ELSEVIER

Speech Communication 33 (2001) 297–303

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

The effect of speech melody on voice quality

Marc Swerts [a,b,*], Raymond Veldhuis [a]

[a] *IPO, Center for User–System Interaction, P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands*
[b] *Universitaire Instelling Antwerpen, CNTS, Universiteitsplein 1, B-2610 Wilrijk, Belgium*

Abstract

This paper explores whether a speaker's voice quality, defined as the perceived timbre of someone's speech, changes as a function of variation in speech melody. Analyses are based on several productions of the vowel 'a', provided with different intonation patterns. It appears that in general fundamental frequency covaries with the 'strength relationship' between the first two harmonics (H1–H2). That relationship determines the voice quality to some extent, and is often claimed to reflect open quotient. However, correlating the H1–H2 measure to parameters of the LF-model reveals that both the open quotient and the skewness of the glottal pulse have an impact on the lower part of the harmonic spectrum. © 2001 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Artikel wird untersucht ob die Stimmqualität eines Sprechers, definiert als das wahrgenommene Timbre, sich als Funktion der Sprechmelodie verändert. Die Analysen basieren auf diversen Realisationen des Vokals 'a', die mit unterschiedlichen Intonationsmustern versehen sind. Die Grundfrequenz scheint im allgemeinen in einem 'Stärkeverhältnis' mit einem der ersten beiden harmonischen Obertöne (H1–H2) zu kovariieren. Dieses Verhältnis bestimmt in einem gewissen Ausmass die Stimmqualität und ist oft für den offenen Quotient verantwortlich gemacht worden. Die Korrelierung der H1–H2-Werte mit Parametern des LF-Modells zeigt, dass sowohl der offene Quotient als auch der Neigungsgrad des glottalen Pulses einen Einfluss auf den tieferen Teil des harmonischen Spektrums hat. © 2001 Elsevier Science B.V. All rights reserved.

Résumé

Cet article explore dans quelle mesure la qualité vocale d'un locuteur, définie comme le timbre perçu de la parole d'un individu, évolue en fonction de la variation mélodique. Les analyses sont fondées sur plusieurs productions de la voyelle 'a' correspondant à différents patrons intonatifs. I1 apparaît qu'en général la fréquence fondamentale covarie avec la 'relation de force' entre les deux premiers harmoniques (H1–H2). Cette relation détermine dans une certaine mesure la qualité vocale et elle est souvent avancée comme le reflet du quotient d'ouverture. Cependant, la corrélation de la mesure de H1–H2 avec les paramétres du modéle LF ŕevéle que le quotient d'ouverture et l'asymétrie de l'impulsion glottique ont tous deux un impact sur la partie basse du spectre harmonique © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Voice quality; Intonation; H1–H2; Glottal pulse

[*] Corresponding author. Tel.: +31-40-2475256; fax: +31-40-2431930.
E-mail address: m.g.j.s@tue.nl (M. Swerts).

0167-6393/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00061-3

1. Introduction

A speaker's voice quality, defined as the perceived timbre of someone's speech, is generally not constant. In the course of talking, it may change from e.g. dark and warm to more tense and sharp. Those attributes are believed to be related mainly to characteristics of the voice source, defined as the volume velocity airflow through the glottis during phonation (Gobl, 1988). Supposedly, various determinants may be held responsible for such alterations in timbre. The most dramatic changes are possibly due to attitudinal or emotional factors, for instance, when a speaker switches from a neutral to a sad or happy speaking mode. The goal of the current paper is to gain insight into more subtle causes. More specifically, it wants to explore potential correlations between variation in speech melody and voice quality.

A first theoretical reason to gain insight in such a relationship is that it could partly explain the so-called graph paper phenomenon: utterances of two different speakers may be spoken at approximately the same fundamental frequency (F_0), yet they may be distinct in that they are differently located in the speaker's pitch range, so that one utterance sounds as relatively high and the other as low. Listeners might be able to locate an utterance in a speaker's speech range on the basis of voice quality.

A more technology-driven motivation for the research is that a better understanding of possible interactions between pitch and voice quality may lead to a more natural output of (glottal-excited) speech synthesizers which are able to generate different voice qualities. Most synthesizers make use of explicit intonation models that specify how pitch should change in utterances according to such variables as stress and accent, sentence structure, sentence length and discourse structure. If there is a simple dependency between pitch and source characteristics, then it might be worth trying to have changes in voice quality be triggered by melodic features. Ideally, in the case of diphone-based synthesis, if such manipulations considerably improve the synthesis quality, one would like to be able to generate voice changes on the basis of only a limited set of diphones, because it would be too costly to record different sets that match different pitch ranges.

Previous work shows that pitch and voice characteristics may indeed be correlated. For instance, natural speech is characterized by deviations in strict periodicity (also known as jitter) and other unstabilities, such as shimmer, but the amount of such irregularities is stronger when subglottal pressure is falling and fundamental frequency is low, and vice versa (Baken, 1987).

Also, pitch differences may influence the shape of the glottal pulse. In that respect, the current paper focuses on one aspect of the voice source, i.e., its open quotient (OQ), which refers to the relative time that the glottis is open within a pitch period. Various comparisons between male and female speakers all reveal that in general the former exhibit smaller OQ values than the latter (e.g., Sluijter, 1995; Baken, 1987). However, within-speaker studies are not very conclusive. Some researchers claim that F_0 and OQ correlate positively. Summarizing a series of studies that used highspeed-filming and photoglottography, Baken (1987) makes the generalization that OQ increases with higher F_0 in a modal speech register. Pierrehumbert (1989) found that H tones have a greater OQ than L tones when uttered at a comparable voice level. Oliveira (1996) measured F_0 values in the middle portion of 3276 vowels and found that these correlated moderately with OQ. Koreman (1995) also reports an increase in OQ for higher F_0's. Others, however, claim that OQ remains remarkably constant under changes in F_0 (see references in Klatt and Klatt, 1990) and some present evidence for inverse dependencies, higher F_0 being reflected in smaller OQ values (Karlsson and Liljencrants, 1996; Cleveland and Sundberg, 1983).

The results from different studies may be at variance with each other, partly because of different estimation procedures with different underlying assumptions, which makes it difficult to compare outcomes. The discrepancies in the results could also be due to the fact that in one type of data a clear correlation between OQ and F_0 is 'disturbed' by influences of other properties of the glottal pulse or other prosodic factors. For instance, it is known that voice level has an opposite effect on OQ in that a louder voice results in a smaller OQ (Baken, 1987; Pierrehumbert, 1989).

The present study chooses to constrain the analysis to relatively short utterances, using the vowel ‘a’ for which the estimation of OQ is comparatively easy. Also, both a simple and more sophisticated method are introduced to automatically determine glottal pulse characteristics from running speech. The focus on OQ (and H1–H2, see below) in this study is largely inspired by the fact that many previous studies have also concentrated on this parameter. This does not imply, however, that it is the most important determinant of voice quality, since other spectral attributes due to the overall spectral tilt of the voice source are likely to be more salient.

2. Procedure

The potential effect of pitch range on voice source characteristics was investigated in two production experiments. Seven male speakers were instructed to produce a series of increasingly more complex utterances that systematically varied with respect to intonation. In one condition, they were asked to produce monotonous vowels (‘a’), spoken at two discrete pitch heights (low and high at an approximate distance of 1 octave). Another set involved the production of vowels with different intonation patterns, i.e., four types of continuous variation between a low (l) and high (h) pitch target, giving lh, hl, lhl and hlh, respectively. (All speakers performed the two production tasks, except HP and JP, who only participated in the first and second test, respectively.) All the utterances were digitized with a 16 kHz sampling frequency. The vowel ‘a’ was chosen for further analysis, because the determination of glottal-pulse characteristics in this vowel is relatively easy because of a high first formant which has a neglectable effect on the lower part of the spectrum (Ní Chasaide and Gobl, 1997).

In one tradition of research, glottal pulse characteristics are derived from concise analyses of spectra. More specifically, OQ is estimated from the relationship between the first two harmonics with a relatively stronger H1 reflecting a bigger open quotient (Klatt and Klatt, 1990; Fant, 1997; Sluijter, 1995; Holmberg et al., 1995). Changes in this relationship result in the perception of distinct voice qualities: a relatively strong H1 leads to a more breathy voice, whereas a strong H2 correlates with tense or creaky voice (Ní Chasaide and Gobl, 1997). However, more recent studies question the claim that OQ is the sole factor to determine the H1–H2 relationship; there is growing evidence that also other characteristics of the voice source have an impact on the lower part of the harmonic spectrum.

Consequently, the present research has a double goal. It will first try to find out whether H1–H2 covaries with pitch. Next, the study explores whether the H1–H2 measure indeed reflects OQ within the constraints of the LF-model.

3. Measurements

We adopt the well-known source-filter model for speech production (Fant, 1960) for our measurements. In a simplified form this model consists of a source producing a signal representing the time derivative of the air flow through the glottis which is input to a filter modelling the transfer function of the vocal tract. We restrict ourselves to voiced speech and choose the LF-model of the glottal-pulse time derivative (Fant et al., 1985) to parametrize one cycle of the source signal. The model parameters that we adopt are OQ, the open quotient, RK, the skewness parameter, and RA, the relative duration of the return phase. (OQ is similar to the ‘reduced’ OQ_i in (Fant, 1997), which excludes the tail end of the glottal flow.) We also want to measure the strengths H1 and H2, expressed in decibels, of the first two harmonics of the source signal. All the quantities OQ, RK, RA, H1 and H2 need to be estimated from running speech.

Measurement of H1 and H2 can be done directly by Fourier transforming consecutive segments of speech, in which case the influence of formants has to be corrected for (Fant, 1960; Sluijter, 1995). However, since we have to perform an inverse-filtering operation for the estimation of the parameters OQ, RK and RA, we estimate the

H1 and H2 parameters from Fourier-transformed segments of the inverse-filtered signal. The segment length is 40 ms and the segments have a 20-ms overlap. The inverse filter is derived from the linear prediction coefficients by removing all poles below 200 Hz. This procedure is similar to the one described in (Childers and Lee, 1991). A procedure to precisely estimate F_0 is also taken from this reference. Each inverse-filtered segment is windowed by a Hanning window, after which its Fourier transform is computed for all integer multiples of F_0 below half the sampling frequency f_s. The spectral lines $U_l,\ l \geqslant 1,\ lF_0 < f_s/2$, which are also required for the estimation of the LF parameters, are the squared moduli of the segment's Fourier transform. We use a power-normalized version of U_l, such that $\sum_l U_l = 1$. The values of H1 and H2 follow by expressing U_1 and U_2 in decibels.

We now turn to the description of the estimation of the LF parameters OQ, RK and RA. Let the time derivative of the glottal air flow be denoted by $\dot{g}(t;\mathrm{OQ},\mathrm{RK},\mathrm{RA})$. This is a periodic function, with period $T_0 = 1/F_0$. A common procedure for estimating OQ, RK and RA is by matching a prototype waveform $\dot{g}(t;\mathrm{OQ},\mathrm{RK},\mathrm{RA})$ to one period of the inverse filtered signal (e.g., Childers and Lee, 1991), and tune the parameters of the prototype wave form until the match is optimal, e.g., in a quadratic sense. This procedure requires a speech recording without phase distortion and often additional manual fine tuning. Instead, we use a magnitude-spectral estimation method that can be used on running speech and that is not sensitive to phase errors. The prototype waveform is periodic with period T_0. This means that it can be described by a Fourier series expansion. We refer to the squared magnitude

$$\dot{G}_l(\mathrm{OQ},\mathrm{RK},\mathrm{RA}) = \left| \frac{1}{T_0} \int_0^{T_0} \dot{g}(t;\mathrm{OQ},\mathrm{RK},\mathrm{RA}) \exp\left(-\mathrm{j}2\pi \frac{t}{T_0} \right) \mathrm{d}t \right|^2, \tag{1}$$

with $l \geqslant 1,\ lF_0 < f_s/2$, as the glottal-pulse spectral lines. We use a power-normalized version of $\dot{G}_l(\mathrm{OQ},\mathrm{RK},\mathrm{RA})$, such that $\sum_l \dot{G}_l(\mathrm{OQ},\mathrm{RK},\mathrm{RA}) = 1$. The parameters OQ, RK and RA are those which minimize the spectral distance measure to the spectral lines $U_l,\ l \geqslant 1,\ lF_0 < f_s/2$,

$$d(U,\dot{G}) = \sum_l \dot{G}_l(\mathrm{OQ},\ \mathrm{RK},\mathrm{RA}) \times \log\left(\frac{\dot{G}_l(\mathrm{OQ},\mathrm{RK},\mathrm{RA})}{U_l} \right). \tag{2}$$

This measure is also known as the Kullback–Leibler (Kullback and Leibler, 1951) distance measure for probability distributions. Frequency domain approaches to determine voice source characteristics, including estimation of spectral tilt, have also been proposed in (Fant et al., 1995; Hanson, 1997; Stevens, 1994; Stevens and Hanson, 1994).

4. Results

The results are presented in the following two subsections: Section 4.1 discusses correlations of H1–H2 with F_0 in different vowel productions; Sections 4.2 explores to what extent H1–H2 can be related to parameters of the LF-model.

4.1. Relation of H1–H2 to F_0

Table 1 gives averages of F_0 and H1–H2 for monotonous low and high vowels of different speakers. It indicates that there is considerable between-speaker variation in H1–H2 values. Within-speaker comparisons, however, reveal that, with one exception, low vowels always have smaller H1–H2 values than high ones. ANOVAs performed on the data of each speaker separately

Table 1
Mean F_0 (in Hz) and H1–H2 (in dB) for /a/'s spoken by different speakers at a low and a high pitch register

	Low		High	
SP	F_0	H1–H2	F_0	H1–H2
DH	99	1.17	198	9.41
EK	93	1.94	122	4.08
GV	129	8.19	257	11.40
HP	98	13.20	193	18.35
MS	90	4.59	179	9.66
RK	101	7.66	197	9.74
RH	115	7.08	228	5.76

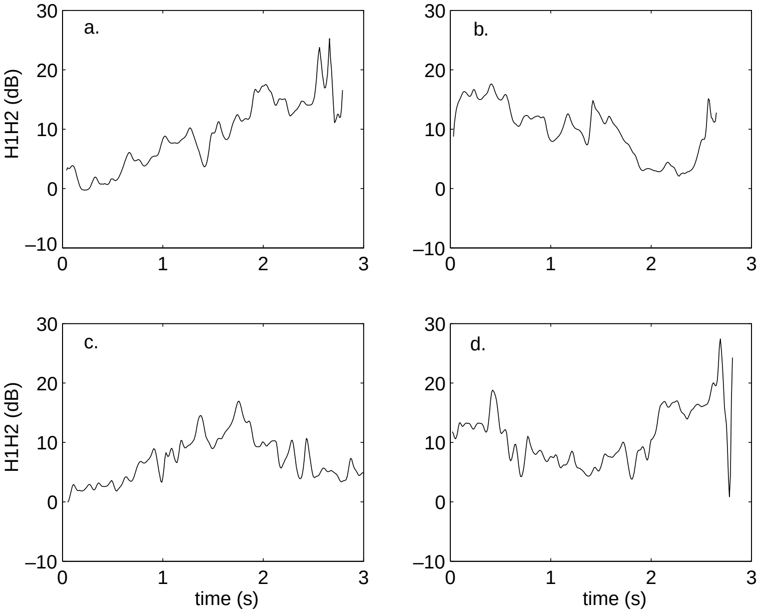

Fig. 1. H1–H2 patterns in different intonation contours of speaker MS: (a) lh; (b) hl; (c) lhl; (d) hlh.

always give significant differences in H1–H2 ($p < 0.001$), except for EK and GV though the trend in their data is the same. Speaker RH appears to show a significant opposite effect.

Fig. 1 gives the H1–H2 patterns of speaker MS for a set of more complex intonation contours. It reveals that the spectral measure covaries continuously with pitch in these data. The Pearson correlation coefficients between melodic and spectral data for the different speakers are given in Table 2, and summarized in Table 3. From these tables, it appears that there is indeed an overall tendency for H1–H2 to covary with F_0, but that there is quite some speaker variation: for most cases, there is a significant positive correlation (18), though not always very high; a minority exhibits no dependency at all between the two variables (4), and a few show a significant opposite relationship (6).

Table 2
Correlations between F_0 and H1–H2 in different intonation patterns for different speakers

	Intonation pattern			
SP	hl	hlh	lh	lhl
DH	0.12	0.60[a]	0.47[a]	0.02
EK	0.65[a]	0.50[a]	0.67[a]	0.70[a]
GV	−0.36[a]	−0.56[a]	−0.45[a]	−0.16[a]
JP	0.82[a]	0.14[b]	−0.57[a]	0.65[a]
MS	0.81[a]	0.78[a]	0.92[a]	0.70[a]
RK	0.16	0.73[a]	0.42[a]	0.26[b]
RH	−0.22[b]	0.18[c]	0.24[b]	0.13

[a] $p < 0.001$.
[b] $p < 0.01$.
[c] $p < 0.05$.

4.2. Relation of H1–H2 to LF parameters

In (Holmberg et al., 1988), different procedures to estimate glottal pulses were compared and it was found that OQ correlates reasonably well with a relatively strong first harmonic. However, Ní Chasaide and Gobl (1997) and Fant et al. (1995) claim that the skewness of the pulse, indicated with the RK parameter of the LF-model, also has a strong effect on the lower part of the harmonic

Table 3
Number of cases that F_0 and H1–H2 are negatively correlated, not correlated, or positively correlated

F_0 and H1–H2 are	Cases
Negatively correlated	6
Not correlated	4
Positively correlated	18

Table 4
Number of cases of H1–H2 patterns that correlate positively, do not correlate, or correlate negatively with the LF-parameters OQ and RK

	With	
H1–H2	OQ	RK
Correlates positively	17	25
Does not correlate	7	2
Correlates negatively	4	1

spectrum. Table 4 reveals that there is indeed no one-to-one relationship of H1–H2 with OQ: although the two measures correlate positively in the majority of the cases, there is sometimes no such dependency at all, or even an opposite one. From the correlation data, it appears that there is a comparatively bigger tendency for RK to correlate positively with H1–H2, though there are also cases where the two measures do not correlate or correlate negatively.

Therefore, in order to explore this somewhat further, we have analysed the behaviour of H1–H2 as a function of the LF parameters OQ and RK, for a fixed value 0.05 of RA. Fig. 2 shows the contours of H1–H2 at the values 0, 5, 10, 15, 20, 30, 45 dB as a function of OQ and RK. The figure shows that the values of H1–H2 in the range plotted in Fig. 1 can be attained for all combinations of OQ and RK along a contour, and that along such a contour, e.g. the 20-dB contour, OQ and RK go through a fairly wide range of values. Moreover, for a common value of $\mathrm{RK} = 0.5$, we can see that H1–H2 is not monotonically increasing with OQ.

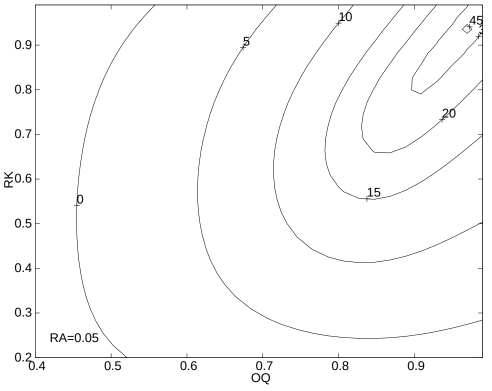

Fig. 2. Contour plot of H1–H2 as a function of OQ and RK.

This is in line with the outcome of recent work by Doval and d'Alessandro (1997) who argue that the H1–H2 ratio is the composite result of both OQ and RK, which is expressed in their equation

$$\mathrm{H1{-}H2} = 12\left(\frac{\mathrm{OQ}}{0.7}\right)^2\left(1-\left(1-\frac{\mathrm{RK}}{0.7}\right)^2\right)-6. \tag{3}$$

Fant (see Fant et al., 1995; Fant, 1997) even argues that all LF parameters have an influence on H1–H2. With normal male voice as a reference with RK = 31%, RG = 118% and RA = 2.4%, a shift in H1–H2 of +1 dB can be reached by either a 1.25% increase of RA to 3.65% or by an increase of RK by 4.5–35.5% or by a decrease of RG by 10–108%. This makes the simple use of H1–H2 as a measure of OQ even more questionable.

5. Conclusions and perspectives

The research reported in this paper reveals that there is some evidence for a dependency of F_0 on voice quality. Usually, changes in speech melody are reflected in the H1–H2 relationship, but it is clear that there is profound speaker variation, in that some exhibit strong relationships, some none, and some opposite dependencies. A new method was introduced to estimate glottal pulse characteristics from running speech. There does not appear to be a one-to-one relationship between H1–H2 and OQ as defined in the LF-model.

There are a few logical follow-up studies to the one presented here. First, since the results here are based on very limited speech data, it remains to be seen to what extent they generalize to 'real' utterances; the picture may be more complex in such data, e.g., because of effects such as the one reported by Oliveira (1996) that syllable duration may influence voice characteristics. There may also be interference of other prosodic features, such as loudness or vocal effort. Fant et al. (1995), for instance, have shown that the H1–H2 increases by about 6 dB when the overall intensity is lowered by 10 dB below normal level. As a matter of fact, this study does not exclude that these variables may even correlate more strongly with the H1–H2 measure than does F_0. It is possible that variation in intensity is responsible for the observed inter-speaker variations in the data presented here. Second, it needs to be explored (a) whether the observed differences in OQ are above a perceptual threshold and (b) whether manipulations in the voice source may lead to a different perception of pitch range. Finally, the implementation in glottal-excited speech synthesis of interactions between pitch and voice quality need to be evaluated to see whether it improves the naturalness of the spoken output.

Acknowledgements

Marc Swerts is also affiliated with the Flemish Fund for Scientific Research (FWO, Flanders). Thanks to Gunnar Fant, Jan van Santen and Jacques Terken for comments on an earlier version of this paper.

References

Baken, R.J., 1987. Clinical Measurements of Speech and Voice. Taylor and Fancis, London.

Childers, D.G., Lee, C.K., 1991. Voice quality factors, analysis, synthesis and perception. J. Acoust. Soc. Am. 90, 2394–2410.

Cleveland, T., Sundberg, J., 1983. Acoustic analysis of three male voices of different quality. STL-QPSR 4, 24–38.

Doval, B., d'Alessandro, C., 1997. Spectral correlates of glottal waveform models, an analytic study. In: Proc. ICASSP'97, pp. 1295–1298.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague.

Fant, G., 1997. The voice source in connected speech. Speech Communication 22, 125–139.

Fant, G., Liljencrants, J., Lin, Q., 1985. A four-parameter model of glottal flow. STL-QPSR 4/85, 1–13.

Fant, G., Liljencrants, J., Karlsson, I., Båvegård, M., 1995. Time and frequency domain aspects of voice source modelling. ESPRIT BR Speechmaps (6975). Deliverable 27 WP 1.3.

Gobl, C., 1988. Voice source dynamics in connected speech. STL-QPSR 1, 123–159.

Hanson, M., 1997. Glottal characteristics of female speakers, acoustic correlates. J. Acoust. Soc. Am. 101 (1), 466–481.

Holmberg, E., Hillman, R.E., Perkell, J.S., 1988. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. J. Acoust. Soc. Am. 84 (2), 511–529.

Holmberg, E., Hillman, R.E., Perkell, J.S., 1995. Measures of the glottal airflow waveform, EGG, and acoustic spectral slope for female voice. In: Proc. ICPhS'95, Stockholm, pp. 178–181.

Karlsson, I., Liljencrants, J., 1996. Diverse voice qualities, models and data. TMH-QPSR 2/1996, 143–146.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis and perception of voice quality variations among female and male speakers. J. Acoust. Soc. Am. 87 (2), 820–856.

Koreman, J., 1995. The effects of stress and F_0 on the voice source. Phonus 1, University of Saarland, pp. 105–120.

Kullback, S., Leibler, R., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–87.

Ní Chasaide, A., Gobl, C., 1997. Voice source variation. In: Hardcastle, W.J., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Oxford.

Oliveira, L.C., 1996. Text-to-speech synthesis with dynamic control of source parameters. In: van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer, New York, pp. 27–39.

Pierrehumbert, J., 1989. A preliminary study of the consequences of the intonation for the voice source. STL-QPSR 4, 23–36.

Sluijter, A., 1995. Phonetic Correlates of Stress and Accent. Holland Ac. Graphics, The Hague.

Stevens, K.N., 1994. Prosodic influences on glottal waveform, preliminary data. In: International Symposium on Prosody, Yokohama, September 1994, pp. 53–64.

Stevens, K.N., Hanson, M., 1994. Classification of glottal vibration from acoustic measurements. In: Fujimura, O., Hirano, M. (Eds.), Vocal Fold Physiology. Singular Publishing Group, San Diego, CA, pp. 147–170.