Publication II

Matti Airas, Paavo Alku, and Martti Vainio, "Laryngeal voice quality changes in expression of prominence in continuous speech." In *Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions in Biomedical Applications (MAVEBA 2007)*, pp. 135–138, Florence, Italy, December 13–15, 2007.

LARYNGEAL VOICE QUALITY CHANGES IN EXPRESSION OF PROMINENCE IN CONTINUOUS SPEECH

M. Airas[1], P. Alku[1], M. Vainio[2]

[1] Helsinki University of Technology, Espoo, Finland
[2] University of Helsinki, Finland

Abstract: **In this study three different prominence and speech melody related effects on voice quality were studied using an inverse filtering based method. The hypothesis that prominence as a function of sentence and word stress is signaled with more pressed voice quality was tested. The results indicated discrepancies in the parameterization results, and the original hypothesis could not be confirmed. Instead, it is suggested that prominence is expressed with a more breathy voice quality. A possible physiological explanation for the phenomenon is also provided.**
Keywords: **voice inverse filtering, prominence**

I I

Local increase of the voice fundamental frequency and attenuation of the respective changes elsewhere is used to signal prominence relations within utterances, phrases, words, or series of utterances. Fundamental frequency variation, however, is not the sole effect in the expression of prominence. Several studies have reported spectral or glottalization changes as an effect of prominence, suggesting that the changes are due to a tenser *voice quality* in prominent vowels [6].

Although Laver's original definition of voice quality [10] attributed it to be caused by both laryngeal and supralaryngeal features of the voice production mechanism, it is nowadays often restricted to only reflect the laryngeal settings of speech. The major physiological source of these changes, in turn, is represented by the airflow generated by the vibrating vocal folds, the *glottal flow*. Unfortunately, direct measurement of this major source of voice quality is not possible from continuous speech due to the hidden position of vocal folds, located deep within the larynx and surrounded by several vital organs. Hence, the only feasible means to estimate the glottal flow from speech is to use a technique called *inverse filtering*. This implies that resonances of the vocal tract are cancelled from the speech pressure signal by feeding it through anti-resonances which have been defined from the underlying speech spectrum. The glottal flow is then parameterized using some time, amplitude, frequency, or model-based techniques to gather numerical data of the studied phenomena.

Research on the function of the glottal flow has concentrated mainly on isolated vowels. In contrast to this, there is surprisingly little evidence on how the glottal flow as the source of voice quality behaves in the sentence and word level in expression of stress. One such study was performed by Gobl, who studied LF model parameter fitting in a sentence, the focus of which varied [7]. The excitation parameter E_e values were found to be larger in focal context, indicating voice quality changes in the expression of focus. Swerts and Veldhuis reported some evidence for correlation of F0 and voice quality expressed by the first harmonic amplitude difference (ΔH_{12}, or H1-H2) in the context of speech melody [13]. However, they also cited several other studies regarding F0 and OQ, in which contradictory views were presented, i.e. F0 and OQ did not correlate, or exhibited negative correlation. Epstein stated that speakers use voice quality, as expressed by LF model parameter changes, to distinguish between prominent and non-prominent words in declarative and interrogative sentences [6].

The understanding of the behaviour of voice quality in expression of stress is limited to large extent by the lack of relevant methodologies to analyze the glottal flow from continuous speech. In order to address this issue, the current study utilizes TKK Aparat, a unified voice inverse filtering and parameterization package. Using this sophisticated speech research tool, the authors further tested the hypothesis that stress is expressed with a pressed voice quality. Hence, this study extends on the previous works on the topic which have utilized only a handful of speakers with restricted utterances or model-based parameters. This is performed using continuous speech and robust voice source parameters together with statistical analyses.

II M M

Speech of healthy, native Finnish speakers was recorded. There were 11 speakers in total, of which 6 were women. The ages of the subjects ranged from 18 to 48 years, mean being 30 years. Two of the speakers smoked regularly or irregularly, while the others were non-smokers.

The recordings were performed in an anechoic chamber. The speakers were standing, reciting the text from a paper attached on a sheet of cardboard.

The speakers were equipped with a headset microphone consisting of a unidirectional Sennheiser electret capsule. The microphone signal was routed through a microphone preamplifier and a mixer to iRiver iHP-140 digital audio recorder. Low-frequency phase distortion introduced by the digital recorder was corrected by acquiring the input impulse response of the device using an MLS measurement [12] and convolving the recorded signals using a time-reversed version of the impulse response.

The speech material consisted of three passages of Finnish text describing past weather conditions. The material was selected so that there were multiple [ɑ] vowels with different levels of prominence suitable for inverse filtering. The three different speech melody–and hence prominence–related conditions using a long [ɑ] or [æ] segment were chosen as follows: (1) A paragraph initial con-

tent word with a relatively high F0 in a lexically stressed syllable (sentence stress condition). (2) The same segment in a later repetition of the word (word stress condition). (3) A long [ɑ] in a lexically unstressed position. Each recitation took about one minute, and was repeated three times. The middle recitation was chosen for further processing. Three vowels from each passage, a total of nine, were then marked using Praat [4]. In total there were $3 \cdot 3 \cdot 11 = 99$ marked vowels.

The phase-corrected recordings were high-pass filtered to remove any low-frequency noise in the signal and then cut into separate files containing only single vowels using the time instants marked in Praat. Further processing of the segmented files was performed using TKK Aparat, which is a comprehensive voice inverse filtering and parameterization software package, supporting two different inverse filtering methods, a multitude of time, amplitude, frequency, and model-based glottal flow parameters, seamless interoperation with the MATLAB environment and easy exporting of data to statistical software packages [1].

The separated vowel files were inverse filtered using the iterative adaptive inverse filtering (IAIF) algorithm [3]. The flow diagram of the current version of IAIF, which is a slightly modified version from the previous ones, is shown in Fig. 1. Most notably, parametric spectral models used in various blocks of the flow diagram are computed with the discrete all-pole modeling (DAP) method [5] instead of the conventional linear predictive analysis. This reduces the bias of the harmonic structure of the speech spectrum in the formant frequency estimates. In block no. 1 of Fig. 1, the speech signal is high-pass filtered using a linear-phase FIR filter to reduce any low frequency fluctuations captured during the recordings. Stages 2–6 form the first glottal flow approximation by making an estimate of the vocal tract transfer function and inverse filtering the signal with that estimate. The first approximation is used as a basis for stages 7–12, which roughly repeat the process of the earlier stages to yield the final glottal flow estimate.

The inverse filtering process yields glottal flow estimates, an example of which is shown in Fig. 2. The glottal flow parameters of the vowel segments were computed automatically from the estimated glottal flow. Even though all parameters implemented in TKK Aparat were acquired, further analysis was restricted to only NAQ and AQ parameters. NAQ, the normalized amplitude quotient, and AQ, the amplitude quotient, measure time-domain characteristics of the glottal closing phase from two amplitude-domain quantities [2]. AQ is defined as $\mathrm{AQ} = \frac{A_{\mathrm{ac}}}{d_{\mathrm{min}}}$, where A_{ac} is the maximum AC amplitude of the flow and d_{min} is the minimum of the flow derivative. Correspondingly, NAQ is defined as $\mathrm{NAQ} = \frac{\mathrm{AQ}}{T_0}$, where T_0 is the period length. Both AQ and NAQ correlate well to the pressedness of voice, which contributes considerably to the voice quality.

The effect of various factors on the NAQ and AQ values was tested using analysis of variance (ANOVA). First, the values were log-transformed to correct the skew in parameter distributions. Then, ANOVA was performed using the vowel running number, speaker sex, sentence stress, and word stress as dependent variables and the log-transformed AQ as the independent variable. All statistical treatments

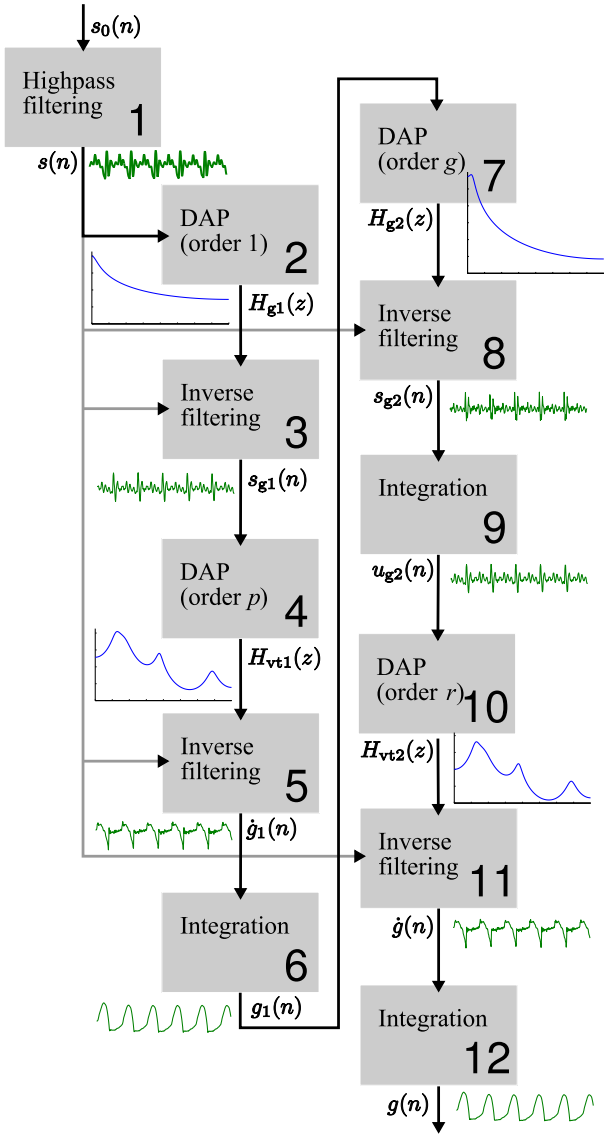

Figure 1: *The block diagram of the IAIF method for the estimation of the glottal excitation $g(n)$ from the speech signal $s_0(n)$. For further details of the different stages, please refer to [3].*

were performed using the R statistical software environment [9].

III R

In general, inverse filtering analysis of continuous speech is problematic due to, for example, rapid changes in formant frequencies. In spite of these inherent difficulties, the analyses conducted in the current study were successful and reliable estimates of the glottal flow could be computed with the IAIF method for all the intended samples. Furthermore, during the inverse filtering process, a subjective quality evaluation on a scale of 0–3 was given for each glottal flow estimate using the general shape of the resulting glottal flow estimate as the criterion. This evaluation yielded a mean value of 2.4, which is considerably higher than in other inverse filtering studies conducted by the authors.

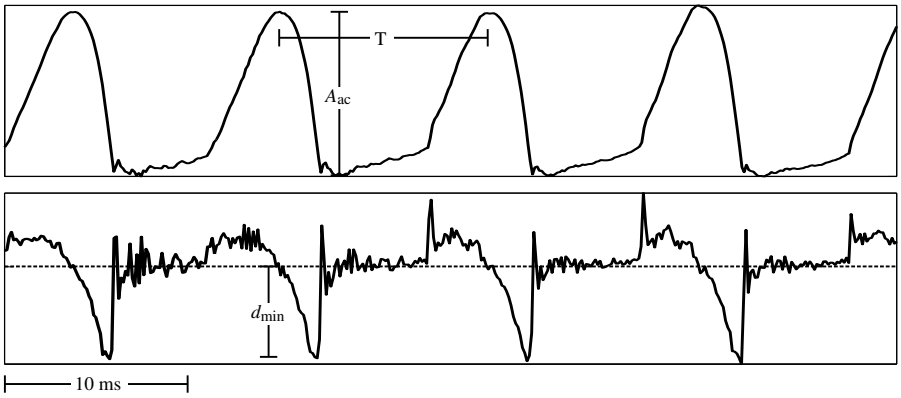

Figure 2: *A representative sample of the glottal flow, acquired from the material of the current study. The sample represents an [ɑ] vowel of a male speaker with no sentence or word stress. The measures required for the computation of NAQ and AQ are also illustrated.*

First, NAQ parameter values were inspected. Box-plots summarizing the values are shown in the left half of Fig. 3. The mean value of the NAQ parameters computed for both genders was 0.108, while the standard deviance (std.dev) equaled 0.024. The respective values for males and females were 0.102 (0.020) and 0.113 (0.025), i.e. the values were smaller for males. For males, in vowels without and with sentence stress, the values were 0.098 (0.020) and 0.109 (0.019). Two-way ANOVA with word and sentence stress as factors indicated that this difference was not statistically significant [$F(1,42) = 3.65,\ p = 0.063$]. For male vowels without and with word stress, the NAQ values were 0.091 (0.016) and 0.107 (0.019), again indicating higher values for stressed cases. This result was statistically significant [$F(1,42) = 4.49,\ p = 0.040$].

For females, the NAQ values were 0.109 (0.025) without sentence stress, and 0.120 (0.025) with it. The respective values without and with word stress were 0.105 (0.020) and 0.117 (0.028). Again, the NAQ values were higher for stressed cases, but the results were not statistically significant [$F(1,51) = 2.72,\ p = 0.105$ and $F(1,51) = 0.532,\ p = 0.469$, respectively].

Due to the NAQ values behaving contrary to the research hypothesis (see Section IV for details), AQ parameter values were also studied. The summary box-plots are shown in the right half of Fig. 3. The mean AQ values for males and females were 0.841 (0.134) and 0.567 (0.123), respectively, the considerably higher values for males stemming mainly from the F0 differences between males and females. For males, the values were 0.890 (0.129) without sentence stress and 0.745 (0.087) with it. This difference was found statistically significant [$F(1,42) = 16.0,\ p < 0.001$]. For males without and with word stress, the values were 0.869 (0.126) and 0.828 (0.138), respectively. This indicated lower AQ values in stressed cases for males. However, the result was not statistically significant [$F(1,42) = 0.858,\ p = 0.360$]. For females, the values were 0.605 (0.114) and 0.489 (0.105) without and with sentence stress, and 0.609 (0.103) and 0.545 (0.128) without and with word stress, respectively. Hence, the values were smaller in the stressed cases for females as well. In the case of sentence stress, the result was statistically significant [$F(1,51) = 14.0,\ p < 0.001$], but not so in word stress [$F(1,51) = 0.0762,\ p = 0.784$].

IV C

The NAQ values were higher in stressed than unstressed cases for both males and females, although for the majority of cases, not statistically significantly so. Still, this suggested that stress would be expressed using a breathier voice quality. This contradicted the original research hypothesis predicting that stress would be expressed with a pressed voice quality. Therefore, AQ values were inspected as well. As shown by the results, the AQ values behaved as expected, exhibiting smaller values in stressed vowels. ANOVA analyses showed this result to be statistically significant in the case of sentence stress, but not in the case of word stress. The authors suspect, however, that since the word stress appears to behave in a similar manner as sentence stress in the box plots, the lack of significance in ANOVA is only due to the small amount of material in the study.

There is plenty of evidence which shows that changing the glottal function from breathy towards pressed in sustained phonation is reflected by increase of F0 and decrease of both the absolute and the relative length of the glottal closing phase [e.g. 8]. In terms of NAQ and AQ, this implies that changing the phonation type from breathy to pressed in sustained phonation results in decrease of both of the parameters [2]. Interestingly, the current results on the glottal function in continuous speech showed a different trend according to which AQ decreased in stressed vowels in comparison to unstressed ones whereas the values of NAQ increased. Hence, the initial hypothesis that stress is expressed with a relatively more pressed voice quality could not be supported in this study. This unexpected, yet highly interesting result might be explained by the behaviour of sub-glottal pressure. In sustained phonation, namely, a speaker is able to produce a long vowel by using a steady-state value of the sub-glottal pressure which, in parallel with glottal adductory forces controlled by the cricothyroid and thyroarytenoid muscles, result in desired value of F0 and voice quality. In continuous

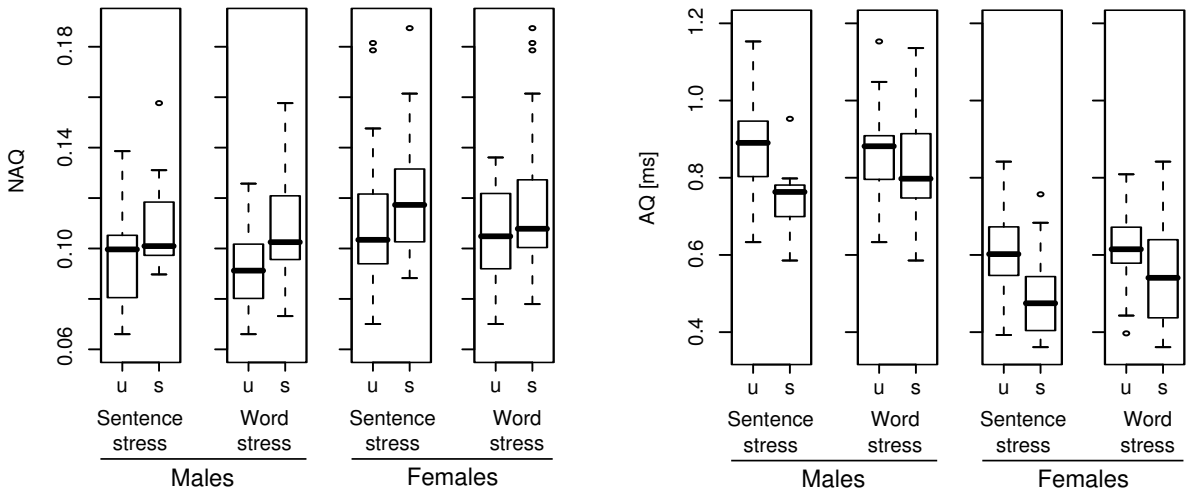

Figure 3: NAQ and AQ box-plots. In the labels, the letter 's' stands for a stressed case and 'u' for a unstressed one. The NAQ values are higher in stressed than in unstressed cases for both males and females. The AQ values show a large difference between males and females due to the intrinsically higher fundamental frequency of the females. Both in males and in females, AQ exhibits lower values in stressed vowels than in unstressed.

speech, however, the speaker has to adjust continuously the function of the vocal apparatus in order to produce different utterances including both voiced and unvoiced sounds. This implies, importantly, that a sustained sub-glottal pressure value is not possible to be held in the production of vowels in continuous speech. However, the speaker is able to change F0 by using the glottal adductory forces and this property can even be used to create fast changes in F0 as evidenced by F0 contours computed from continuous speech [e.g. 11]. With respect to the current results, the authors argue that the change from unstressed to stressed vowels caused a decreasing trend of AQ simply due to the increase of F0, that is, the shortening of the entire length of the glottal cycle. However, due to the lack of a sufficient level of sub-glottal pressure the shape of the glottal pulse became smoother when its cycle length was reduced. In other words, the speakers seemed to be unable to shorten the length of the glottal closing phase as effectively as they seem to be able to affect to the length of the entire glottal cycle. This, in turn, resulted in a breathier voice quality indicated by the higher NAQ value.

Remarkably, when the references given by Swerts and Veldhuis regarding the effect of F0 on OQ in inverse filtered speech are inspected more carefully, OQ appears to correlate positively with F0 or remain constant only when samples of continuous speech are used. This supports the findings of this study. The studies performed on sustained vowels or artificial voicing tasks, on the other hand, are more conflicting. These notions suggest, in the authors' opinion, that the results acquired by the study of sustained vowels should not be considered directly applicable to continuous speech.

Clearly, more work is required to gather comprehensive data regarding the voice source behaviour in natural speech. Such research should concentrate on recordings of continuous speech and should apply robust inverse filtering methods and reliable glottal flow parameterization methods. The authors believe that TKK Aparat, the freely available glottal flow examination software used in this study, provides tools suitable for further research on the topic.

A

This research was supported by the Emil Aaltonen foundation.

R

[1] M. Airas. HUT Aparat: An environment for voice inverse filtering and parameterization. *Logoped Phoniatr Vocol*, 2007. Submitted for review.

[2] P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *J Acoust Soc Am*, 112(2):701–710, August 2002.

[3] P. Alku, B. Story, and M. Airas. Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production. *Folia Phoniatr Logo*, 58:102–113, 2006.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.5.15) [computer program]. http://www.praat.org/, 2007. Visited 22-Feb-07.

[5] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans Sig Process*, 39:411–423, February 1991.

[6] M. A. Epstein. *Voice Quality and Prosody in English*. PhD thesis, University of California, Los Angeles, USA, 2002.

[7] C. Gobl. Voice source dynamics in connected speech. *STL-QPSR*, 29(1):123–159, 1988.

[8] E. B. Holmberg, R. E. Hillman, and J. S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *J Acoust Soc Am*, 84(2):511–1787, August 1988.

[9] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *J Computat Graphical Stat*, 5(3):299–314, 1996.

[10] J. Laver. *The phonetic description of voice quality*. Cambridge University Press, 1980.

[11] J. Pierrehumbert. Synthesizing intonation. *J Acoust Soc Am*, 70(4):985–995, 1981.

[12] D. D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *J Audio Eng Soc*, 37:419–444, 1989.

[13] M. Swerts and R. Veldhuis. The effect of speech melody on voice quality. *Speech Commun*, 33:297–303, 2001.