ON THE RELATION BETWEEN VOICE SOURCE PARAMETERS AND PROSODIC FEATURES IN CONNECTED SPEECH

H. Strik and L. Boves

Dept. of Language and Speech, University of Nijmegen,

Nijmegen, the Netherlands

H. Strik

Dept. of Language and Speech

Erasmusplein 1

P.O. Box 9103

6500 HD Nijmegen

The Netherlands

Unusual symbols: None

Number of pages: 21 + 5 (for the 5 figures)

Number of tables: 3

Number of figures: 5

Proposed running head: voice source parameters and prosody

Keywords: inverse filtering; LF-model; voice source

Abstract

The behaviour of the voice source characteristics in connected speech was studied. Voice source parameters were obtained by automatic inverse filtering, followed by automatic fitting of a glottal waveform model to the data. Consistent relations between voice source parameters and prosodic features were observed.

Zusammenfassung

Das Verhalten der Stimm-Quellcharakteristik in kontinuierlicher Sprache wurde untersucht. Stimm-Quellparameter wurden durch automatisches inverses Filtern ermittelt. Anschließend wurden die Daten über ein automatisches Fehlerminimierungsverfahren in ein Modell der glottalen Wellenform eingepaßt. Es wurden konsistente Zusammenhänge zwischen Stimmquellecharakteristiken und prosodischen Merkmalen festgestellt.

Résumé

Le comportement de la source vocale dans la parole continue a été examiné. Des paramètres de la source vocale ont été obtenus à l’aide de filtrage automatique, suivi par un approchement automatique d’un modèle d’onde de débit glottique aux observations. Des relations consistentes entre les paramètres de la source vocale et des traits prosodiques ont été trouvées.

1. Introduction

Modern text-to-speech systems produce speech that is intelligible, but not quite natural. This lack of naturalness is at least in part due to the absence of adequate prosody control. Prosody does not only include fundamental frequency (F_0) and duration, but it also affects more subtle aspects of the speech signal that can be subsumed under the cover term ’voice quality’. Completely satisfactory prosody will therefore require the use of adequate voice source control rules. This opinion is reflected by the fact that many rule based text-to-speech systems are now being updated, in order to replace a static voice source with a source that can be dynamically controlled. A number of different voice source models have been proposed, each with its own specific advantages and drawbacks. However, it is not our intention to compare different models. Even the most sophisticated voice source model will not improve speech quality if it is not being controlled by the right rules. These rules, on the other hand, cannot be derived without a large amount of data on the behaviour of the voice source in natural speech, or more specifically, of the behaviour of those characteristics of the source that can be mapped onto the model parameters. Fortunately, most modern source models share a large number of parameters, so that most of the results obtained with one model should be easy to generalise to other models.

In a text-to-speech synthesis framework all relevant properties of the voice source can be described in terms of the glottal volume flow signal, and its time derivative. Those glottal flow signals can be approximated, starting from the acoustic speech signal, via inverse filtering. Model parameters can then be estimated by fitting the model waveform to the inverse filtered waveforms. Inverse filtering and model fitting could in principle be done interactively. However, interactive measurements would take an inordinate amount of time, because rule development requires one to process large quantities of speech. Moreover, interactive measurements are difficult to reproduce. For these reasons a procedure was developed to derive the voice source parameters automatically. That procedure is explained in Section 2.

Up to now, most research on voice source characteristics has dealt with sustained vowels, produced in different ways. For sustained vowels, recorded with a high SNR, automatic extraction of the voice parameters is fairly easy. But it is difficult to extrapolate from data acquired from isolated speech sounds to rules for connected speech. Therefore, our aim is to study the behaviour of the voice source in connected, preferably spontaneous speech. And in addition to steady state portions of vowels we also want to extract source parameters for voiced consonants, as well as for voiced/unvoiced (V/UV) and UV/V transitions. The results of our work are presented in section 3.

The strategy that we adopt to find relations between several voice source parameters on the one hand, and between voice source parameters and prosody on the other, is the following: first we will derive general relations by averaging over all data; after that we will look for local deviations from these general relations. Special attention is given to the relation between voice source parameters and prosodic features like F_0, intensity (Int), and voice quality.

2. Method and material

2.1. Speech material

To study voice source characteristics data were collected for four male subjects. For all subjects recordings were made of the speech signal, electroglottogram (EGG), subglottal (P_{sub}) and oral (P_{or}) pressure, lung volume, and electromyographic activity of some laryngeal muscles (mostly crycothyroid, vocalis, and sternohyoid). The signals were stored on wide band FM-tape. All recordings were made at the ENT-clinic of the University Hospital "Sint Radboud" in Nijmegen, in a room in which no special acoustic precautions were made. For the current article only data of one subject were used (Strik and Boves, in press). Near the end of a recording session he was asked to produce an utterance spontaneously. His response was: "*Ik heb het idee dat mijn keel wordt afgeknepen door die band*" ("I have the feeling that my throat is being pinched off by that band"). He then repeated this utterance 29 times. The 30 utterances had an average length of 2.3 seconds. For this paper inverse filter results of the first four utterances were analyzed.

2.2. Inverse filtering

The speech signal was transduced by a condensor microphone (B&K 4134) placed about 10 cm in front of the mouth, pre-amplified at the microphone (B&K 1619), and amplified by a measuring amplifier (B&K 2607) using the built-in 22.5 Hz high-pass filter to suppress low frequency noise. The speech signal was A/D converted off-line at a 10 kHz sampling rate, and processed with a phase correction filter in order to undo the low frequency phase distortion caused by the high-pass filter.

Closed glottis interval covariance LPC analysis was used to estimate the parameters of the inverse filter. In de Veth, Cranen, Strik & Boves (1990) it was shown that this technique for estimating the inverse filter is as powerful as more sophisticated techniques, like Robust ARMA analysis. The moment of glottal closure was determined from the EGG, and it is used to position the analysis window. Inverse filtering yields an estimate of the differentiated glottal volume flow (dU_g); integration of dU_g gives the flow signal (U_g).

Closed glottis interval inverse filtering is a complex process; its implementation requires several choices to be made to fix parameters. The most important parameters are the length and exact position of the analysis window, the pre-emphasis factor, and the order of the analysis. In general, there seems to be no combination of these parameters that is optimal for each individual pitch period in a normal speech utterance. However, a 12th order LPC analysis with a pre-emphasis factor of 0.95 worked satisfactorily for almost all pitch periods.

Thus window position and window length were left as the parameters to be varied. Instead of trying to formulate criteria that would allow one to determine the unique optimal combination of window length and position for each period, we decided to try a large number of combinations and to leave it to a simple statistical procedure to make the final selection (see section 2.4.).

2.3. Voice source parameters

For automatic fitting of a glottal waveform model to inverse filtered flow signals we used a special software package (Jansen, Cranen, and Boves, 1991). The fit is done pitch synchronously. The periods are defined by the minima in dU_g, because these time points can be located most reliably. This software package allows one to use different glottal waveform models, different definitions of the error function, and different optimization routines. The choices made for this study are given below.

The so called LF-model was used, because it seems useful for synthesis, and because it has already been studied in great detail (see e.g. Fant, Liljencrants, and Lin, 1985). The model and its parameters are presented in Fig. 1. The relations between the dimensionless wave shape parameters of the LF-model and the spectrum are well-known (see e.g. Fant and Lin, 1988): R_g has a small influence on the amplitude relations of the lower harmonics, R_k influences the spectral balance, and R_a influences the spectral tilt.

- insert Figure 1 about here -

The error function describes the difference between the model and the measured signals. It can be defined in the time domain, the frequency domain, or in both domains simultaneously. For this study the error function is based on the time signals of flow and flow derivative. In a pilot experiment it was found that this error definition minimises the number of discontinuities in the signals fitted to U_g and dU_g. For a given pitch period the error function is calculated by subtracting the modelled signals from the measured signals. The best fitting model waveform is found by adapting the model parameters in such a way that the energy in the error function is minimised.

An adaptive nonlinear least-squares optimisation algorithm called NL2SNO (Dennis, Gay, and Welsch, 1981) was used to find the best fit. The algorithm returns the (minimised) error energy, and the parameters for which that optimum is found. If the minimal error is smaller than a pre-defined threshold, then the fit is said to be good. But if the minimal error remains above the threshold, then all LF-parameters for that pitch period are set to -1 to indicate that the fit is not successful.

2.4. Averaging the results

Inverse filtering was done for all 25 combinations of 5 window lengths (33, 34, 35, 36, and 37 samples) and 5 window shifts (-2, -1, 0, 1, and 2 samples relative to the moment of glottal closure). The LF-parameters were obtained for all 25 resulting inverse filter signals, by fitting the LF-model to the data. For each pitch period median values for all parameters in the LF-model were calculated.

The median value of a parameter for a pitch period can become negative (-1), if at least 13 of the 25 values of that parameter are equal to -1. This occurs if in more than half of the cases the fit was not successful. The data of all pitch periods in which the median value of one of the LF-parameters is equal to -1 were discarded. In total 128 periods were discarded, and the data of 613 pitch periods were used for further analysis. The disadvantage of using such a conservative criterion is that a lot of data have to be discarded, but the advantage is that the risk of errors in the final data is reduced. We are convinced that keeping more of the data for the consonants and onsets/offsets would not have changed our results and conclusions.

3. Results

- insert Figure 2 about here -

The audio signal, automatically calculated inverse filter results, and automatically obtained fits for five consecutive pitch periods of a vowel /e/ are given in Figure 2. The differentiated flow signals often contain a pronounced ripple. It is clear from this figure that attempts to measure the LF-parameters from the raw dU_g or U_g signals would result in noisy estimates. For instance, the maximum of dU_g (E_i) and the place of this maximum (T_i) are to a large extent determined by the ripple. By fitting a LF-model to the data the measurements are made more robust. The fit procedure is almost always able to find a combination of LF-parameters that generates a model signal that closely resembles the measured flow signal.

- insert Figure 3 about here -

In Fig. 3 the median values of the most relevant parameters are given for a voiced interval of one of the utterances. For some pitch periods the median values of all LF-parameters are -1, indicating that for the majority of the 25 combinations the fit was not successful for these periods. There are two causes that could hinder a good fit. Sometimes the estimate of the vocal tract transfer function was not correct, in which case inverse filtering did not yield a flow signal that resembles a LF-pulse even remotely. There were also cases, however, in which inverse filtering produced a reasonable estimate of dU_g, but where the optimization routine did not converge. Not surprisingly, estimation problems occurred more often in voiced consonants, and during voice onset and offset (the first and last periods of a voiced segment) than during the steady parts of vowels.

Furthermore, it was observed that estimates of the parameters of the first part of the LF-model (the exponentially growing sine wave, i.e. T_p, T_e, E_e) varied less than those of the return phase (i.e. T_a). Partly this is due to the fact that the duration of the first part is longer than the duration of the return phase. But another cause is that the return phase often is not smooth and contains a ripple (see Fig. 2). This pronounced ripple often affects the automatic fitting process for the return phase. In many cases a reasonable fit could be reached for the first part of the LF-model, but not for the return phase. The result is that the median value of T_a often is -1, while the other parameters are not (see Fig. 3).

For the moment we do not know whether the failure of the fit procedure to converge to an acceptably small error is due to computational problems or to the failure of the LF-model to approximate all glottal flow pulse forms that occur in real speech.

3.1. General behaviour

Typical behaviour of the LF-parameters can be observed in Fig. 3. During transitions from vowel to consonant T_0, T_a, and T_n generally increase, while transglottal pressure (P_{tr}), U_o, E_e, and Int decrease. The consistent reciprocal relation between the parameters in these two sets is reflected in the correlation coefficients (see Table I), which are all negative and highly significant ($p<0.0001$). For these and all following correlation coefficients the level of significance for a two-tailed test was calculated (Ferguson, 1987). The correlation coefficient between two sets of 613 samples is said to be significant at the 0.01% level ($p<0.0001$) if its absolute value is larger than 0.16.

- insert Table I about here -

The rationale behind this very general behaviour is most probably the following. For vowels the impedance of the glottis is much higher than the impedance of the vocal tract, and thus P_{tr} is almost equal to P_{sub}. For consonants there is a constriction in the vocal tract, causing a pressure build-up above the glottis and a drop in P_{tr}. In order to keep vibration going (with a lowered P_{tr}) during these voiced consonants, some adjustments must be made: the vocal folds are slackened and abducted, and the consequence is that T_a and T_n are raised. Lowering of P_{tr} and slackening of the folds will lower F_0, and thus raise T_0. Although the folds are slackened, the decrease in P_{tr} is such that the amplitude of vibration of the folds decreases, and with it the modulation of the flow (U_o), and eventually E_e and Int.

The observed reciprocal relation provides a natural way for dividing the LF-parameters into two sets. The first set consists of T_i, T_p, T_e, T_n, T_a, and T_0, and will be referred to as the ’time parameters’, while the second set (P_{tr}, U_o, E_e, Int) will be referred to as the ’amplitude related parameters’. Relations within the first set are described in section 3.2, and relations within the second set in section 3.3. The relations between F_0 and other parameters can be derived directly from the relations of these parameters with T_0. Therefore, they are not treated separately, but are part of section 3.2. The behaviour of the wave shape parameters R_g, R_k, and R_a is described in section 3.4.

3.2. Time parameters

It was already mentioned that during transitions from vowels to consonants T_0, T_a, and T_n are generally raised (see Fig. 3). The following question than emerges: How does a change in T_0 affect the time parameters, or, in other words, how does the shape of the pulse change with F_0? In this section we try to answer this question by looking at the relations between T_0 and the other time parameters.

The five time parameters T_i, T_p, T_e, T_a, and T_n were first plotted as a function of T_0 on a double logarithmic scale, and the best linear fits were calculated. The resulting lines are of the form:

$$\log T_x = \log a_0 + a_1 . \log T_0 \Leftrightarrow T_x = a_0 . T_0^{a_1}, \ x \in \{i, p, e, a, n\}$$

The regression lines for T_i, T_p, T_e, T_a, and T_n are shown in Fig. 4. All correlations between the logarithm of the five time parameters and the logarithm of T_0 are positive and highly significant ($p<0.0001$). So, on the average, all time parameters increase with increasing T_0, and the glottal pulse is stretched. However, this stretching is not distributed uniformly over the entire period.

- insert Figure 4 about here -

If a time parameter changes linearly with T_0, then its regression line in Fig. 4 should have a slope of 1. In that case it would run parallel to the reference line for T_0 that is also given in Fig. 4 ($T_0 = 1.T_0^1$), which obviously has a slope of 1. This is the case for T_e, so generally the duration of the first part of the LF-pulse changes linearly with T_0. However, the increase in T_i and T_p is less than linear, and the increase in T_a and T_n ($T_n = T_e - T_p$) is more than linear (see Fig. 4). The ordering of the time parameters with ascending power is T_i, T_p, T_e, T_n, T_a. It seems as if the amount of stretching increases when going towards the end of the LF-pulse. With regard to the shape of the LF-pulse, the consequence is that the skewing decreases more than linearly with T_0.

3.3. Amplitude related parameters

A constantly high covariance between the amplitude related parameters was found for all data (see Table II and Fig. 5). At first sight the high covariance of these parameters does not seem surprising, as an increase in P_{tr} alone (everything else being equal) would increase the amplitude of vibration of the vocal folds, and therefore lead to an increase in U_o and E_e. Increasing U_o and E_e by roughly the same amount would lift the spectrum (see Fant and Lin, 1988), and thus increase Int. However, our data form a mix of voiced consonants, stressed and unstressed vowels. Thus one might expect large variations, both in the glottis and in the vocal tract. For instance, for voiced consonants T_a and T_n are generally higher than for vowels (see section 3.2). A change in T_a has little effect on Int, but an increase in T_n (i.e. less skewing) combined with a decrease in U_o would lead to a decrease in E_e that is relatively larger than the decrease in U_o. Given the large variation in articulatory gestures, it is surprising that the covariance between P_{tr}, U_o, E_e, and Int is invariably high.

- insert Figure 5 about here -

- insert Table II about here -

Regression lines were calculated for the amplitude related parameters. The procedure used was analogous to the procedure used for the time parameters, as described in section 3.2. The regression lines are of the form:

$$\log X = \log a_0 + a_1 \log P_{tr} \Leftrightarrow X = a_0 P_{tr}^{a1},\ X \in \{U_o, E_e, Int\}$$

The slope of the regression line for U_o in Fig. 5 is 1.0, indicating that the relation between U_o and P_{tr} is approximately linear. In the LF-model E_e is a function of U_o and the skewing of the glottal pulse. The fact that both U_o and skewing increase with increasing P_{tr} explains why the slope for E_e (of 1.6, see Fig. 5) is larger than the slope for U_o. The slope of the regression line for Int (of 3.0) is about twice the value found for E_e, which is not surprising, because the Int of a freely travelling spherical sound wave is proportional to the square of the derivative of the mouth flow (Beranek, 1954). However, without the use of a proper production model it is difficult to unravel the exact underlying relations between the parameters.

3.4. Wave shape parameters

For the dimensionless wave shape parameters R_g, R_k, and R_a the following general relations can then be derived. R_g is almost constant; the correlation of R_g with T_0 is positive but very small (see Table III). For the range of R_g values found in this study, the influence of this parameter on the spectrum (and thus on voice quality) is very small. The correlations of R_a and R_k with T_0 (see Table III) are positive and highly significant ($p<0.0001$), which implies that voice quality changes with T_0 and consequently with F_0. The correlations of R_a and R_k with Int and P_{tr} were even higher (see Table III), so voice quality also changes with Int. The average values of R_g, R_k, and R_a were 108%, 41%, and 6.5% respectively and are in accordance with the values given by Carlson et al. (1989).

3.5. Deviations from the general behaviour

The fact that we have a large data set in which most parameters display consistent relations allows us to identify the outliers, i.e. the instances that do not fit in with the general pattern. Pitch periods that show different relations between the parameters are mainly found during voice onset and voice offset, and in the last syllable of an utterance.

The values of U_o for voice onset and offset generally fall below the regression line of U_o on P_{tr} that is given in Fig. 5, but there are also differences between voice onset and offset. The average P_{tr} during an UV/V transition (5.0 cm H_2O) is higher than the average P_{tr} during a V/UV transition (3.7 cm H_2O). It seems that higher P_{tr} values are needed to initiate vibration of the vocal folds, than to keep vibration going towards the end of a voiced interval. At the beginning of a voiced interval the average values of Int and F_0 (59 dB and 131 Hz) are also higher than those at the end of a voiced interval (57 dB and 120 Hz). Furthermore, a rise in T_a and T_n was found both towards beginning and end of a voiced interval.

Near the end of all 30 utterances there was a substantial decrease in P_{sub}, P_{tr}, Int, and F_0; and a marked increase in the activity of the sternohyoid. Also, for the final vowel U_o was relatively high, compared to the general trend. The deviating behaviour of the voice source during the final syllable was also observed by Klatt and Klatt (1990). This is described in more detail in Strik and Boves (in press).

4. Conclusions

In general, the method of automatic inverse filtering and fitting worked satisfactorily. Most problems were encountered with attempts to obtain a good approximation for the T_a parameter in pitch periods taken from consonants. For some glottal periods our method did not succeed in finding a combination of LF-parameters that define a LF-model that closely resembles dU_g. This could be a shortcoming of the inverse filter or the fitting procedure, but also of the LF-model. It remains to be seen if the LF-model can describe all variations in the glottal pulse that occur in different kinds of speech.

Consistent relations were found within the set of the time parameters and the set of amplitude related parameters, but also between the parameters of both sets. The highest correlations were found between P_{tr}, U_o, E_e, and Int. The behaviour of the voice source during voice onset, voice offset, and the last syllable was different from the general behaviour. When relating LF-parameters to prosody the general picture is that voice quality is mainly affected by R_k and R_a (or T_n and T_a), and that Int is mainly affected by E_e (or U_o).

All these fluctuations in the voice source parameters are likely to have perceptual consequences. To improve the naturalness of synthetic speech, these effects have to be taken into account.

Acknowledgements

This research was supported by the foundation for linguistic research, which is funded by the Netherlands Organization for the Advancement of Scientific Research N.W.O. Special thanks are due to dr. Philip Blok who inserted the hooked-wire electrodes and the pressure catheter in the experiments.

References

L. Beranek (1954), Acoustics (McGraw-Hill Book Company, New York), pp. 23-115.

R. Carlson, G. Fant, C. Gobl, B. Granstrom, I. Karlsson and Q. Lin (1989), "Voice source rules for text-to-speech synthesis", Proc. ICASSP, Vol. 1, pp. 223-226.

J.E. Dennis, D.M. Gay, and R.E. Welsch (1981), "An adaptive nonlinear least-squares algorithm", ACM Transactions on Mathematical Software, Vol. 7, pp. 348-368.

G. Fant, J. Liljencrants, and Q. Lin (1985), "A four-parameter model of glottal flow", STL-QPSR, Vol. 4, pp. 1-13.

G. Fant and Q. Lin (1988), "Frequency domain interpretation and derivation of glottal flow parameters", STL-QPSR, Vol. 2-3, pp. 1-21.

G.A. Ferguson (1987), Statistical analysis in psychology and education (McGraw-Hill Book Company, Singapore), pp. 195.

J. Jansen, B. Cranen, and L. Boves (1991), "Modelling of source characteristics of speech sounds by means of the LF-model", Proc. of EUROSPEECH ’91, Vol. 1, pp. 259-262.

D.H. Klatt and L. Klatt (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am., Vol. 87, pp. 820-857.

H. Strik and L. Boves (in press), "Control of fundamental frequency, intensity and voice quality in speech", J. of Phon.

J. de Veth, B. Cranen, H. Strik and L. Boves (1990), "Extraction of control parameters for the voice source in a text-to-speech system", Proc. of ICASSP-90, paper 21.S6a.2.

Table I. Correlations between 4 amplitude related parameters (P_{tr}, U_o, E_e, Int) and 3 time parameters (T_o, T_n, T_a) for 613 voiced periods.

	P_{tr}	U_o	E_e	Int
T_o	-0.44	-0.17	-0.35	-0.45
T_n	-0.41	-0.19	-0.48	-0.42
T_a	-0.31	-0.36	-0.50	-0.36

Table II. Correlations between $\log P_{tr}$, $\log U_o$, $\log E_e$, and logInt for 613 voiced periods.

	$\log P_{tr}$	$\log U_o$	$\log E_e$
$\log U_o$	0.60		
$\log E_e$	0.63	0.86	
logInt	0.81	0.78	0.81

- Figure captions -

Fig. 1. Glottal flow (Ug) and glottal flow derivative (dUg) with the parameters of the LF-model.

U_o: maximum of U_g

E_i: maximum of dU_g

E_e: absolute value of the minimum of dU_g

t = 0: time of glottal opening

T_c: time of glottal closure

T_i, T_p, T_e: time points of E_i, U_o, and E_e respectively

T_a: the time between T_e and the projection of the tangent of dU_g in $t=T_e$

$T_n = T_e - T_p$

The dimensionless wave shape parameters than can be derived from the LF-parameters are:

$R_g = T_0/2T_p$

$R_k = T_e/T_p - 1 = T_n/T_p$

$R_a = T_a/T_0$

Fig. 2. Results of the automatic fitting procedure for five periods of a vowel /e/. Shown are, from top to bottom, audio signal, glottal flow derivative (dU_g, solid line) with fitted signal (dotted line), and glottal flow (U_g, solid line) with fitted signal (dotted line).

Fig. 3. Results for a voiced interval to illustrate the behaviour of the voice source parameters. Given are, from top to bottom, phonetic transcription, audio signal, trans-glottal pressure (P_{tr}), median values of E_e and U_o, intensity (Int), and median values of T_0, T_a, and T_n. Although /p/ is phonologically an unvoiced plosive, it is observed that voicing continues in this utterance.

Fig. 4. The relation between the time parameters and T_0. Given are the regression lines of the time parameters as a function of T_0. Note that both the horizontal and the vertical axis have a logarithmic scale.

Fig. 5. Scatterplots of the amplitude related parameters U_o, E_e, and Int as a function of P_{tr}, with regression lines. Note that both the horizontal and the vertical axis have a logarithmic scale.