical data analysis due to the fact that it does not regard the number of false alarms or bias in the use of the response categories. He proposed the *unbiased hit rate*, denoted H_u, as an improved measure of accuracy. The unbiased hit rate for stimulus/response category i is calculated from the stimulus/response confusion matrix as follows:

$$H_u(i) = \frac{c_{ii}}{\sum_k c_{ik}} \frac{c_{ii}}{\sum_k c_{ki}} \tag{3}$$

where c_{ij} is a cell in the confusion matrix, with i being the stimulus category and j the response category. As in the hit rate, the values of H_u fall between 0 and 1. To compare performance with that to be expected as a result of chance, unbiased hit rates may be compared with respective unbiased chance proportions, given as follows:

$$p_c(i) = \frac{\sum_k c_{ik}}{N} \frac{\sum_k c_{ki}}{N} \tag{4}$$

where N is the total number of observations. The unbiased hit rates of the raters may be compared to the unbiased chance proportions using pairwise t tests.

The raters' performances in the auditory discrimination test of the study are compared using unbiased hit rates, unbiased chance proportions and pairwise t tests, as suggested by Wagner [1993]. Cohen's kappa was used to assess the intrarater reliability of the raters.

3 Results

The mean, the standard deviation (SD) and the range of the NAQ values were calculated. These are given in table 3. The respective box plots are shown in figure 4. The measured mean of all NAQ values was 0.110 (SD 0.037), ranging from 0.043 to 0.271. The data show that the NAQ mean, when averaged over all emotion types, was approximately of the same order for males (0.103) as for females (0.118). However, the variation of NAQ was clearly larger for females (SD 0.043) than for males (SD 0.031). Both the minimum and the maximum values of NAQ were found to be slightly smaller for males than females.

ANOVA indicated that both emotion and gender had a significant effect on the NAQ value ($p < 0.001$). The results of the ANOVA analysis are given in table 4. In addition, the recitation number had a significant effect on the NAQ value ($p < 0.05$). Further ANOVA analysis was performed on the results of both genders separately. These results are shown in tables 5 and 6. It can be noted that the recitation number has an effect on the NAQ value for males, but not for females.

Table 3. Mean, SD and range of NAQ for different emotions and genders

Gender	Emotion	Mean	SD	Range
All	all	0.110	0.038	0.043–0.271
All	neutral	0.098	0.024	0.057–0.166
All	sadness	0.123	0.044	0.047–0.258
All	joy	0.109	0.032	0.043–0.228
All	anger	0.088	0.022	0.051–0.157
All	tenderness	0.134	0.044	0.069–0.271
Males	all	0.103	0.031	0.043–0.250
Females	all	0.118	0.043	0.055–0.271
Males	neutral	0.097	0.026	0.059–0.166
Males	sadness	0.107	0.034	0.047–0.177
Males	joy	0.100	0.028	0.043–0.181
Males	anger	0.089	0.022	0.051–0.157
Males	tenderness	0.122	0.036	0.069–0.250
Females	neutral	0.098	0.021	0.057–0.154
Females	sadness	0.140	0.047	0.061–0.258
Females	joy	0.120	0.034	0.058–0.228
Females	anger	0.086	0.021	0.055–0.157
Females	tenderness	0.148	0.048	0.079–0.271

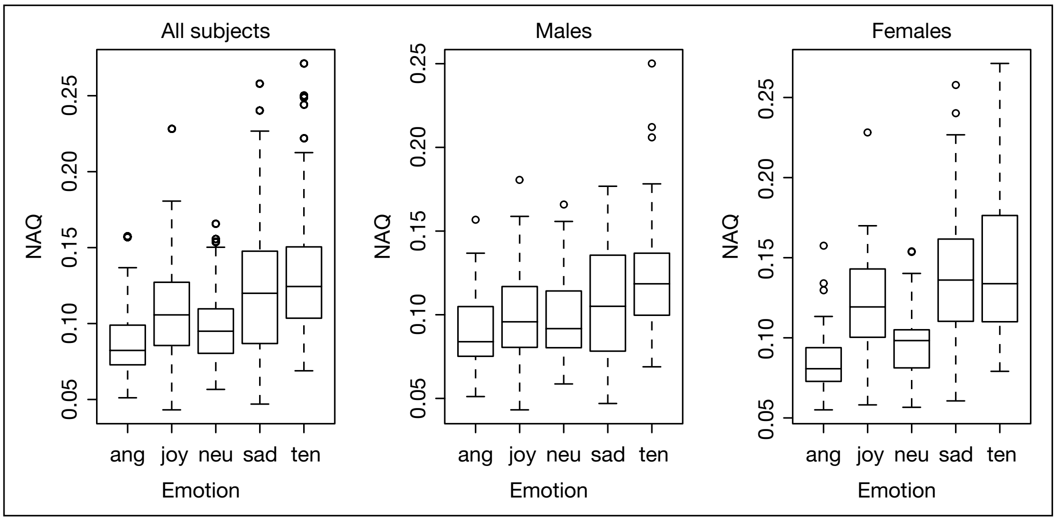

Fig. 4. Box plots of NAQ values with regard to different genders and emotions.

Table 7 shows the results of Tukey's honestly significant difference test. It can be seen that when both genders are combined, there are significant differences between all emotions except neutral-anger, neutral-joy and tenderness-sadness.

By analysing figure 4 and table 7, it becomes apparent that in males only four out of ten emotion pairs differ significantly, with the differing emotions lying at the opposite ends in the box plot. In females, seven out of ten pairs differ significantly. In the combined genders, anger exhibited the smallest average NAQ values of all the emotions although the difference between it and the neutral emotion was not statistically

Table 4. ANOVA of the effect of different factors on the NAQ value

	Df	Sum Sq	Mean Sq	η^2	*F* value	*Pr*(>*F*)	
Emotion	4	0.128	0.032	0.196	30.76	$<2 \cdot 10^{-16}$	***
Gender	1	0.024	0.024	0.038	22.99	$2.2 \cdot 10^{-6}$	***
Recitation number	1	0.007	0.007	0.009	6.36	0.01199	*
Emotion: gender	4	0.023	0.006	0.034	5.50	0.00025	***
Residuals	444	0.463	0.001	0.707			

Statistically insignificant factors are omitted.
Significance codes: 0…0.001***, 0.01…0.05*.

Table 5. ANOVA of the effect of different factors on the NAQ value: male subjects

	Df	Sum Sq	Mean Sq	η^2	*F* value	*Pr*(>*F*)	
Emotion	4	0.0316	0.0079	0.130	9.37	$4.7 \cdot 10^{-7}$	***
Recitation number	1	0.0069	0.0069	0.029	8.25	0.0044	**
Residuals	241	0.2029	0.0008	0.836			

Significance codes: 0…0.001***, 0.001…0.01**.

Table 6. ANOVA of the effect of different factors on the NAQ value: female subjects

	Df	Sum Sq	Mean Sq	η^2	*F* value	*Pr*(>*F*)	
Emotion	4	0.1176	0.0294	0.304	22.90	$1.2 \cdot 10^{-15}$	***
Recitation number	1	0.0009	0.0009	0.002	0.72	0.40	
Residuals	203	0.2605	0.0013	0.673			

Significance code: 0…0.001***.

significant. By contrast, the other three emotions were significantly different statistically from anger. In males, however, the difference between anger and joy was not statistically significant.

Statistically, the neutral emotion differed significantly in the combined genders from two other emotions, namely, tenderness and sadness, but among males, in tenderness only. Likewise, joy differed significantly from anger, sadness and tenderness in the overall group, but among males only in tenderness.

Interestingly, joy exhibited the smallest minimum NAQ values of all, although according to mean and median, it was the most centered emotion.

Sadness scored the second highest average NAQ value and the second highest maximum NAQ value differing significantly, among the combined genders, from anger, joy and neutral emotions while in males it differed only from anger. Moreover, sadness also had the highest standard deviation of NAQ of all emotions. Tenderness, on the other hand, scored the highest NAQ value mean of all emotions, and the standard deviation was nearly as large in tenderness as in sadness. Tenderness alone differed significantly from anger, joy and neutral emotions among both males and females.

Table 7. Multiple comparison of different emotions with regard to their NAQ values

	Joy	Neutral	Sadness	Tenderness	Joy	Neutral	Sadness	Tenderness	Joy	Neutral	Sadness	Tenderness
Angry	X	–	X	X	–	–	X	X	X	–	X	X
Joy		–	X	X		–	–	X		–	X	X
Neutral			X	X			–	X			X	X
Sadness				–				–				–
	Both genders				males				females			

The results are shown for both genders combined, and for males and females separately. The letter 'X' indicates a significant difference between given emotions at 95% confidence level.

Table 8. Confusion matrix and the marginal sums for the auditory discrimination test ratings

Emotion	Rating					Sum
	anger	joy	neutral	sadness	tenderness	
Anger	311	332	349	146	50	1,188
Joy	188	406	339	189	63	1,185
Neutral	109	128	596	207	140	1,180
Sadness	251	127	294	364	150	1,186
Tenderness	214	110	318	306	253	1,201
Sum	1,073	1,103	1,896	1,212	656	5,940

3.1 Auditory Discrimination Test Results

In the auditory discrimination test, the raters categorised 40-ms vowel samples according to the perceived emotion. A confusion matrix for the ratings is given in table 8. The row marginals given in the last column are not constant, but vary slightly due to the random insertion of duplicate samples for the intrarater reliability assessment. The strong bias towards the neutral rating indicated by the column marginals may be neutralised by calculating the estimated conditional distributions for the discrimination test ratings. These are given in table 9. The table gives probabilities for a sample representing different emotions after the rater has chosen a specific rating.

Wagner's unbiased hit rates, as described in Section 2.6, were calculated using the data in table 8. The values were 0.076, 0.126, 0.159, 0.092, and 0.081 for the emotions anger, joy, neutral, sadness, and tenderness, respectively. For comparison, the respective unbiased chance proportions were 0.036, 0.037, 0.063, 0.041, and 0.022. In order to compare the values statistically, unbiased hit rates were calculated for each rater separately. The results are summarised in table 10. These values were compared to the respective unbiased chance proportions, also described in Section 2.6, using a pairwise t test. In every emotion category, the difference between the unbiased hit rate and unbiased chance proportion was found to be significant on a 95% confidence level. The p values for different emotion ratings were 0.00617, $3.63 \cdot 10^{-5}$, $6.844 \cdot 10^{-5}$, 0.000176, and $3.826 \cdot 10^{-5}$ for anger, joy, neutral, sadness, and tenderness, respectively.

Table 9. Estimated conditional distributions for the auditory discrimination test ratings

Emotion	Rating				
	anger	joy	neutral	sadness	tenderness
Anger	0.290	0.301	0.184	0.120	0.076
Joy	0.175	0.368	0.179	0.156	0.096
Neutral	0.102	0.116	0.314	0.171	0.213
Sadness	0.234	0.115	0.155	0.300	0.229
Tenderness	0.199	0.100	0.168	0.252	0.386
Sum	1.0	1.0	1.0	1.0	1.0

For example, when the rating 'anger' is selected, the sample most commonly represents the angry emotion (p = 0.290), next commonly the sad emotion (p = 0.234), and so on.

Table 10. Summary of unbiased hit rates calculated for each rater separately

	Anger	Joy	Neutral	Sadness	Tenderness
Min.	0.021	0.017	0.088	0.044	0.008
1st quartile	0.054	0.111	0.118	0.064	0.046
Median	0.071	0.148	0.163	0.083	0.096
Mean	0.082	0.130	0.166	0.095	0.082
3rd quartile	0.119	0.166	0.196	0.123	0.115
Max.	0.151	0.176	0.285	0.155	0.137

Table 11. Confusion matrix for intrarater reliability assessment

1st classification	2nd classification				
	anger	joy	neutral	sadness	tenderness
Anger	48	9	13	21	5
Joy	11	49	28	7	2
Neutral	15	20	96	18	13
Sadness	20	6	32	49	16
Tenderness	7	2	15	14	24

3.1.1 Discrimination Test Reliability Measures

As described in Section 2.1, 10% of the samples were randomly duplicated for intrarater reliability assessment. Table 11 shows the confusion matrix for the duplicate vowel samples, with rows indicating the classification when the sample was heard for the first time, and columns indicating the classification when the sample was heard for the second time. The confusion matrix indicates 49% of the classifications were in the same category on both the first and the second repetition. Intrarater reliability was measured using Cohen's kappa, which gave a kappa value of 0.35. This suggests only fair agreement between the repetitions.

Inter-rater agreement was measured using Light's kappa. The kappa value for the discrimination test results was 0.00078, indicating poor agreement between the raters.

4 Discussion

As shown in table 3 and in figure 4, the NAQ values for females were larger than for males, suggesting, on average, a smoother glottal pulse. The standard deviation of NAQ for females was also larger (both in absolute and relative terms) than for males, suggesting a wider variation of voice quality for females. These features can also be seen from the box plots in figure 4, as the boxes of males are clumped more tightly together, while both the boxes and their tails are spread across a wider range for females. Although not shown here, the means of the NAQ values also exhibited considerable variation between the subjects.

Regarding the different emotions, anger is the only emotion with a smaller NAQ mean than neutral emotion, although even that difference was found to be statistically null. This is consistent with the findings of Cummings and Clements [1995] with regard to closing duration ratios. Johnstone and Scherer [1999] also reported similar ClQ relationships among tense, irritated, and neutral emotions. Joy, sadness and tenderness all exhibited larger NAQ value means than the neutral, indicating that they were expressed on average with a smoother glottal pulse than neutral. Although Cummings and Clements [1995] did not specifically study these emotions, their results showed similar behaviour between soft and normal speech. Johnstone and Scherer [1999] reported ClQ values for depressed, which was considerably higher than for neutral, also consistent with the present study. However, happiness exhibited a very low ClQ value, which is not consistent with the present study. The considerably higher standard deviation values for joy, sadness and tenderness than for neutral in the present study indicate a greater variation in the voice quality within these emotions. A qualitative inspection of the samples supports these findings, as neutral emotion was nearly always expressed with a modal, even assertive and broadcast-like voice quality.

The relative order of the NAQ value means between different emotions was the same for both males and females. The order from the smallest to the largest was anger, neutral, joy, sadness and tenderness. Importantly, when comparing males and females, it can be readily noted that the variation of the NAQ values among different emotions was greater in females than in males. This is consistent with the findings in the gender differences of the sending accuracy of the facial expression of emotions [Manstead, 1992].

The number of actors in the study was rather small, and individual differences in expression of emotion were found to be rather large. Thus, the possibility remains that for a larger set of actors the results would not be the same. However, since the findings are supported by previous studies, the authors are confident that the results are not markedly incorrect.

Qualitative inspection of figure 4 yields similar results as the ones given by the multiple comparison and table 7. Two very similar emotion pairs, anger-neutral and sadness-tenderness, appear. They have similar NAQ value medians and quartile levels, as well as means and standard deviations. Although not as easily apparent in the box plot, joy-neutral is another emotion pair with non-significantly different NAQ values. Qualitative inspection also reveals the clear gender differences, again consistent with the statistical analyses.

The relationship of voice intensity, F_0 and NAQ has been repeatedly demonstrated. It has been shown by Bäckström et al. [2002] that at SPL values below 80 dB, NAQ decreases as the SPL increases. This behaviour of NAQ is similar to that of ClQ in intensity regulation of speech; both of them demonstrate that raising the vocal intensity typically corresponds to the shortening of the glottal flow closing phase [Bäckström et al., 2002; Holmberg et al., 1988; Sulter and Wit, 1996]. Furthermore, it has also been reported in many studies that raising the vocal intensity corresponds to an increase in F_0 [Sulter and Wit, 1996]. In light of these previous studies on intensity regulation of speech, one can expect NAQ values to correlate with both F_0 and SPL also in expression of vocal emotions. If the present study is compared, for example, to Banse and Scherer [1996], the close similarities of the behaviour of NAQ and SPL for respective emotions become apparent. The similarities of NAQ and F_0 can be seen as well, although differences arise in sadness and between joy and elation. One should, however, not conclude that NAQ is just a superfluous counterpart of the SPL, since NAQ quantifies characteristics of the glottal flow closing phase, which is known to be one of the major means underlying intensity regulation. Therefore, the authors regard SPL as a consequence of the vocal effort, the causes of which can be measured more directly with NAQ at the laryngeal level. This argument can also be justified by studies on perception of speech with varying SPL values. For example, Eriksson and Traunmüller [2002] noted that listeners are able to distinguish changes in vocal effort from changes in listening distance, which suggests that the changes of the SPL alone are mainly perceived as changes in the distance, whereas voice quality variation is required for the perception of vocal effort to change. In addition, some studies of emotional speech synthesis support the conceptual separation of NAQ and SPL. For example, Schröder [1999] noted that when the voice quality is held constant, energy modelling appeared to be a negligible factor for emotional speech synthesis, indirectly indicating the importance of voice quality in the perception of vocal effort and emotions.

Gobl and Ní Chasaide [2003b] studied the associations of different emotions and voice qualities using perceptual testing of synthetic speech samples. They found that emotions with high activation and/or high power (confident, interested, happy, angry, stressed) were associated with tense and harsh voice qualities. Correspondingly, the other non-modal group of their synthetic stimuli (breathy, whispery, creaky, lax-creaky) were associated with emotions with low activation (relaxed, content, intimate, friendly, sad, bored). Interestingly, NAQ values show strong differentiation along that axis: low values correlate with a pressed or tense voice quality, while high values correlate with a breathy or whispery voice quality. While the set of emotions in the present study is somewhat different from that of Gobl and Ní Chasaide [2003b], comparisons can still be made. In their study, angry was strongly associated with tense and harsh voice qualities, while in the present study, anger exhibited very low NAQ values. The NAQ values for joy were larger than those for angry, but smaller than those for sadness or tenderness. In Gobl and Ní Chasaide [2003b], happy was associated with tense and harsh voice qualities, but not quite as strongly as angry and formal, for example. While they did not study neutral emotion, formal and confident could be considered similar in nature to neutral in the present study. Formal and confident emotions exhibited an association with tense and harsh voice qualities, consistent with the low NAQ values for the neutral emotion. In the current study indicating the highest NAQ values together with tenderness, sadness was associated with breathy and whispery voice qualities. Again, Gobl and Ní Chasaide [2003b] did not study tenderness, but intimate may be considered somewhat similar. Intimate was strongly associated with breathy and whispery voice qualities, which is consistent with the high NAQ values of tenderness. Thus, given that the counterparts for neutral and tenderness are similar enough, the results of Gobl and Ní Chasaide [2003b] appear to be supported by the present study for every emotion studied. This also indicates that NAQ would presumably correlate better with activation rather than valence, as suggested by Laukkanen et al. [1996, 1997].

While it appears that NAQ correlates well with emotional and voice quality changes, it has to be noted that it is only a single parameter, and therefore cannot alone represent all the rich features embedded in vocal emotions. The focus of this study was to assess whether NAQ varies with regard to different emotions. Further work is required to assess the relative importance of NAQ and more traditional features such as F_0 and the SPL. However, in the authors' opinion, any single parameter cannot possibly be sufficient to describe the emotional content in speech completely. To a lesser extent, the same also applies to emotional speech synthesis. Although changes in the NAQ parameter alone probably could convey information regarding the activation dimension, NAQ by itself is not sufficient to depict discrete emotions accurately.

The results of the auditory discrimination test indicated considerably lower recognition ratings than those quoted in the literature [for a review, see e.g. Scherer, 2003]. From the confusion matrix in table 8 it can be seen that in anger and tenderness, the correct discrimination result (entry in the diagonal of the matrix) was not even the most common one. In anger, neutral and joy were chosen more often, while in tenderness, neutral and sadness were selected more often. Even in joy, neutral, and sadness, the hit rate is not too impressive. If the evident bias in the ratings is removed by calculating estimated conditional distributions for the ratings as given in table 9, the correct entry becomes most common within any given rating category. In every stimulus/rating category, however, the calculation of unbiased hit rates and their comparison to the respective unbiased chance proportions corroborates that the null hypothesis of the unbiased hit rates not differing from chance proportions has to be rejected on a 95% confidence level. Yet, the chance level inter-rater agreement results indicate that the emotions could not be reliably discriminated. Although the fair intrarater agreement levels suggest that the individual raters were able to perceive some differences in the content of the extremely short samples, these differences might have not correlated with the emotional content of the samples. Even though the number of raters in the discrimination study was quite small, it is not very probable that the results would have changed even in the case of more raters. Several possible reasons for the low recognition rates in the auditory discrimination test emerge. The sample length in the auditory discrimination test was shorter than in any previous work found in the literature. Pollack et al. [1960] reported emotion recognition accuracy of 50% with 60-ms samples. In this work, 40-ms samples were used. The sample energies were equalised to remove loudness cues from the samples. In the literature, loudness has been found to be an important parameter in emotion recognition. Removing this cue would expectedly lower the discrimination accuracy. Also, instead of using explicit scenarios for the emotions or otherwise controlling the emotion induction method, the actors were asked to give simulated emotion portrayals. Without these instructions, many of the given emotions can be considered to be emotion families [Ekman, 1992]. For example, sadness encompasses widely different emotions and moods ranging from mild depression to open desperation. Also, the expression of joy was performed using an expanse of variation, ranging from mild enjoyment to full-blown elation. Anger was also expressed by widely different emotions ranging from mild irritation to hot anger and rage. This wide variation within the stimulus categories naturally increases the difficulty of the discrimination task, especially when no prosodic features or loudness information is available. It should also be noted that this assumed overlapping of stimulus categories would not only degrade the discrimination test results, but hinder acoustic differentiation of the emotions with NAQ or any other acoustic parameter as well.

5 Conclusions

Samples of vocal expression of emotion in continuous speech were gathered using simulated emotion portrayals of professional stage actors. Short continuous speech /a:/ vowel segments of 40 ms duration were extracted from these samples. Glottal source was estimated from the segments using a sophisticated semi-automatic inverse filtering method. In order to quantify the voice source characteristics during the closing phase of the glottal cycle, the acquired glottal flow was then analysed using a robust parameter, the NAQ, which describes the properties of the closing phase of the glottal cycle.

Importantly, it was repeatedly indicated that females express wider variations not only between emotions, but also within the emotions. This result was consistent with previous findings in the gender differences of the sending accuracy of the facial expression of emotions [e.g. Manstead, 1992]. We have no ready explanation for the gender differences on the effect of recitation number. Apparently females were less affected by the preceding emotions in their acting tasks than males.

By comparing the results of this study with some previous publications, it becomes apparent that the NAQ is able to convey information on the activation dimension of the activation-evaluation space. This result differs from those of some other studies, in which the connection of voice quality and valence had been suggested [e.g. Laukkanen et al., 1996, 1997].

The results of the current study are promising in that they show that estimation of the glottal flow from continuous speech can be used for analysis of emotional content of speech. Inverse filtering was successfully applied to obtain voice source signal estimates in emotional speech. This study further showed that estimation of the glottal flow could be performed from continuous speech segments as short as 40 ms. The NAQ parameter could be acquired automatically from the inverse filtered glottal flow and a clear connection between the NAQ values and different emotions was demonstrated. Most emotion pairs also exhibited significant differences with regard to NAQ values. With the development of fully automatic inverse filtering methods, NAQ could become a viable candidate as a feature in automatic recognition of emotions. Hence, the present study indicates that the application of modern inverse filtering techniques together with the NAQ parameter may aid in overcoming the methodological difficulties of voice quality studies applied to the empirical study of vocal expression of emotion.

Acknowledgements

This research was supported by the Academy of Finland (project no. 200 859).

References

Alku, P.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. *11:* 109–118 (1992).

Alku, P.; Bäckström, T.; Vilkman, E.: Normalized amplitude quotient for parameterization of the glottal flow. J. acoust. Soc. Am. *112:* 701–710 (2002).

Alku, P.; Strik, H.; Vilkman, E.: Parabolic spectral parameter: a new method for quantification of the glottal flow. Speech Commun. *22:* 67–79 (1997).

Alku, P.; Tiitinen, H.; Näätänen, R.: A method for generating natural-sounding speech stimuli for cognitive brain research. Clin. Neurophysiol. *110:* 1329–1333 (1999).

Alku, P.; Vilkman, E.: Estimation of the glottal pulseform based on discrete all-pole modeling. Proc. Int. Conf. Spoken Lang. Process. (ICSLP), Yokohama 1994, pp. 1619–1622.

Alku, P.; Vilkman, E.: Amplitude domain quotient of the glottal volume velocity waveform estimated by inverse filtering. Speech Commun. *18:* 131–138 (1996a).

Alku, P.; Vilkman, E.: A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. Folia phoniat. logop. *48:* 240–254 (1996b).

Allen, J.B.; Curtis, T.H.: Automatic extraction of glottal pulses by linear estimation. J. acoust. Soc. Am. *55:* 396 (1973).

Bäckström, T.; Alku, P.; Vilkman, E.: Time-domain parametrization of the closing phase of glottal airflow waveform from voices over a large intensity range. IEEE Trans. Speech Audio Process. *10:* 186–192 (2002).

Banse, R.; Scherer, K.R.: Acoustic profiles in vocal emotion expression. J. Pers. soc. Psychol. *70:* 614–636 (1996).

Childers, D.G.; Lee, C.K.: Vocal quality factors: analysis, synthesis, and perception. J. acoust. Soc. Am. *90:* 2394–2410 (1991).

Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. *70:* 213–220 (1968).

Conger, A.J.: Integration and generalisation of kappas for multiple raters. Psychol. Bull. *88:* 322–328 (1980).

Cowan, J.M.: Pitch and intensity characteristics of stage speech. Arch Speech *1:* suppl., pp. 1–92 (1936).

Cowie, R.; Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Commun. *40:* 5–32 (2003).

Cummings, K.E.; Clements, M.A.: Analysis of glottal waveforms across stress styles. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Albuquerque 1990, pp. 369–372.

Cummings, K.E.; Clements, M.A.: Analysis of the glottal excitation of emotionally styled and stressed speech. J. acoust. Soc. Am. *98:* 88–98 (1995).

Darwin, C.: The expression of the emotions in man and animals; 3rd ed. (Oxford University Press, Oxford 1872, reprint 1998).

Dromey, C.; Stathopoulos, E.T.; Sapienza, C.M.: Glottal airflow and electroglottographic measures of vocal function at multiple intensities. J. Voice *6:* 44–54 (1992).

Ekman, P.: An argument for basic emotions. Cognition Emotion *6:* 169–200 (1992).

Elfenbein, H.A.; Ambady, N.: On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychol. Bull. *128:* 203–235 (2002).

Eriksson, A.; Traunmüller, H.: Perception of vocal effort and distance from the speaker on the basis of vowel utterances. Perception Psychophysics *64:* 131–139 (2002).

Fairbanks, G.: Recent experimental investigations of vocal pitch in speech. J. acoust. Soc. Am. *11:* 457–466 (1940).

Fairbanks, G.; Hoaglin, L.W.: An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monogr. *8:* 85–91 (1941).

Fairbanks, G.; Pronovost, W.: Vocal pitch during simulated emotion. Science *88:* 382–383 (1938).

Fairbanks, G.; Pronovost, W.: An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monogr. *6:* 87–104 (1939).

Fant, G.: Some problems in voice source analysis. Speech Commun. *13:* 7–22 (1993).

Fant, G.: The LF-model revisited: transformations and frequency domain analysis. STL-QPSR, No. 2–3, pp. 119–156 (Speech, Music and Hearing, Royal Institute of Technology, Stockholm 1995).

Fant, G.: The voice source in connected speech. Speech Commun. *22:* 125–139 (1997).

Fant, G.; Kruckenberg, A.; Liljencrants, J.; Båvegård, M.: Voice source parameters in continuous speech: transformation of LF-parameters. Proc. Int. Conf. Spoken Lang. Process. (ICSLP), Yokohama 1994, pp. 1451–1454.

Fant, G.; Liljencrants, J.; Lin, Q.: A four-parameter model of glottal flow. STL-QPSR, No. 4, pp. 1–13 (Speech, Music and Hearing, Royal Institute of Technology, Stockholm 1985).

Frank, M.G.; Stennett, J.: The forced-choice paradigm and the perception of facial expressions of emotion. J. Pers. soc. Psychol. *80:* 75–85 (2001).

Gobl, C.: Voice source dynamics in connected speech. STL-QPSR, No. 1, pp. 123–159 (Speech, Music and Hearing, Royal Institute of Technology, Stockholm 1988).

Gobl, C.: A preliminary study of acoustic voice quality correlates. STL-QPSR, No. 4, pp. 9–21 (Speech, Music and Hearing, Royal Institute of Technology, Stockholm 1989).

Gobl, C.; Ní Chasaide, A.: Amplitude-based source parameters for measuring voice quality. Proc. ISCA VOQUAL ’03 Workshop on Voice Quality: Functions, Analysis and Synthesis, Geneva 2003a, pp. 151–156.

Gobl, C.; Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Speech Commun. *40:* 189–212 (2003b).

Hertegård, S.; Gauffin, J.; Karlsson, I.: Physiological correlates of the inverse filtered flow waveform. J. Voice *6:* 224–234 (1992).

Holmberg, E.B.; Hillman, R.E.; Perkell, J.S.: Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. J. acoust. Soc. Am *84:* 511–1787 (1988).

Javkin, H.R.; Antoñanzas-Barroso, N.; Maddieson, I.: Digital inverse filtering for linguistic research. J. Speech Hear. Res. *30:* 122–129 (1987).

Johnstone, T.; Scherer, K.R.: The effects of emotions on voice quality. Proc. XIVth Int. Congr. Phonetic Sci., San Francisco 1999, pp. 2029–2032.

Kaiser, L.: Communication of affects by single vowels. Synthese *14:* 300–319 (1962).

Laukkanen, A.-M.; Vilkman, E.; Alku, P.; Oksanen, H.: Physical variations related to stress and emotional state: a preliminary study. J. Phonet. *24:* 313–335 (1996).

Laukkanen, A.-M.; Vilkman, E.; Alku, P.; Oksanen, H.: On the perception of emotions in speech: the role of voice quality. Logoped. Phoniatr. Vocol. *22:* 157–168 (1997).

Laver, J.: The phonetic description of voice quality (Cambridge University Press, Cambridge 1980).

Laver, J.; Hanson, R.: Describing the normal voice; in Darby, Speech evaluation in psychiatry*,* pp. 51–78 (Grune & Stratton, New York 1981).

Leinonen, L.; Hiltunen, T.; Linnankoski, I.; Laakso, M.-L.; Aulanko, R.: Vocal communication between species: man and macaque. Lang. Commun. *11:* 241–262 (1991).

Levin, H.; Lord, W.: Speech pitch frequency as an emotional state indicator. IEEE Trans. Syst. Man Cyb. *5:* 259–273 (1975).

Manstead, A.S.R.: Handbook of individual differences: biological perspectives*,* chapter Gender differences in emotion (Wiley, Chichester 1992).

Mataušek, M.R.; Batalov, V.S.: A new approach to the determination of the glottal waveform. IEEE Trans. Acoust. Speech Signal Process. *28:* 616–622 (1980).

Metfessel, M.: Technique for objective studies of the vocal art. Psychol. Monogr. *36:* 1–40 (1926).

Milenkovic, P.: Glottal inverse filtering by joint estimation of an AR system with a linear input model. IEEE Trans. Acoust. Speech Signal Process. *34:* 28–42 (1986).

Miller, R.L.: Nature of the vocal cord wave. J. acoust. Soc. Am. *31:* 667–677 (1959).

Murray, I.R.; Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. acoust. Soc. Am. *93:* 1097–1108 (1993).

Pollack, I.; Rubenstein, H.; Horowitz, A.: Communication of verbal modes of expression. Lang. Speech *3:* 121–130 (1960).

Rothenberg, M.: A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. J. acoust. Soc. Am. *53:* 1632–1645 (1973).

Scherer, K.R.: Speech evaluation in psychiatry*,* chapter Speech and emotional states, pp. 189–203 (Grune & Stratton, New York 1981).

Scherer, K.R.: Vocal affect expression: a review and a model for future research. Psychol. Bull. *99:* 143–165 (1986).

Scherer, K.R.: The neuropsychology of emotion*,* chapter Psychological models of emotion, pp. 137–162 (Oxford University Press, Oxford 2000).

Scherer, K.R.: Vocal communication of emotion: a review of research paradigms. Speech Commun. *40:* 227–256 (2003).

Scherer, K.R.; Banse, R.; Wallbott, H.G.: Emotion inferences from vocal expression correlate across languages and cultures. J. Crosscult. Psychol. *32:* 76–92 (2001).

Schröder, M.: Can emotions be synthesized without controlling voice quality? Phonus, vol. 4, Res. Rep. Inst. Phonet., pp. 37–55 (University of Saarland, 1999).

Skinner, E. R.: A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. Speech Monogr. *2:* 81–137 (1935).

Strube, H.W.: Determination of the instant of glottal closure from the speech wave. J. acoust. Soc. Am. *56:* 1625–1629 (1974).

Sulter, A.M.; Wit, H.P.: Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age. J. acoust. Soc. Am. *100:* 3360–3373 (1996).

Sundberg, J.; Cleveland, T.F.; Stone, R.E., Jr.; Iwarsson, J.: Voice source characteristics in six premier country singers. J. Voice *13:* 168–183 (1999).

Timcke, R.; von Leden, H.; Moore, P.: Laryngeal vibrations: measurements of the glottic wave. Archs. Otolar. *68:* 1–19 (1958).

Titze, I.R.; Sundberg, J.: Vocal intensity in speakers and singers. J. acoust. Soc. Am. *91:* 2936–2946 (1992).

van Bezooyen, R.: Characteristics and Recognizability of Vocal Expressions of Emotion (Dordrecht, Foris, 1984).

Wagner, H.L.: On measuring performance in category judgment studies of nonverbal behavior. J. nonverbal Behav. *17:* 3–28 (1993).

Westermann, R.; Spies, K.; Stahl, G.; Hesse, F.W.: Relative effectiveness and validity of mood induction procedures: a meta-analysis. Eur. J. Soc. Psychol. *26:* 557–580 (1996).

Williams, C.E.; Stevens, K.N.: On determining the emotional state of pilots during flight: an exploratory study. Aerospace Med. *40:* 1369–1372 (1969).

Williams, C.E.; Stevens, K.N.: Emotions and speech: some acoustical correlates. J. acoust. Soc. Am. *52:* 1238–1250 (1972).

Wong, D.Y.; Markel, J.D.; Gray, A.H., Jr.: Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Acoust. Speech Signal Process. *27:* 350–355 (1979).