Time- and Amplitude-Based Voice Source Correlates of Emotional Portrayals

Irena Yanushevskaya, Michelle Tooher, Christer Gobl, and Ailbhe Ní Chasaide

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland
{yanushei,mtooher,cegobl,anichsid}@tcd.ie

Abstract. A detailed analysis of glottal source parameters is presented for emotional portrayals which included both low and high activation states: *neutral*, *bored*, *sad*, and *happy*, *surprised*, *angry*. Time- and amplitude-based glottal source parameters, F0, RG, RK, RA, OQ, FA, EE, and RD were analysed. The results show statistically significant differentiation of all emotions in terms of all the glottal parameters analysed. Results furthermore suggest that the dynamics of the individual parameters are likely to be important in differentiating among the emotions.

Keywords: Voice source parameters, emotion, inverse filtering, LF model.

1 Introduction

The communication of emotion has been a major focus of many in the speech research community over recent years [1]. However, there has been limited analysis of how the voice source varies in emotionally coloured speech, so that we lack a true understanding of how the encoding of emotion is reflected in voice source parameters.

It has been suggested that acoustic correlates of emotion can be described in terms of f0 contour, utterance timing, intensity and voice quality, e.g. [2]. In [1, 4] f0 and temporal characteristics, among other measures, are reported to be correlated with the activation levels of emotions. However, it was also proposed that discrete emotions cannot in fact be modelled using features such as f0 and temporal characteristics alone. [3, 4] comment on the limitations with respect to the acoustic parameters studied in the vocal communication of emotions research and suggest that in order to understand encoding and communication of emotions, more weight must be given to voice quality. The results of listening tests of synthesised tokens of an utterance with different voice qualities, reported in [5], showed that such voice quality differences alone are capable of imparting differences in affective colouring. In some analytic studies [6, 7], a global glottal parameter, NAQ, an amplitude parameter derived in the time domain, has been considered in relation to emotional speech. NAQ claims to provide a similar but more robust measure of the closing phase of a glottal pulse than time-domain parameters. Short vowel segments of acted emotional speech [6], as well as a large corpus of naturally occurring speech [7] have been analysed using the NAQ parameter. In [6], NAQ was found to vary with respect to emotion and gender of the speaker. In [7], NAQ showed consistent voice quality variation to signal

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 159–170, 2007.

paralinguistic information. [8] tested acoustic correlates of emotional speech through formant synthesis and perception tests, and concluded that a combination of parameters is needed to convey emotions. Tests reported in [9] showed that the inclusion of spectral measures, such as drop-off of spectral energy above 1000Hz, spectral flatness measures, and the Hammarberg Index, was beneficial to sinusoidal modelling of expressive speech, whereas inclusion of jitter, shimmer and HNR was found to be relatively unimportant. [10] attempted to simulate seven basic emotions, using prosodic parameters (F0, duration, intensity) and voice quality measures (aspiration noise, jitter, shimmer, OQ, SQ, RQ) suggested in the literature. However, they commented on the lack of published data relating to measurements of the glottal source variation in emotional speech.

The present study focuses on glottal source parameters rather than other measures traditionally employed to quantify vocal expression of emotions, such as, for example, jitter, shimmer and energy distribution of the spectrum [1, 4]. Although the scope is quite limited in terms of the quantity of data and the nature of the emotionally coloured speech analysed, the study aims to contribute towards a more detailed specification of the voice source in the expression of emotion. The data analysed here involved a number of emotion-portraying utterances of a male speaker. The current analysis methods impose many constraints on the type of data we can work with and on the recording materials and conditions. Therefore, it was necessary to work with portrayed rather than naturally occurring emotions, and the results reported here make no claim that these portrayals are representative of truly occurring emotional states. Rather, the study focuses on the more limited objective of detailing how these instances of emotion portrayal are differentiated in terms of glottal source parameters. We are particularly interested in those parameters which may be used for resynthesis, and would hope eventually to be able to synthesise narrations (e.g., of children's stories) where such emotional portrayals would be appropriate.

2 Data and Methodology

The recorded data consisted of repetitions of a short all-voiced utterance 'We were aWAY a YEAR ago' produced by a male speaker of Irish English. The speaker's neutral mode of phonation corresponded to modal voice of Laver's taxonomy [11]. The recording was made in a semi-anechoic chamber; the distance from the microphone was kept constant at 30 cm. The signal was recorded directly to computer, at the sampling frequency 44.1 kHz.

With the help of short frame stories[1], the speaker portrayed the following basic emotions: *sad, happy, bored, surprised, disgusted, angry*, *afraid* as well as the *neutral* state. In the course of the recording, the speaker was advised to keep the peaks of prominence (accents) on the same syllables in all repetitions. For each portrayed emotion, four repetitions were recorded. On the basis of the following auditory

[1] E.g., when portraying 'sad' the speaker had to admit that even though going away on holidays was possible a year ago, dire financial circumstances would not allow such a trip any time soon; in 'neutral' an elderly lady who had left her reading glasses at home asked him to read a product label in a shop for her, etc.

analysis, one repetition was selected for the instrumental analysis; two major factors influencing the choice being the overall quality of the signal and the authenticity of the emotional portrayal. To confirm that the recorded portrayals indeed represented the targeted emotions, a listening test was conducted.

2.1 Listening Test

The stimuli for the listening test consisted of the 7 portrayals of emotions (*happy, sad, bored, angry, surprised, disgusted,* and *afraid*) and three repetitions of the *neutral* utterance. The stimuli were presented to the listeners on a computer screen in randomised order, 5 randomised lists were generated. 16 volunteers participated in the listening test. The participants were asked to listen to each of the sound files as many times as they wanted, to ascertain which emotion is expressed in each one, and to mark their choice by clicking the radio button next to the emotion label listed next to the sound file. The emotive labels included *happy, sad, bored, angry, surprised, disgusted, afraid* and *no emotion*. A blank box was also added to the list of labels so that the listeners could provide their own emotive label should none of the available labels prove adequate. The results of the listening test are presented in Table 1.

Table 1. Results of the perception test (%). Emotions recognised as targeted in 70% or more cases are shown in bold type.

Perceived as Target	Sad	Happy	**Bored**	**Surprised**	Disgusted	**Angry**	Afraid	**Neutral**	Other
Sad	58	-	8	-	1	-	6	21	6
Happy	-	46	-	5	5	12	2	28	2
Bored	18	-	**71**	-	-	-	1	9	1
Surprised	-	6	-	**93**	-	-	1	-	-
Disgusted	1	1	12	6	24	-	-	51	5
Angry	-	1	1	1	2	**91**	-	2	2
Afraid	56	1	1	-	2	1	28	6	5
Neutral	8	3	6	-	6	2	2	**71**	2

The analysis of the listening test showed that the targeted emotion was recognised at a level of 70% or higher in only 4 out of 8 portrayals. The most readily recognised were *surprised* (93%) and *angry* (91%), followed by *bored* and *neutral* (both with 71% 'correct' recognition). *Bored* was perceived as *sad*, but only in 18% of cases. *Sad* received relatively lower recognition rates: only in 58% of cases did the listeners recognise it as such. *Sad* was identified as *neutral* in 21% of cases. *Afraid*, recognised as such in only 28% of cases, was most readily perceived as *sad* (56%). *Disgusted* got the lowest recognition rates compared to the other emotions, only 24% listeners identified it as such, and it was identified primarily as *neutral* (51%). As *disgust* is more often than not expressed in affect bursts rather than in longer utterances and with facial expression acting as a strong recognition cue, its low recognition rates were somewhat expected. *Happy* was also among the emotions that received relatively low recognition rating, 46%, and it was confused principally with *neutral* (28%) or *angry* (12%).

The utterances that were recognised by the listeners as conveying the targeted emotions in more than 70% of all cases were selected for the further analysis involving inverse filtering. Furthermore, despite the relatively lower recognition rates, both *sad* and *happy* were also included in the glottal source parameter analysis, in order to represent a broader range of affective states.

2.2 Data Analysis: Inverse Filtering and Glottal Source Parameters

The selected six utterances (*angry, surprised, bored, sad, happy* and *neutral*) were inverse filtered. Prior to inverse filtering, each sound file was resampled at a sampling frequency of 10 kHz and high-pass filtered using a phase linear high-pass filter with a cut-off frequency of 40 Hz, to ensure the correct zero pressure line.

Each utterance was initially inverse filtered using automatic inverse filtering software based on closed phase covariance LPC, to obtain a first estimate of the differentiated glottal flow. Subsequent manual, interactive fine-tuning of the inverse filter was performed, pulse by pulse, for all utterances. The details of the system are described in [12]. The number of pulses in the utterances varied across the different utterances/emotions, e.g. there were 81 pulses in *bored* and *neutral*, 86 in *sad*, 96 in *surprised*, 105 in *happy* and 129 in *angry*. It should be noted that that the last syllable [ɡo] was excluded from the analysis due to the difficulties the presence of the obstruent posed for the software.

The same interactive software [12] was used to manually fit the LF model to obtain measures of glottal source parameters. The LF model is a well established parametric voice source model, which is described in detail in [13] and [14]. The model matching procedure for extracting data on source parameters is advantageous in that it provides a means for optimisation in both the time and frequency domains. A further reason to use the model is that the LF model is incorporated in the KLSYN88 [15] formant synthesiser, which facilitates the task of resynthesis of emotionally coloured speech.

The following glottal source parameters were included in the analysis: F0, EE, RK, RG, RA, FA, OQ and RD. Note that the first five parameters are sufficient to characterise the basic glottal pulse shape.

F0 is the fundamental frequency and is calculated as the inverse of the glottal pulse duration, T_0. EE is a measure of the strength of the main glottal excitation. RK is a measure of the symmetry/skew of the glottal pulse. A higher RK value indicates a more symmetrical glottal pulse. RG is the glottal frequency FG normalised to F0, i.e. RG = FG/F0, where FG is the characteristic frequency of the glottal pulse during the open phase. RA is T_a normalised to T_0, where T_a is a measure of the effective duration of the return phase after the main excitation, prior to full or maximum glottal closure. Acoustically, its importance lies in its relation to spectral tilt. FA is a parameter related to RA, as it is also a measure capturing spectral tilt. It is inversely proportional to T_a and the FA value indicates the frequency in the source spectrum at which there may be additional downward tilt. Thus, a high FA value indicates a source spectrum with relatively strong higher harmonics. OQ is the duration of the glottal open phase in relation to the duration of the whole glottal period. The OQ value, which can be derived from the parameters RG and RK, is linked to the strength of the lowest harmonics of the source spectrum. RD is a global waveshape parameter [14], which is

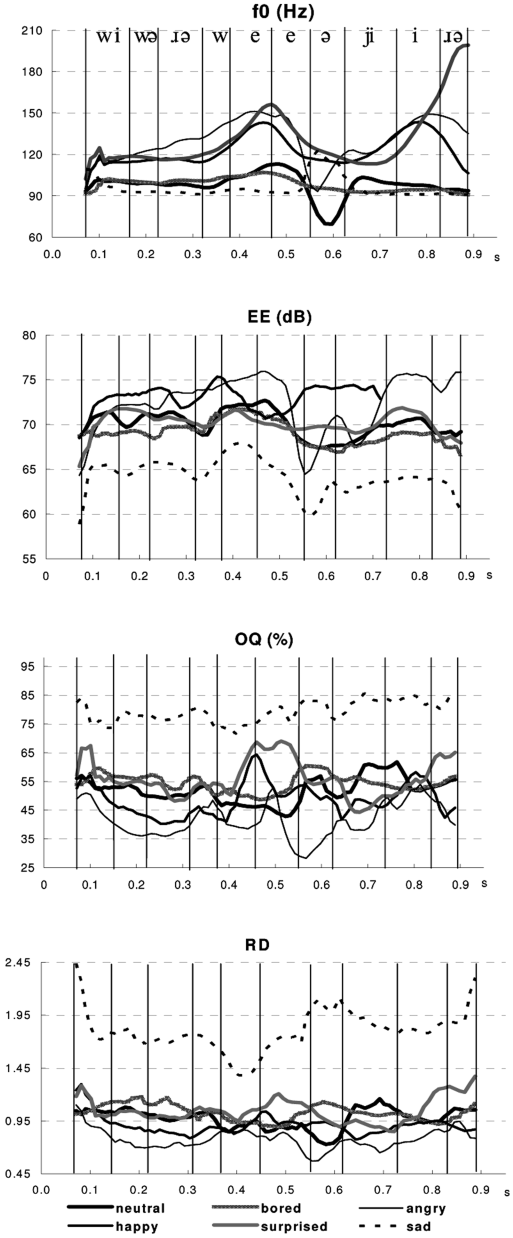

Fig. 1. Dynamics of glottal parameters

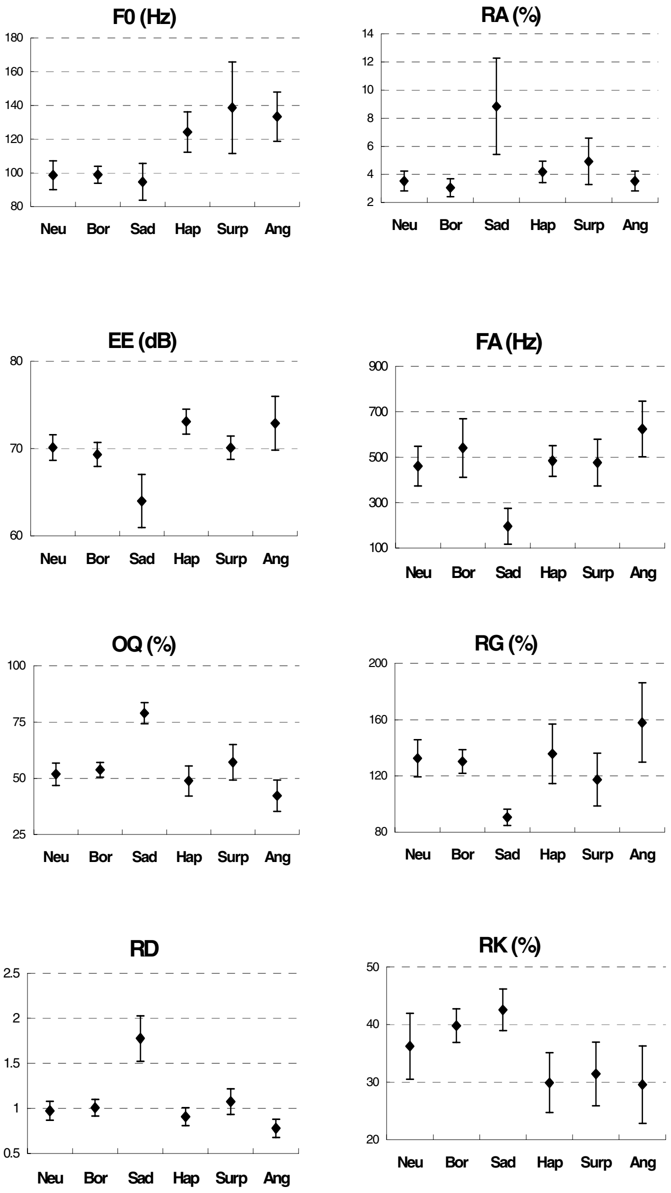

Fig. 2. Mean and standard deviation of glottal parameters across emotions

derived from F0, EE and UP, where UP is the peak amplitude of the glottal flow pulse. It has been suggested that there is a high correlation between the RD value and voice quality variation on the tense to lax continuum [16]. Note that RD is essentially the same as the NAQ parameter [6]. For more detailed descriptions of these parameters and their spectral correlates, see [5, 12].

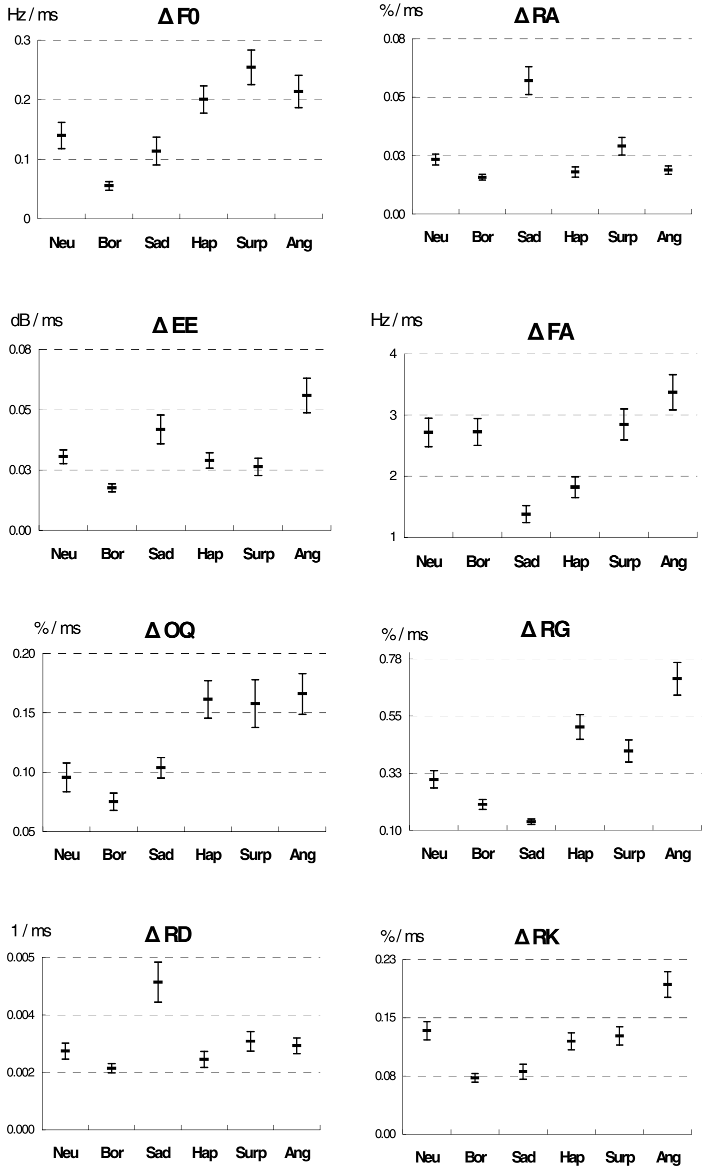

Fig. 3. Mean and standard error of the rate of change of parameters across emotions

2.3 Statistical Analysis and Data Processing for Presentation of Results

In order to aid visual inspection of the parameter dynamics of each emotion-coloured utterance (Fig. 1), the time axis of each utterance was normalised to that of *neutral*. A number of anchor points were chosen in the *neutral* utterance marking syllable boundaries. For vowels from the accented syllables, additional anchor points were placed at their midpoints. For each part of the utterance between these anchor points, 11 in total, the time axis was scaled to be of the same duration as that of the corresponding *neutral* one. As the utterances also had a different number of pulses, linear interpolation was performed in order to plot all emotive utterances to the same time axis points as that of the *neutral* utterance.

To reduce noise caused by small pulse to pulse parameter variation, a moving average of parameters was calculated. The frame spanned 3 pulses with a 1 pulse frame-shift. This served to smooth the plots, while preserving the overall parameter dynamics (see Fig. 1).

For each of the analysed utterances, mean and standard deviation values of each parameter were calculated (see Fig. 2). To explore parameter variability as a function of emotion, a one-way ANOVA with subsequent Tukey's HSD test was conducted (see Table 2). There was a statistically significant difference at the $p<0.05$ level in parameter values for all emotions ($p < 0.0001$): F0 [F (5, 573)=161.31; η_p^2=0.58], EE [F (5, 573)=220.32; η_p^2=0.68], OQ [F (5, 573)=401.41; η_p^2=0.78], RD [F (5, 573)=588.53; η_p^2=0.84], RA [F (5, 573)=158.59; η_p^2=0.58], RG [F (5, 573)=142.51; η_p^2=0.55], RK [F (5, 573)=103.19; η_p^2=0.47], FA [F (5, 573)=195.34; η_p^2=0.63]. High partial eta squared (η_p^2) values are suggestive of significant effect size: over 50% of the variance in mean parameter values is explained by emotion.

The mean values for parameters in Fig. 2 only give a very global picture of variation found for the different emotions. To better capture the parameter dynamics across the duration of the utterance, the rate of change in each parameter was obtained by calculating the first order difference from the smoothed parameter values. Fig. 3 shows the mean and standard error of the absolute rate of change for the analysed glottal parameters.

3 Results and Discussion

ANOVA results showed statistically significant differences among all emotions in terms of all glottal parameters measured. Table 2 details the results of Tukey's HSD test. Fig.1 illustrates the dynamics of the parameters across emotions. Fig. 2 shows the mean and standard deviation for each of the parameters across emotions. Fig. 3 complements by showing the mean and standard error values of the absolute first order difference (rate of change) of glottal parameters.

According to the level of activation, affective states fall into two distinct groups; low activation: *neutral*, *bored* and *sad*, and high activation: *happy*, *surprised* and *angry*. To facilitate the discussion, we will focus the analysis on each of these activation groups with respect to the measured parameters.

Statistical analysis (Table 2) showed that parameters F0, OQ, RD and RK, have relatively high potential to differentiate between high and low activation groups within this dataset. RK is one of the glottal parameters that best differentiates between activation groups, and between emotions within the low activation group. The results observed for F0 were somewhat expected, as F0 is an established correlate of affect,

Table 2. Significance level of the difference in parameters across emotions (* represents p<0.05; p values showing no significant difference between activation groups are in bold type)

F0	neu	bor	sad	hap	surp	**RA**	neu	bor	sad	hap	surp
bor	1.00					bor	0.52				
sad	0.57	0.50				sad	*	*			
hap	*	*	*			hap	**0.09**	*	*		
surp	*	*	*	*		surp	*	*	*	*	
ang	*	*	*	*	0.08	ang	*	**0.38**	*	*	*
EE	neu	bor	sad	hap	surp	**FA**	neu	bor	sad	hap	surp
bor	0.26					bor	*				
sad	*	*				sad	*	*			
hap	*	*	*			hap	**0.79**	*	*		
surp	**1.00**	**0.25**	*	*		surp	**0.97**	*	*	1.00	
ang	*	*	*	0.99	*	ang	*	*	*	*	*
OQ	neu	bor	sad	hap	surp	**RG**	neu	bor	sad	hap	surp
bor	0.34					bor	0.99				
sad	*	*				sad	*	*			
hap	*	*	*			hap	**0.92**	**0.39**	*		
surp	*	*	*	*		surp	*	*	*	*	
ang	*	*	*	*	*	ang	*	*	*	*	*
RD	neu	bor	sad	hap	surp	**RK**	neu	bor	sad	hap	surp
bor	0.85					bor	*				
sad	*	*				sad	*	*			
hap	*	*	*			hap	*	*	*		
surp	*	*	*	*		surp	*	*	*	0.40	
ang	*	*	*	*	*	ang	*	*	*	1.00	0.12

for example [1]. OQ and RD, a global waveshape parameter, differentiated between all emotions, apart from *neutral* and *bored*. The RD data were compared with the NAQ (essentially the same global parameter as RD) values in [6]. In [6], NAQ values for all emotions barring *angry* were higher than *neutral*. Similar patterns were observed here for RD, except *happy* was also lower than *neutral*.

EE, RA, FA and RG were not significantly different for emotions from the two different activation groups. For example, EE of *surprised* is not significantly different from that of *bored* and *neutral*, and RA for *angry* and *bored* yielded similar values. Overlap between emotion groups in EE, RA, FA and RG suggests that no one parameter can provide sufficient information about the voice source in certain affective states and that these parameters in particular should be considered in conjunction with other glottal parameters.

Analysis of parameter values suggests that within-group emotion differentiation is parameter specific, some parameters distinguishing between emotions more clearly than others. Emotion differentiation is also activation-group specific. Within the high activation group, emotions are differentiated in terms of OQ, RD, RA and RG, whereas within the low activation group the only differentiator is RK. All parameters except RK showed no statistically significant difference between *bored* and *neutral*, which in itself is noteworthy.

It is obvious not only from the means in Fig. 2, but also from the smoothed parameter trajectories in Fig. 1, that parameter values of *angry* and *sad* almost always appear as extremes (highest and lowest), each representative of the most extreme emotion within their relevant activation group. It is interesting to note that these affective states have been associated with very different voice qualities, which is supported by the combination of parameters found here.

It is often the case that when parameters show no emotion differentiation in terms of mean values, it is the parameter dynamics that distinguish emotions. It is especially evident in the example of *bored* and *neutral* that yield similar mean values for the majority of parameters (except RK), but show very different rates of change, parameters for *neutral* being of more dynamic nature. Within the high activation group, for example, *angry* is not differentiated from *happy* and *surprised* in terms of mean RK, but it shows markedly higher rate of change for this parameter.

In emotion research literature, there has been an increasing awareness that to achieve affective sounding speech synthesis, we must adequately represent the contribution of the voice source. The results presented here show that we must take into account how the voice source parameters vary, not only in broad terms but also in terms of their dynamics. It goes without saying that ultimately this information needs to be combined with other variables such as speech rate, duration, etc., which are also important in the signalling of emotion. Similar suggestions were made earlier, e.g. [8].

A summary of parameter combinations for each emotion that could also serve as a first approximation of parameter settings for each affect is shown in Table 3. Parameter levels are calculated as a percentage difference relative to *neutral.* These parameters can be readily converted into synthesis parameters, for example those of the KLSYN88 [15] formant synthesiser. Note that the combination of parameter settings is different for each emotion and we would tentatively conclude that the voice source difference between affective states will not be captured by a single measure or a combination of static parameters. Rather, a combination of dynamically varying parameters needs to be considered.

Table 3. Suggested levels for LF-parameters for synthesis: summary* (shaded are other parameters considered in the paper)

Affect		**F0**	**EE**	**RG**	**RK**	**RA**	**OQ**	**FA**	**RD**
Low activation	Neutral	M	M	M	**M**	M	M	M	M
	Bored	M	M	M	H	L	M	H	M
	Sad	M	**L**	LL	H	**HH**	HH	LL	**HH**
High activation	Happy	**H**	M	**H**	**L**	H	**L**	M	L
	Surprised	**HH**	M	**L**	**L**	HH	**H**	M	H
	Angry	**HH**	**M**	**H**	**L**	M	**L**	**HH**	L

* LL = [< -25%] (very low), L = [-25%, -5%] (lower than *neutral*), M = [-5%, 5%] (within the *neutral* range), H = [5%, 25%] (higher than *neutral*), HH = [> 25%] (very high). Bold type shows parameters demonstrating high dynamic variation.

4 Conclusion

Glottal parameters were analysed for emotion-portraying utterances, with the aim of 1) describing how glottal parameters can vary across the emotions, and 2) identifying which glottal parameters or combination of such may be more important for the resynthesis of emotion. The parameters chosen were a combination of those that give an overall picture of the glottal pulse and those that can be incorporated into synthesis of emotive speech. As already mentioned, the analysis is on limited data, and results can only be considered tentative. The parameters demonstrate unequal potential in emotion differentiation. Parameters that differentiate between activation groups are F0, RK, RD and OQ. Four out of eight parameters analysed – RA, FA, RG, and EE demonstrate overlap between the activation groups. *Bored* and *neutral* show similar parameter values, except for RK and FA. *Sad* is significantly different from all other emotions in terms of all parameters except F0 (F0 *sad* is similar to *neutral* and *bored*).

In the high activation group, there is good within-group differentiation in terms of the mean values of RA, RG, RD and OQ, but no differentiation in F0 (*angry* and *surprised*), RK (all high activation emotions), EE (*angry* and *happy*) and FA (*happy* and *surprised*). It is obvious that parameter dynamics should be considered, as they further demonstrate differences among emotions. For example, *neutral* and *bored* are not differentiated in terms of mean values of all parameters except one (RK), but are well differentiated in terms of the rate of change. Glottal parameters can be used to describe voice quality patterns pertaining to various emotions, especially those that prove to demonstrate good differentiation between high and low activation emotions. However, one cannot expect to reproduce emotion in synthesis by setting voice source parameter values to static values, the dynamics must also be considered.

The results presented here are only preliminary, and much further work will be required that will include analysis of more speech samples obtained from more speakers, with a broader range of affective states, where arousal differences within emotions of the same family are taken into account in the emotion elicitation process. One aim of the present research is to provide data for the synthesis of narrated stories, capable of generating at least a small repertoire of narrator's emotions, something we hope to explore in future work.

Acknowledgments. The research was funded by the EU Sixth Framework Network of Excellence HUMAINE and by the Irish Research Council for Science and Engineering Technology-EMBARK initiative.

References

1. Juslin, P., Scherer, K.R.: Vocal Expression of Affect. In: Harrigan, J., Rosenthal, R., Scherer, K.R. (eds.) The New Handbook of Methods in Nonverbal Behavior Research, pp. 65–135. Oxford University Press, Oxford (2005)
2. Murray, I., Arnott, J.: Toward 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 (1993)
3. Scherer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication 40, 227–256 (2003)
4. Banse, R., Scherer, K.R.: Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology 70(3), 614–636 (1996)
5. Gobl, C., Ní Chasaide, A.: The Role of Voice Quality in Communicating Emotion, Mood and Attitude. Speech Communication 40, 189–212 (2003)

6. Airas, M., Alku, P.: Emotions in Short Vowel Segments: Effects of the Glottal Flow as Reflected by the Normalised Amplitude Quotient. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 13–24. Springer, Heidelberg (2004)
7. Campbell, N., Mokhtari, P.: Voice Quality: the 4th Prosodic Dimension. In: Proceedings of the 15th International Congress of Phonetic Sciences, pp. 2417–2420 (2003)
8. Burkhardt, F., Sendlmeier, W.F.: Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis. In: Proc. ISCA Workshop (IRTW) on Speech and Emotion, pp. 151–156 (2000)
9. Drioli, C., Tisato, G., Cosi, P., Tesser, F.: Emotions and Voice Quality: Experiments with Sinusoidal Modelling. ITRW VOQUAL'03, Switzerland, pp. 127–132 (2003)
10. Cabral, J.P., Oliveira, L.C.: EmoVoice: a System to Generate Emotion in Speech. Interspeech 2006, Pittsburgh (2006)
11. Laver, J.: The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge (1980)
12. Gobl, C., Ní Chasaide, A.: Techniques for Analysis the Voice Source. In: Hardcastle, W.J., Hewlett, N. (eds.) Coarticulation: Theory, Data and Techniques, pp. 300–320. Cambridge University Press, Cambridge (1999)
13. Fant, G., Liljencrants, J., Lin, Q.: A Four Parameter Model of Glottal Flow. STL-QPSR, Speech, Music and Hearing, Royal Institute of Technology, Stockholm 1, 1–13 (1985)
14. Fant, G.: The LF-model Revisited: Transformations and Frequency Domain Analysis. STL-QPSR, Speech, Music and Hearing, Royal Institute of Technology, Stockholm 156, 2–3 (1995)
15. Klatt, D.H., Klatt, L.C.: Analysis, Synthesis and Perception of Voice Quality Variations among Male and Female Talkers. Journal of the Acoustical Society of America 87, 820–856 (1990)
16. Fant, G.: The Voice Source in Connected Speech. Speech Communication 22, 125–139 (1997)