ON THE RELATION BETWEEN VOICE SOURCE CHARACTERISTICS AND PROSODY

Helmer Strik & Louis Boves

Department of Language and Speech, Nijmegen University,
P.O. Box 9103, 6500 HD Nijmegen, the Netherlands

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 the LF-model to the data. Consistent relations between voice source parameters and prosody were observed.

Keywords: inverse filtering; LF-model; voice source

1. INTRODUCTION

Many present day text-to-speech systems produce speech that is intelligible, but doesn't sound natural. This lack of naturalness is at least in part due to the absence of voice source control rules. Numerous different voice source models have been proposed, some of which could be very useful for speech synthesis. But if a sophisticated voice source model is to improve speech quality, one has to be able to control its parameters. Therefore, there is a need for data on the behaviour of the voice source, or more specifically of the behaviour of those characteristics of the source that can be mapped onto the model parameters. These data can be used to extract rules for the model parameters, that can then be used to improve synthesis.

To extract rules a large amount of data is required. Both inverse filtering of the speech, and fitting a model to the inverse filter results could be done by hand. However, this is time consuming, subjective and thus probably not reproducible. Therefore a procedure is developed to derive the voice source parameters automatically.

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 from these data obtained from isolated speech segments it is difficult to formulate rules for whole utterances. Therefore, our aim is to study the behaviour of the voice source in connected, preferably spontaneous speech. And apart from the vowels we also want to extract source parameters for voiced consonants, V/UV and UV/V transitions. Research on these topics is now in progress. In this article some results are presented. Special attention is given to the relation between voice source dynamics and prosody.

2. METHOD AND MATERIAL

2.1. SPEECH MATERIAL

To study voice source characteristics data were obtained 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). For the current article only data of one male subject were used. Near the end of a recording session he was asked to produce an utterance spontaneously. He then repeated this utterance 29 times. The experiment is described in more detail in Strik and Boves (in press). For this paper inverse filter results were obtained for two of the 30 utterances.

2.2. INVERSE FILTERING

The speech signals were transduced by a condensor microphone (B&K 4134) placed about 10 cm in front of the mouth, 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 digitized speech signal was processed with a phase correction filter in order to undo the low frequency phase distortion. Closed glottis interval covariance LPC was used to estimate the parameters of the inverse filter. In Veth, Cranen, Strik & Boves (1990) it was shown that this technique is as powerful as more sophisticated techniques, like Robust ARMA analysis. The moment of glottal closure was determined from the EGG. Inverse filtering yields an estimate of the differentiated glottal volume flow (dU_g/dt); integration gives the flow signal (U_g).

2.3. VOICE SOURCE PARAMETERS

Voice source parameters were obtained by fitting a voice source model to the data. 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). In terms of the LF-model the maximum in U_g (U_o) is reached at time T_p, the minimum in dU_g/dt (E_e) at time T_e, and T_a is defined by the tangent of dU_g/dt at the beginning of the return phase. In this research U_o and E_e were not calibrated, and are given in arbitrary units. T_n is the length of the interval between T_p and T_e, and it is related to the skewing of U_g. T_a is related to the spectral tilt (see e.g. Fant and Lin, 1988).

For automatic fitting of a model to the signals use was made of a special software package (details are given in Jansen, 1990). The fit is done pitch synchronously. The periods are defined by the minima in dU_g/dt, because these time points can be located most reliably. Automatic fitting seems possible, although for the return phase it is difficult to obtain reliable, stable parameters (see Jansen, 1990).

2.4. AVERAGING THE RESULTS

For inverse filtering a number of choices must be made. The most important are the length and exact position of the analysis window, and the order of the analysis. Generally, there seems to be no combination of these parameters that is optimal for each individual pitch period in a normal speech utterance. LPC-12 worked satisfactorily for almost all voiced frames.

The following strategy was adopted. For each utterance inverse filtering was done with a number of different analysis windows, i.e. for all 15 combinations of 5 window lengths and 3 window shifts. In addition, inverse filtering was done for closed glottis intervals (of variable length) that were derived automatically from the EGG. Voice source parameters were extracted for all 16 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 values were used for further analysis.

3. RESULTS

Inverse filtering of the vowels proved to be fairly easy; the voiced consonants, and especially the first and last periods of a voiced segment gave more difficulties. In general it was observed that the lower transglottal pressure (P_{tr}), the more difficult it is to obtain reliable inverse filter results. For all data the parameters obtained for the return phase (T_a) were less stable than those of the exponential growing sine wave.

Rapid changes in the voice source parameters were observed at the beginning and end of voiced intervals. At voice onset P_{tr}, U_o, E_e, and intensity level (IL) increase until they reach a more or less steady level; at voice offset the reverse happens. The data for UV/V and V/UV transitions were separated from the other data by visual inspection. Because there were also differences between voice onset and offset, these data were analyzed separately (see section 3.2). The data of the final vowel are presented in section 3.3. All remaining data fall into the category called steady phonation. These data are treated first (see section 3.1.) and serve as a reference against which the other data are compared.

3.1. STEADY PHONATION

During the course of all 30 utterances a gradual decline in P_{sub}, P_{tr}, IL, and fundamental frequency (F_o) was observed; in the individual voiced segments P_{sub} was almost constant and P_{or} covaried with P_{tr}, IL, and F_o. Consequently, a large covariance between P_{tr}, IL, and F_o was found for all data of the 30 utterances (Strik and Boves, in press).

For the voice source parameters U_o and E_e the same tendencies were observed (see Table I and Fig. 1). At first sight the high covariance of P_{tr}, U_o, E_e, and IL 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

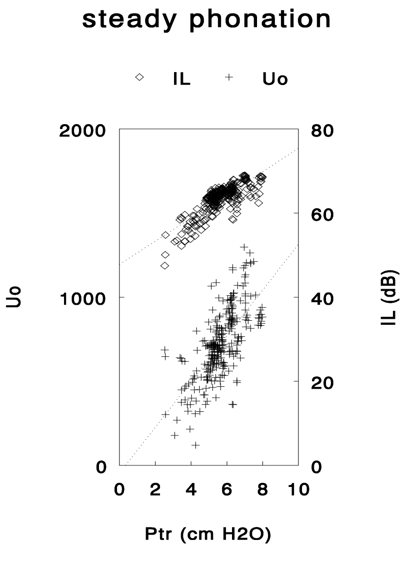

Figure 1. Scatterplot of U_o and IL as a function of P_{tr}, with regression lines.

amount would lift the spectrum (see Fant and Lin, 1988), and thus increase IL. However, the data of steady phonation 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. Given the large variation in articulatory gestures, it is surprising that the covariance between P_{tr}, U_o, E_e, and IL is consistently high. Further research is needed to unravel the underlying relations.

	P_{tr}	U_o	E_e	IL
U_o	0.69			
E_e	0.66	0.81		
IL	0.80	0.70	0.69	
F_o	0.42	0.29	0.32	0.47

Table I. Correlations between P_{tr}, U_o, E_e, IL, and F_o for steady phonation (N = 280, |R| > 0.24 for p<0.0001)

The correlation between P_{tr} and F_o is much lower than the correlations between P_{tr}, U_o, E_e, and IL (see Table I). Strik and Boves (1989) studied the relation between P_{tr} and F_o in connected speech, and found that the activity of laryngeal muscles is an important factor in this relation. Probably, the variables that are not used in the present article (like activity of laryngeal muscles) have more effect on the relation between P_{tr} and F_o (and most likely also on T_a and T_n), than on the relations between P_{tr}, U_o, E_e, and IL.

3.2. VOICE ONSET AND OFFSET

Scatterplots of U_o and IL as a function of P_{tr} are given in Fig. 2 and Fig. 3, for voice onset and offset respectively. Also given are the regression lines for steady phonation (see Fig. 1). It is observed that for UV/V and V/UV transitions U_o is relatively lower compared to steady phonation, but that there are also differences between voice onset and voice offset.

The average P_{tr} for voice onset (4.9 cm H_2O) is higher than the average P_{tr} for voice offset (3.6 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 IL and F_o (59 dB and 130 Hz) are also higher than those at the end of a voiced interval (57 dB and 120 Hz).

Both towards beginning and end of a voiced interval a rise in T_a and T_n was observed. A higher T_a would make the spectral slope more steep, and thus lower IL. The result of increasing T_n alone is that the flow pulses are less skewed, decreasing E_e and IL. Both changes would therefore affect IL and the spectrum of the audio signal.

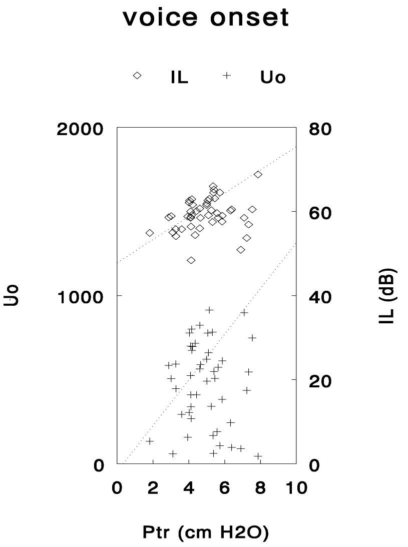

Figure 2. Scatterplot of U_o and IL as a function of P_{tr}, and regression lines for the steady phonation data.

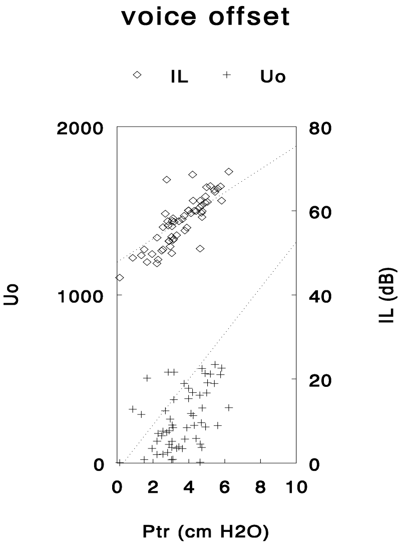

Figure 3. Scatterplot of U_o and IL as a function of P_{tr}, and regression lines for the steady phonation data.

3.3. FINAL VOWEL

Near the end off all 30 utterances there was a substantial decrease in P_{sub}, P_{tr}, IL, and F_o; and a marked increase in the activity of the sternohyoid. Also, for the final vowel U_o was relatively high, compared to the data for steady phonation (see Fig. 4). The deviating behaviour of the voice source during the utterance final syllable (see also Klatt and Klatt, 1990) was studied by comparing the inverse filter data of the last vowel /a/ to the data of the first vowel /a/. The results of this comparison are that U_o of the last vowel is higher, although P_{tr}, E_e, IL, and F_o are considerably lower. Furthermore, when all parameters are expressed in percentages of the period duration, no major differences were found between the relative time parameters of the fitted flow signals for both vowels. This means that apart from time stretching (increase of T_o), there were no significant differences in the shape of the flow pulses.

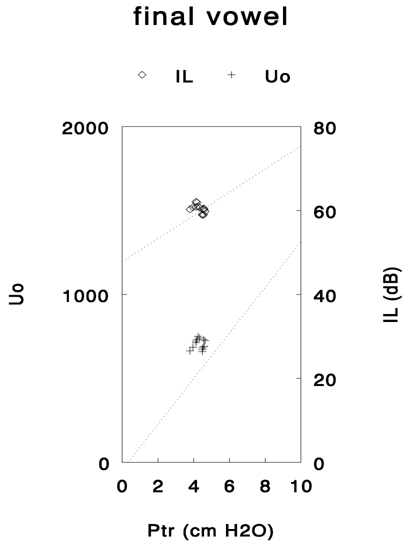

Figure 4. Scatterplot of U_o and IL as a function of P_{tr}, and regression lines for the steady phonation data.

The fact that U_o of the final vowel is higher, while P_{tr} is about 1.5 cm H_2O lower, indicates that the impedance of the glottis must have been lowered. Because the shape of the flow pulses is almost the same for both vowels, no large differences in degree of adduction and open quotient are expected. The latter was confirmed with measurements from the EGG. Probably the amplitude of vibration of the vocal folds is increased by slackening the vocal folds. This could be done by diminishing the antero-posterior tension of the folds. Strik and Boves (1989) indeed found a suppressed activity of the cricothyroid and vocalis near the end of a declarative utterance.

As the relative time parameters for both vowels are about equal, the decrease in E_e (± 0.7 dB) is the result of the increase in U_o (± 0.3 dB) and the increase in T_o (± 1.0 dB) alone. The effect of these changes on the spectrum is described in Fant and Lin (1988). The increase in U_o increases the amplitudes of the lower harmonics; the decrease in E_e causes a lowering of the high-frequency part of the spectrum; while the increase of T_a causes an increase in the spectral tilt (T_a/T_o is about the same for both vowels, but T_a is much larger for the last vowel). When the spectra of both vowels are compared these differences are clearly visible.

4. CONCLUSIONS

A consistently high covariance between P_{tr}, U_o, E_e, IL, and F_o has been observed for steady phonation. Increasing U_o and E_e would lift the spectrum and thus increase IL, and increasing F_o would change the position of the harmonics in the spectrum. Furthermore, both T_a and T_n rise, when going towards the beginning or end of a voiced interval, or from a vowel to a voiced consonant. All these fluctuations in the voice source parameters, and especially those during the final vowel, would probably have perceptual consequences. To improve the naturalness of synthetic speech, these effects have to be taken into acount.

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

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

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

Jansen, J. (1990) Automatische extractie van parameters voor het stembron-model van Liljencrants & Fant. Unplubished masters thesis, Nijmegen University.

Klatt, D.H. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America 87, pp. 820-857.

Strik, H. & Boves, L. (1989) The fundamental frequency - subglottal pressure ratio. In Proceedings of EUROSPEECH-89, Vol. 2, pp. 425-428.

Strik, H. & Boves, L. (in press) Control of fundamental frequency, intensity and voice quality in speech. Journal of Phonetics.

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