The Relationship between Voice Source Parameters and the Maxima Dispersion Quotient (MDQ)

Christer Gobl, Irena Yanushevskaya, Ailbhe Ní Chasaide

Phonetics and Speech Laboratory, Trinity College Dublin, Ireland

cegobl@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie

Abstract

This study examines the relationship between the Maxima Dispersion Quotient (MDQ), a recently proposed measure of the tense-lax dimension of voice quality, and voice source parameters manually measured from glottal flow data, where there is linguistically and paralinguistically determined voice source modulation. MDQ was found to correlate most closely to the open quotient (OQ) and the RD parameter. The paralinguistically varying data, which involved more extensive voice source modulation than the linguistic, also showed a higher degree of correlation with these parameters, and higher correlations overall with a range of voice source parameters. The high correlation with OQ and RD found in these analyses would suggest that MDQ can be a useful, additional parameter for the analysis of glottal source dynamics.

Index Terms: Maxima Dispersion Quotient (MDQ), glottal source, voice source parameters, voice quality, voice prosody

10.21437/Interspeech.2015-506

1. Introduction

This paper looks at the correlation of the recently proposed voice parameter, the Maxima Dispersion Quotient (MDQ) with voice source parameters, measured from speech data where the voice source varies as a function of both linguistic and paralinguistic factors. The MDQ parameter was proposed by [1], and is suggested as a parameter that should yield a measure of the tense-lax dimension of voice quality. It is part of a system we are developing, GlóRí [2], which aims to incorporate more robust and accurate methods for voice source analysis [2-6].

Accurate analysis of voice source variation is essential to our study of the communicative functions of the voice – the 'voice prosody'. We aim to model within a single framework, both the linguistic (intonation-related) voice modulation [7-9], as well as the paralinguistic, affect-related modulation [10-12]. The linguistic voice modulations present in our view a *baseline* voice prosody which is further perturbed or modified for paralinguistic signalling. The linguistically determined voice modulations, though considerable, are typically not extreme, and are heard not as shifts in voice quality per se, but rather as an inherent part of the prosody of the utterance. In contrast, paralinguistically relevant shifts can be quite extreme and tend to impinge on the listener's consciousness as changes in voice quality – in layman terms, tone of voice. From this perspective, both aspects together constitute the voice prosody (of which intonation is a major aspect), essential to both the linguistic and affective content of the message. It should be noted that, as voice prosody is realised relative to the individual speaker's intrinsic voice quality, long term characteristics of the speaker's voice also need to be taken account of.

The biggest roadblock in describing voice prosody is the difficulty in obtaining accurate voice source measures, capable of capturing accurately even fine-grained modulations of the voice. In most of our studies to date, we have employed a two-step procedure described in [13]: (i) inverse filtering of the speech pressure waveform, to yield the glottal flow wave, and (ii) measurement of voice source parameters, on the basis of a voice source model (the LF model [14]) matched to the derivative of the glottal flow.

Automatic methods for inverse filtering and source parameter extraction tend to be inaccurate, especially for the analysis of connected speech, and for non-modal qualities. For that reason our analyses mostly use an interactive procedure whereby manual pulse-by-pulse editing is carried out (for details, see [13]). Though this procedure yields greater accuracy, it has drawbacks: it requires a high degree of experimenter skill and is hugely time consuming, so that only limited quantities of data can thus be analysed. Furthermore, the inverse filtering technique requires stringently high standards in recording conditions, which are often not met. In this context, the MDQ parameter has been proposed as a voice measure that might be reasonably robust even in less-than-perfect recording conditions [1]. Developing more accurate and robust automatic analysis is a major aim of the GlóRí system which is under development [2]: progress in this area would allow for analysis of large corpora, and will open up many fields of potential application.

The aims of the present pilot study is to examine (i) to what extent MDQ correlates with specific voice source measures, (ii) for which voice source parameters, the highest correlations are found, and (iii) whether the correlations are different in the linguistic and paralinguistic datasets, given that the ranges might differ for these.

2. The MDQ parameter

The Maxima Dispersion Quotient (MDQ) was recently put forward as a measure for capturing differences in voice quality [1]. Experiments showed that this parameter is effective in the discrimination and categorisation of voice qualities such as breathy, modal and tense voice.

In image processing, the maxima of the output signals produced from wavelet-based filtering have been found to be effective in the detection of edges, as these maxima tend to appear in the vicinity of edges. This property is exploited in the calculation of the MDQ parameter, which involves wavelet-based filtering of the glottal excitation signal, derived using autocorrelation LPC and inverse filtering. The assumption is that if the main excitation produced by the glottal pulse

is very sharp and impulse-like, the maxima will appear near the time point of the excitation. This type of excitation would be expected for tense voice. Lax or breathy voice, however, involves less rapid change in the glottal airflow as the vocal folds close, resulting in a much less impulse-like excitation. For this type of excitation, one would therefore expect the maxima from the wavelet decomposition to be more dispersed.

The calculation of the MDQ parameter is fully automatic. First, the time points of the main glottal excitations are determined using the GCI (Glottal Closure Instant) detection algorithm SE-VQ [5]. The LPC residual is then derived using autocorrelation LPC. This residual represents an estimate of the glottal source signal, which is subsequently processed by a dyadic wavelet transform. Here seven scaled versions of the wavelet function are used, which results in an octave band, zero-phase filter bank, with filter centre-frequencies of 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz.

For each glottal excitation found by the GCI detection algorithm, a search interval is defined. The locations of the maxima are determined within this interval, and their distances (durations) from the time point of the excitation are measured. The mean of these durations is then calculated and, finally, the MDQ value is obtained by normalising this mean value to the glottal period. For further details, see [1].

MDQ is essentially a measure linked to the sharpness of the glottal excitation. Of the voice source parameters we use, the ones which most directly capture specific characteristics of the glottal excitation are EE, RA and RK (see Section 3.2). Therefore, one might expect to find strong correlations with these parameters.

In [1], the MDQ parameter was evaluated in classification tests using data (isolated vowels, and voiced sentences) which had been produced with tense, modal and breathy voice quality. For the evaluation, these data were 'screened' by trained listeners, so that only utterances were selected where they agreed with a high degree of confidence that the voice quality was distinctly *breathy*, *modal* or *tense.* MDQ was found to yield good separation of these qualities for the isolated vowel data. Results were less clear-cut for the sentences. Note that in running speech (sentences) there is considerable prosodic modulation of the voice source, not found in steady state vowel productions.

In the present study MDQ is correlated with voice source data, not elicited (or judged) in terms of intended (or perceived) voice quality as such. Rather, the data all involved continuous speech (sentences), where the voice was modulated as a function of varying linguistic or paralinguistic factors.

3. Test data and parameters

3.1. Speech data

Two sets of data were examined, where the inverse filtering and source parameterisation were carried out using the manual interactive methodology mentioned above. The analysis procedures are fully described in [13]. In the first dataset, linguistic prosody was varied (we term this the Linguistic dataset). The all-voiced sentence 'WE WERE aWAY a YEAR ago' was elicited so that in different repetitions, focus fell on one or other of the potentially accented syllables, shown in caps. Source modulations associated with focus are described in [7, 8]. For this dataset, we recorded six male speakers with four focus conditions, and each was produced with either a falling (six speakers) or a rising (five speakers) intonation contour. Additionally, the utterance was elicited with broad focus and as deaccented (both with falling pitch pattern). The dataset comprised 56 utterances (6 speakers × 6 sentences × 1 falling intonation contour + 5 speakers × 4 utterances × 1 rising intonation contour) with a total of 5829 glottal pulses.

The second dataset involved paralinguistic variation (the Paralinguistic dataset) recorded for a single speaker (one of the speakers recorded for the first dataset), and involved repetitions of the same all-voiced sentence, 'We were away a year ago', produced so as to portray differing emotions/affective states. These portrayed emotions included *angry, surprised, sad, bored* as well as a *neutral* rendition. The different 'emotive' renderings were strongly differentiated in terms of voice quality [10]. The dataset comprised 5 utterances totalling 649 glottal pulses. Note that in comparing paralinguistic and linguistic results, we limit ourselves to the single speaker for whom both datasets were available.

3.2. Voice source parameters

The analysis involved manual pulse-by-pulse inverse filtering and subsequent source parameterisation of the full utterances using the software systems described in [13, 15]. In addition to the fundamental frequency, f_0, the following voice source parameters were extracted:

EE (excitation strength), which is the negative amplitude of the differentiated glottal flow signal at the time point of maximum waveform discontinuity. The EE value is closely related to the overall strength of the glottal excitation.

UP (peak glottal flow), which is a measure of the maximum glottal airflow rate of the glottal pulse.

RA (return phase), which is the normalised effective duration of the return phase of the glottal pulse after the main excitation. The RA value relates to the source spectral slope, a higher RA value corresponding to a greater spectral slope.

RG (normalised glottal frequency), which is a measure of the characteristic frequency of the glottal pulse (FG), normalised to f_0. RG mainly affects the relative amplitudes of the first couple of harmonics of the source spectrum.

RK (glottal pulse symmetry), which is defined as the duration of the closing portion of the pulse normalised to the duration of the opening portion of the pulse. Thus, a smaller RK value means a more skewed glottal pulse, whereas a greater RK reflects a more symmetrical pulse.

OQ (open quotient), which is a measure of the duration of the open phase (excluding the return phase) of the glottal pulse normalised to the glottal period. This definition of the open quotient is determined by RG and RK according to OQ = (1+RK)/(2RG). Therefore OQ tends to be positively correlated with RK and negatively correlated with RG. Mainly the amplitudes of the lower components of the source spectrum are affected by changes in OQ.

RD, which is a global waveshape parameter that captures some of the main features of the glottal pulse in one single measure. It is derived from f_0, EE and UP as follows: $(1/0.11)\times(f_0 \cdot \mathrm{UP/EE})$, where UP/EE is equivalent to the glottal pulse declination time during the closing phase of the glottal cycle. The scale factor (0.11^{-1}) makes the numerical value of RD equal to the declination time in milliseconds when f_0 is 110 Hz. For further details, see [16-19, 13].

3.3. MDQ estimation and statistical analysis

For the correlation analysis, MDQ values were automatically estimated using the automatic component of the GlóRí system [2]. Prior to the correlation analyses, the data were smoothed by first applying median filtering, followed by moving average filtering (both filters spanned 5 pulses).

Pearson product-moment correlation coefficients (Pearson's *r*, two-tailed) were computed to explore the correlation between MDQ and the other parameters.

4. Results: Correlations of MDQ and voice source parameters

The *r* values for the Linguistic and Paralinguistic datasets are shown in Table 1. On the left are shown values for the entire linguistic dataset with six speakers. The two rightmost panels compare values for Linguistic and Paralinguistic data for the single speaker for whom both sets were available.

Figure 1 shows for the six-speaker Linguistic dataset, the correlations for the individual voice source parameters. The highest correlations were found with the parameters OQ (r = 0.72), RD (r = 0.67) and RG (r = –0.62). Our initial expectation that EE, RA and RK would be the highly correlated with MDQ was not borne out: these yielded moderate to relatively low correlations, EE (r = –0.49), RA (r = 0.47) and RK (r = 0.27). Note that, even where the correlations are relatively weak, they are all significant.

Table 1. Voice source parameter correlations with MDQ for the 6-speaker Linguistic dataset (left) and the 1-speaker Linguistic and Paralinguistic datasets (two rightmost columns). *Correlations are significant at p< 0.01.

Parameter	Linguistic (6 speakers) Pearson's *r*	Linguistic (1 speaker) Pearson's *r*	Paralinguistic (1 speaker) Pearson's *r*
F0 (Hz)	0.40*	0.54*	-0.27*
EE (dB)	-0.49*	-0.48*	-0.65*
UP (dB)	-0.46*	-0.59*	-0.07
RD	0.67*	0.69*	0.79*
OQ (%)	0.72*	0.67*	0.79*
RG (%)	-0.62*	-0.58*	-0.73*
RA (%)	0.47*	0.59*	0.58*
RK (%)	0.27*	0.14*	0.64*

Figure 2 shows MDQ correlations with source parameters for the Paralinguistic dataset (in red) superimposed on the Linguistic dataset (in green) for the single speaker. It is striking that correlations overall are higher in the Paralinguistic dataset. For OQ, r = 0.79 compared to 0.67 in the Linguistic dataset; for RD, r = 0.79 compared to 0.69; for RG, r = –0.73 compared to –0.58. Furthermore, the parameters EE, RA and RK, which did not show strong correlations with MDQ in the Linguistic dataset, now show a relatively high degree of correlation: for EE, r = –0.65 compared to –0.48; for RK, r = 0.64 compared to 0.14. However, the correlation of MDQ with RA was rather similar for the two datasets: r = 0.58 (Paralinguistic dataset) compared to 0.59 (Linguistic dataset). As in the Linguistic dataset, all parameters in the paralinguistic data showed highly significant correlations with MDQ, with the striking exception of UP, which is not found to be correlated in the paralinguistic set: r = 0.07 (Paralinguistic dataset) compared to 0.59 (Linguistic dataset).

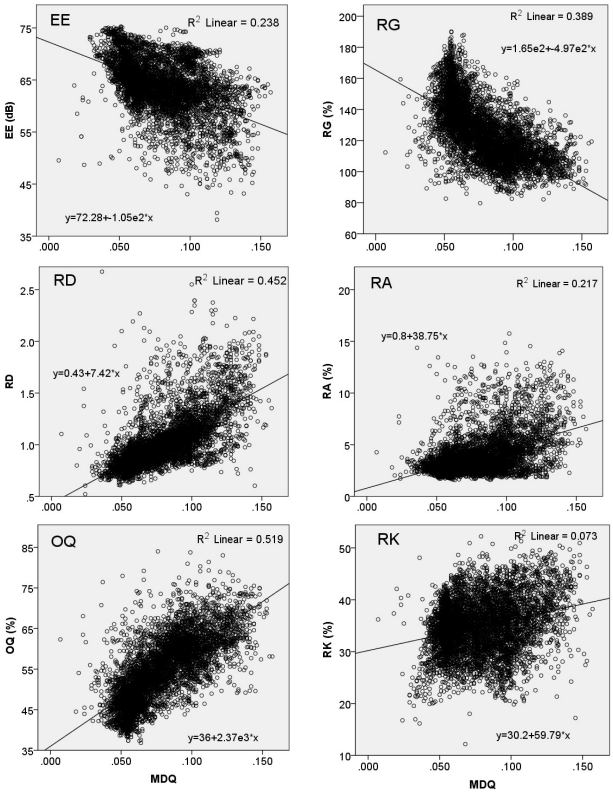

Figure 1: Correlations of MDQ and selected voice source parameters in the Linguistic dataset.

5. Discussion

Overall, the strongest correlations with MDQ were observed for the OQ and RD parameters, with RG also showing a high correlation. The initial expectation that those parameters most closely relating to the sharpness of the glottal excitation, i.e. EE, RA and RK, would be strongly correlated with MDQ was not clearly borne out, at least not for the Linguistic dataset. One reason for this could be it is the combined contribution of these parameters that determine the overall sharpness of the excitation and that the individual parameters are capturing only part of this. Both OQ and RD capture multiple aspects of the glottal pulse (see Section 3.2), which could partly explain the high correlation values obtained for these parameters.

It is also striking in Figure 2 and Table 1 that the correlations are higher for the Paralinguistic dataset. In this dataset, it is also clear from Figure 2 that the range of voice source modulation is greater. Note in Figure 2 that RD, OQ and RK values extend to lower values, while RG values go higher. These changes suggest particularly more use of tenser voice in the Paralinguistic dataset. This shows up equally in the extended lower end of the MDQ values for this dataset.

In addition to the extension of the parameter ranges in the paralinguistic data, there are differences in the distribution of values relative to what is found in the Linguistic dataset. Figure 3 illustrates for the linguistic vs. paralinguistic single-speaker datasets the distribution of values for OQ and RG, the two parameters for which the differences were most striking. The histograms have been superimposed to facilitate comparison. Note that whereas for the Linguistic dataset there is a compact distribution around the mean, in the Paralinguistic dataset, there is a greater use of the extreme ends of the range.

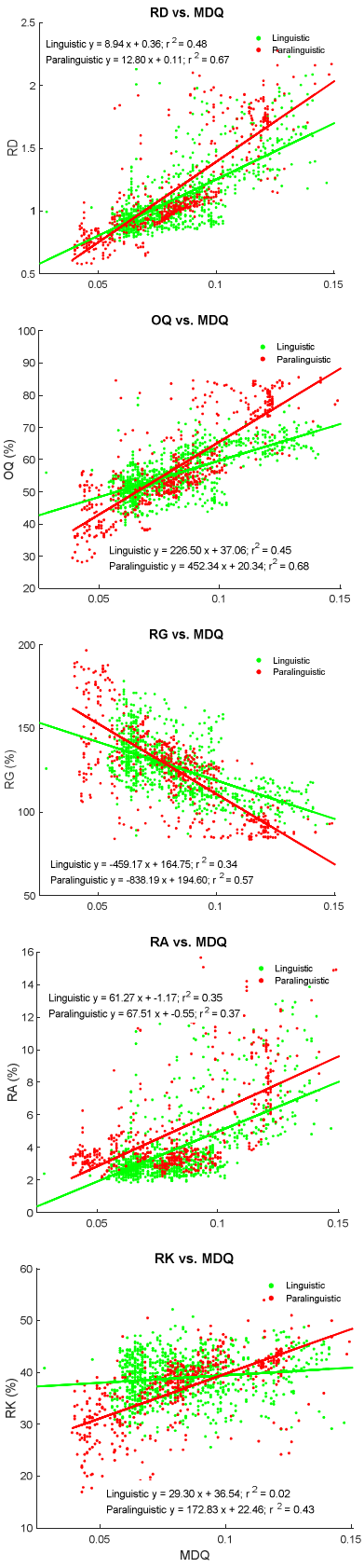

Figure 2: Correlations of MDQ and selected voice source parameters in the linguistic dataset (green) in comparison to the paralinguistic dataset (red).

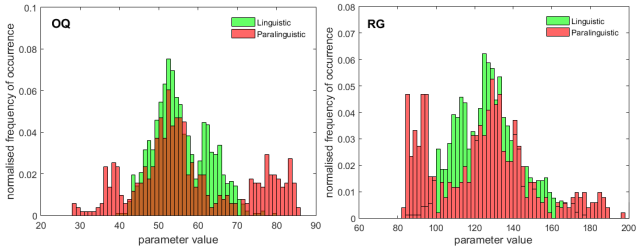

Figure 3: Histograms of the OQ and RG parameters in the Linguistic (green) and Paralinguistic (red) datasets.

As was mentioned earlier, paralinguistic modulations tend to be consciously heard as voice quality shifts (tone of voice changes) on the part of the speaker, while the modulations associated with linguistic prosody are simply heard as part of the prosody, signalling in this case, information structure. These observed correlations do indicate that the MDQ measure is indeed picking up on voice quality differences, particularly when they involve the rather larger effects associated with paralinguistic signalling of emotion and affect.

In Table 1, we note a further striking difference in the correlations for the two datasets. In the Linguistic dataset, there is a strong significant positive correlation of MDQ with f_0. Broadly speaking, this would appear to suggest that, on the whole, higher f_0 is associated with a laxer voice quality. For the Paralinguistic dataset, there is a much weaker, negative correlation, suggesting that there is less linkage of f_0 and voice quality, and that insofar as they are correlated, higher pitch is correlated with a tenser voice quality.

In some ways, it seems intuitively right that the linguistic prosody should be more closely coupled to f_0 modulation. After all, virtually all linguistic research on linguistic prosody is focussed on pitch variation. Although the f_0 modulation does not predict voice source modulation as such, it makes sense that they would be rather more closely correlated. It is interesting that for the paralinguistically varying data, such a correlation no longer holds, and that the polarity is different. We would tentatively suggest, on the basis of these findings that the voice source modulation in linguistic prosody works synergistically with intonation structure, whereas for paralinguistic prosody there is a large degree of decoupling of the intonational and voice quality dimensions of the voice.

6. Conclusions

This study found that the voice parameter MDQ correlated particularly with the OQ and RD voice source parameters. Results also indicate that the correlation is stronger when there is more extensive variation in the voice source, as with the Paralinguistic dataset analysed here. Overall, these results are encouraging, and suggest that MDQ is likely to prove useful insights into the behaviour of the glottal source signal. The fact that correlations are less good with the more limited Linguistic dataset does nonetheless suggest that for now, one must proceed with caution. Further analyses will ascertain to what extent it may capture the finer details of voice source dynamics.

7. Acknowledgements

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET) and the ABAIR project funded by the Department of Arts, Heritage and the Gaeltacht, Ireland.

8. References

[1] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, pp. 1170-1179, 2013.

[2] J. Dalton, J. Kane, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "GlóRí - the glottal research instrument," presented at the Speech Prosody 2014, Dublin, Ireland, 2014.

[3] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech & Language,* vol. 27, pp. 1028-1047, 6// 2013.

[4] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication,* vol. 55, pp. 397-414, 2013.

[5] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication,* vol. 55, pp. 295-314, 2013.

[6] C. Gobl and A. Ní Chasaide, "Amplitude-based source parameters for measuring voice quality," presented at the VOQUAL'03, Geneva, Switzerland, 2003.

[7] I. Yanushevskaya, C. Gobl, J. Kane, and A. Ní Chasaide, "An exploration of voice source correlates of focus," presented at the Interspeech 2010, Makuhari, Japan, 2010.

[8] A. Ní Chasaide, I. Yanushevskaya, and C. Gobl, "Voice source dynamics in intonation," presented at the XVIIth International Congress of Phonetic Sciences, Hong Kong, China, 2011.

[9] A. Ní Chasaide and C. Gobl, "Decomposing linguistic and affective components of phonatory quality," presented at the Interspeech 2004, Jeju Island, Korea, 2004.

[10] I. Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Voice parameter dynamics in portrayed emotions," presented at the 6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA 2009), Florence, Italy, 2009.

[11] I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Universal and language-specific perception of affect from voice," presented at the XVIIth International Congress of Phonetic Sciences, Hong Kong, China, 2011.

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

[13] C. Gobl and A. Ní Chasaide, "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds., 2 ed Oxford: Blackwell Publishing Ltd, 2010, pp. 378-423.

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

[15] C. Gobl and J. Mahshie, "Inverse filtering of nasalized vowels using synthesized speech," *Journal of Voice*, vol. 27, pp. 155-169, 2013.

[16] G. Fant, "The voice source in connected speech," *Speech Communication,* vol. 22, pp. 125-139, 1997.

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

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

[19] C. Gobl, "Exploring voice source dynamics and its signalling function in speech: techniques and data," *Proceedings of the 6th International Conference on Voice Physiology and Biomechanics ICVPB 2008*, Tampere, Finland, pp. 27-49, 2008.