VOICE PARAMETER DYNAMICS IN PORTRAYED EMOTIONS

I. Yanushevskaya, C. Gobl, A. Ní Chasaide
Centre for Language and Communication Studies, Trinity College Dublin, Ireland

Abstract: **This paper is concerned with voice source variation associated with different emotional portrayals of an utterance: *bored, sad, happy, surprised, angry* and *neutral*. The source analyses involved pulse-by-pulse inverse filtering to yield the differentiated glottal flow, and subsequent parameterisation of the source signal using the LF model. The glottal source parameters included in the analysis were F0, EE, RK, RG, RA, FA, OQ and RD. For the data set analysed, each emotion seems to have its own distinct pattern of source parameter settings. Analysis of the dynamics of the source variation illustrated here on the RD parameter suggests that to better understand source variation we need to study it in terms of the prosodic components of the utterance.**

Keywords : **Voice source, dynamics, emotion**

I. INTRODUCTION

This paper deals with voice source variation which is associated with different emotional portrayals of an utterance. Our broad concern is the study of how voice source parameters vary as a function of linguistic (prosodic and segmental) aspects of an utterance [1, 2], as well as how such source differences may signal affective states. Here we consider source variation for a single utterance produced by a male speaker, repeated so as to convey six targeted affective states: *bored, sad, happy, surprised, angry* and *neutral*. The sentence read by the speaker was 'We were aWAY a YEAR ago', and the stressed syllables are shown in capitals.

Note that we do not claim that these represent 'true emotions' as might occur in spontaneous interactions, but rather the type of portrayals one might use when, for example, reading a bedtime story to a child. As such they represent 'feigned' emotion, but we would argue that such feigned emotion is not only part and parcel of narrative reading but is also used in discourse, e.g., when a mother feigns being cross to influence a child's behaviour or when one feigns being calm while truly agitated in a stressful social encounter. In many true-life situations effective social interactions depend more on the ability to feign emotion than to reveal true underlying emotion. Further example of the use of feigned emotion in discourse is when tone-of-voice is mismatched to the utterance as a humorous device or to express sarcasm, etc.

II. METHOD

The source analyses used involved pulse-by-pulse inverse filtering to yield the differentiated glottal flow, and subsequent parameterisation of the source signal using the LF model [3]. These techniques involve a manual interactive analysis system and are described in [4], as are the source parameters. The glottal source parameters included in the analysis were F0, EE, RK, RG, RA, FA, OQ and RD. F0 is the fundamental frequency. EE is a measure of the strength of the main glottal excitation. RK is a measure of the skew of the glottal pulse; e.g., a higher RK value indicates a more symmetrical glottal pulse. RG is the glottal frequency FG normalised to F0, where FG is the characteristic frequency of the glottal pulse during the open phase. RA and FA are related parameters capturing spectral tilt. Thus, a high FA (or low RA) 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, and is linked to the strength of the lowest harmonics of the source spectrum. RD is a global wave shape parameter, and is thought to be highly correlated with voice quality variation on the tense to lax continuum [5, 6].

III. RESULTS AND DISCUSSION

A. *Overall vocal parameter settings*

Fig. 1 illustrates the global source parameter settings for the different affective states extending the preliminary analysis reported in [7]. Note that for the single utterance in question, depending on the emotion expressed, 81 to 120 individual glottal pulses were analysed. Parameter levels are calculated as a percentage difference relative to the *neutral*, based on mean values for the entire utterance.

The scaling allows one to see the extent to which a particular parameter deviates from the *neutral*: from -2 = [< -25% of neutral value set at 0] (very low) to +2 = [> 25% of neutral value] (very high). Note also that filled black circles show parameters demonstrating relatively high dynamic variation as indicated by the mean rate of change (delta) values. These were obtained by calculating

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 6th international workshop: December 14-16, 2009*, ISBN 978-88-6453-094-9 (print), 978-88-6453-096-3 (online),

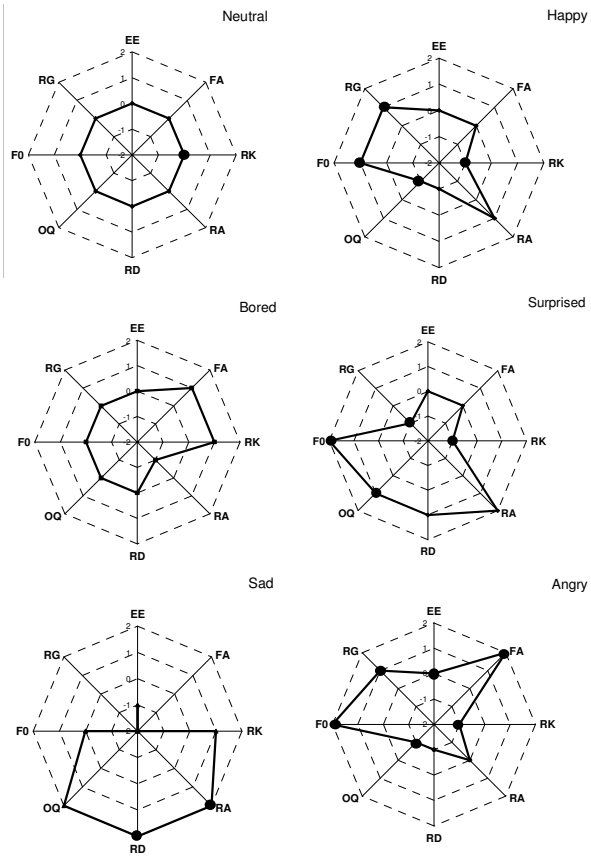

Fig. 1. *Levels for glottal parameters for different emotions: -2 = [< -25%] (very low), -1 = [-25%, -5%] (lower than neutral), 0 = [-5%, 5%] (within the neutral range), 1 = [5%, 25%] (higher than neutral), 2 = [> 25%] (very high). Filled black circles show parameters demonstrating relatively high dynamic variation.*

the first order difference from the smoothed parameter values. The smoothing, which allowed us to decrease the amount of pulse-to-pulse noise while preserving the overall parameter dynamics, involved calculating the moving average of parameter values with a three pulse frame and a one pulse frame-shift.

As evident in Fig. 1, the combination of parameter settings is different for each emotion. In these renditions, each emotion seems to have its own distinct pattern.

Sad relative to *neutral* shows an overall pattern of weak glottal pulses (very low EE), very leaky, breathy voice quality (as suggested by the very high RA, RD and OQ parameters) and large attenuation of high frequency components in the signal (very low FA).

Surprised shares some of these characteristics of the *sad* repetition. Although there is an indication of greater breathiness (rather high OQ, RD and RA values), there is overall less weakening of the glottal pulse excitation, or of the higher frequencies in the signal (more modal-like EE and FA values). It also has strikingly high mean F0.

Angry and *happy* both show broadly opposite deviations from the *neutral* baselines, as can be deduced from the generally upward shift in parameter values. Note that *happy* and *angry* are frequently confused in perception experiments on vocal expression. The raised RG and the lowered OQ, RD and RK values suggest an overall more tense voice quality setting. *Angry*, however, differs here from *happy* in having extreme FA and F0 values, suggesting more extreme vocal tension.

Bored differs least from the *neutral* setting, showing mainly somewhat more strength in the higher frequencies (FA/RA values).

B. Source dynamics

As mentioned above, filled black circles in Fig. 1 denote dynamic variation. As can be seen from the prevalence of such circles in the case of *happy, surprised* and *angry*, there is more dynamic variation in source parameters than for the relatively low activation states of *sad, bored* and even *neutral*. This seems intuitively in keeping with what we might expect for these more aroused high activation states.

Although Fig. 1 gives some idea of the global trends for these different renditions of the utterance, it does not adequately show the considerable dynamic variation in source parameters in the course of the utterance. To illustrate this, in Fig. 2 we show the dynamic course of the RD parameter for these utterances (as represented by the smoothed parameter trajectories). Note that RD tends to be viewed as indicative of the tenseness/laxness of the voice [5, 6]. To facilitate the inspection and comparison of parameter trajectories, the time axis of each

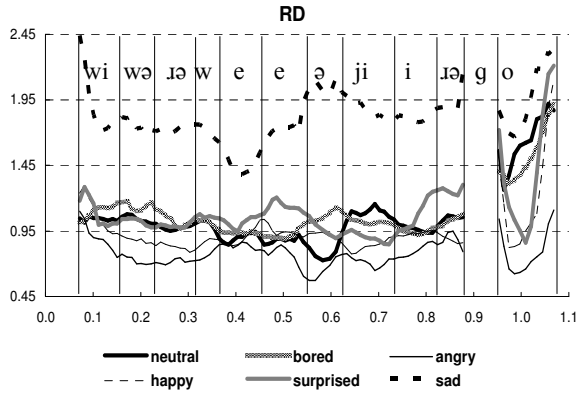

Fig. 2. *Dynamic variation of the RD parameter for different emotions across the utterance.*

emotionally coloured utterance was normalised to that of *neutral* according to a number of anchor points (shown in Fig. 2 as vertical lines). These included utterance and syllable boundaries as well as midpoints of the vowels in the accented syllables and the approximate boundary of [w] and [e] in the stressed syllable WAY. The [g] segment was excluded from the analysis as it had not been consistently realised as voiced across the utterances. For each part of the utterance between anchor points, the time axis was scaled to be of the same duration as the corresponding *neutral* one. As the utterances had a different number of pulses linear interpolation was used to plot all utterances to the same time axis points as the *neutral* utterance.

Note that in Fig. 2 RD values overall are much higher for the *sad* repetition (a very lax quality) and highest (tense) for the *angry* repetition. Also evident are complex parameter variation over time depending on the segmental characteristics of the utterance (consonants tend to lower RD values as higher degree of constriction in the vocal tract have upstream influences on vocal fold vibration). The large differentiation among the emotions in the final syllable of the utterance, as well as the rapid increase in RD values at the end of the utterance are likely to be linked to the realisation differences in the final accent as well as to the transition into breathiness as the vocal folds open prepausally. This suggests that affect related voice differences may be strongly anchored to prosodically important aspects of the utterance. In the *sad* utterance, RD dips in the accented vowels of WAY and YEAR, indicating a less breathy quality. Across all the emotions looked at here there is a distinct trend for these vowels to be associated with relatively strong glottal pulses with more stable parameter settings.

Fig. 3 provides further information on this last point. It shows the mean and standard deviation values of the RD parameter (panel A), as well as the rate of change (delta values) of RD in the course of the utterance, for each affect with an indication of the standard error (panel B). Note that while *bored* and *neutral* have similar means and standard deviations, there is overall much less dynamic variation of the *bored* RD values.

The mean and standard deviation of parameter values are shown separately for the stressed and unstressed syllables in each affect in Fig. 3 (panel C). Similarly, in panel D, the mean rate of change of the RD parameter is shown separately for the stressed and unstressed syllables together with the standard error values. This illustrates again the point made above that for this parameter, although the average values do not differ greatly in the stressed/unstressed conditions, there is considerably more dynamic variation in the unstressed than in the stressed syllables.

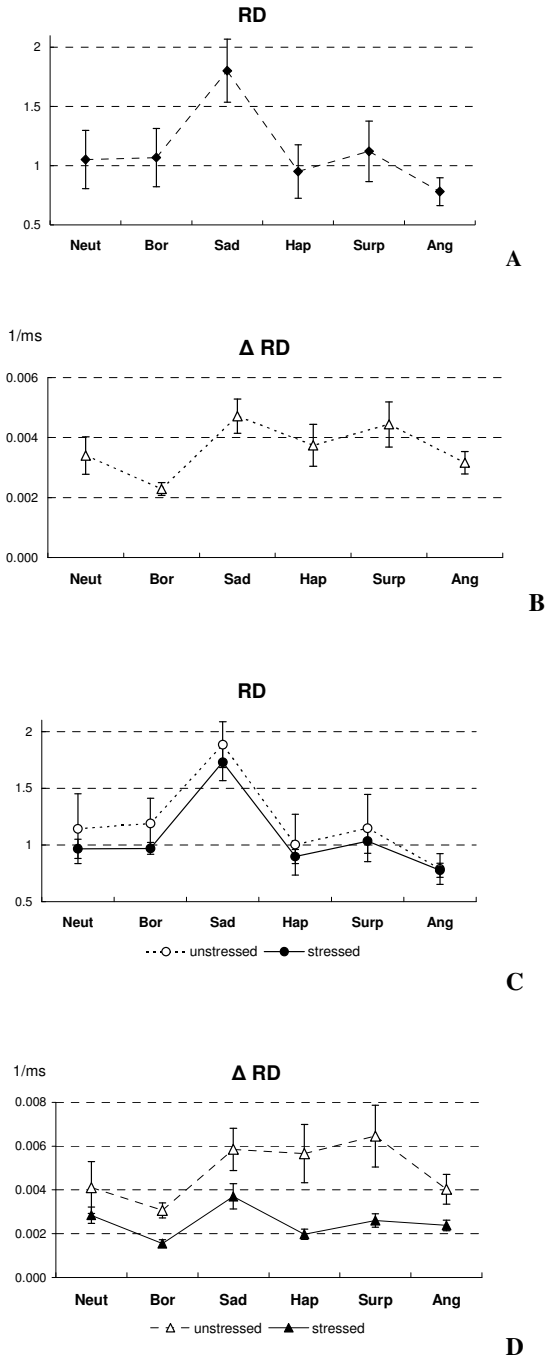

Fig. 3. *RD parameter for different emotions: A- mean values, B - mean delta (rate of change) values, C – mean values for stressed and unstressed syllables, D - mean delta values for stressed and unstressed syllables.*

IV. Conclusion

Although based on very limited sample utterances, we feel that a detailed study can nonetheless yield new insights and prompt us to look at what might be important in the analysis of source variation. Similarly we would argue that being able to analyse (and eventually hopefully to resynthesise) these kinds of simulated portrayed emotions might have many applications in the use of speech technology.

This illustration highlights the need to look closely at the utterance internal dynamics of source variation. We would suggest that these dynamics will be best understood if studied in terms of the prosodic components of the utterance. Differences between stressed and unstressed syllables have been pointed out and there are indicators in the present data that accentuation and in particular the nucleus and the post-nucleus material may be of particular importance. In our future work we hope to examine in greater detail the linkage between prosodic structure and voice source variation, as a basis for understanding how and where source variation may signal affect.

Acknowledgements

The authors acknowledge the stimulating interaction with voice researchers in COST 2103, as well as the EU-funded Network of Excellence on Emotion, HUMAINE, through which this research was partially funded.

References

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

[2] A. Ní Chasaide and C. Gobl, "Voice quality and f0 in prosody: towards a holistic account," in *Speech Prosody 2004*, Nara, Japan, 2004.

[3] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR* vol. 4, pp. 1-13, 1985.

[4] C. Gobl and A. Ní Chasaide, "Techniques for investigating laryngeal articulation (Section B: Techniques for analysing the voice source)," in *Coarticulation: Theory, Data and Techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge University Press, 1999, pp. 300-321.

[5] G. Fant, "The LF-model revisited: transformations and frequency domain analysis," *STL-QPSR* vol. 2-3, pp. 119-156, 1995.

[6] C. Gobl and A. Ní Chasaide, "Amplitude-based source parameters for measuring voice quality," in *VOQUAL'03*, Geneva, Switzerland, 2003, pp. 151-156.

[7] I. Yanushevskaya, M. Tooher, C. Gobl, and A. Ní Chasaide, "Time- and amplitude-based voice source correlates of emotional portrayals," in *Affective Computing and Intelligent Interaction: Proceedings of the ACII 2007*. vol. 4738, A. Paiva, R. Prada, and R. W. Picard, Eds. Lisbon, Portugal: Springer-Verlag, 2007, pp. 159-170.