ANALYSIS AND ARTICULATORY SYNTHESIS OF DIFFERENT VOICING TYPES

C. Scully, K. Stromberg & D. Horton (1), P. Monahan, A. Ní Chasaide & C. Gobl (2)
(1) University of Leeds, Leeds, England
(2) Trinity College, Dublin, Ireland

ABSTRACT

Results of inverse filtering from both airflow and sound pressure signals for a real speaker have been combined to develop the Leeds phenomenological model of voicing. Three controlling physiological variables are mapped onto each of three voice source waveform parameters for nine voicing types, with the aim of simulating variations of the voice source with phonetic context.

INTRODUCTION

The Leeds phenomenological model of voicing relates voice source waveshapes to the physiological conditions which generate it, for a particular speaker. A three-parameter representation of the voice source developed by Fant is used [1]. The approach has been described elsewhere [2]. Previously, voice source waveforms for natural speech were obtained by inverse filtering the total volume flowrate of air through the mouth and nose combined, Ut. This method has the advantage of giving an estimate for the dc flow through a glottal chink; but has the disadvantage that one of the acoustic source parameters, the asymmetry factor K, is difficult to estimate.

Here the same kind of real speech data were obtained but for the inverse filtering two signals were used: simultaneously recorded airflow Ut and sound pressure Sp. Robust features from each method were combined, with the aim of improving the reliability of the voice waveshape description.

Three physiological controlling parameters are used, closely related to those for the two-mass model of voicing [3]. They are the low frequency component of the glottal area AG, the pressure drop across the glottis PDIFF and the mass-tension factor for the vocal folds Q.

METHODS

Subject

The main male speaker for the SPEECH MAPS project, PB, who has a general French accent without Southern French features, provided the speech data.

Signals

An undivided Rothenberg mask and an orally inserted pressure tube with a Gaeltec pressure transducer were used, combined with a B & K condenser microphone outside the mask and a laryngograph. Four channels of data were recorded at Leeds directly onto a pc with a sampling frequency of 10 kHz: sound pressure Sp, laryngograph signal Lx, total output airflow Ut and intra-oral air pressure Po.

Speech Material

Corpus G of the SPEECH MAPS project is central to the Leeds voicing model. It contains multiple repetitions on a single expiratory breath of [pœ] for nine voicing types. These are three levels: medium, soft, loud; three pitches: mid, high, low; three phonation types: normal, breathy pressed. The aim is to describe a range approximating to the normal limits of the voice source for speech. The mid-open, central quality vowel is chosen with the aim of achieving an approximation to the following conditions: [œ] has a fairly high F1 frequency which is suitable for inverse filtering; the jaw movement down from that for [p] to that for the vowel

is not large, so that large changes in vocal tract volume are avoided. The vowel is made rather long, so that a quasi-static jaw and tongue configuration can be assumed near mid-vocoid, where the analyses are performed. It is hoped that, as a result, the errors introduced by assuming that output airflow is a good approximation to transglottal airflow when estimating glottal area and when performing inverse filtering from airflow are minimised.

The [p-p] context is needed for estimates of subglottal pressure as described below. The speaker tried to avoid pitch changes and other phonetic exponents of stress, with the aim of keeping subglottal pressure changes as small as possible. The multiple repetitions, about 10 to 11 for each voicing type, permit multiple analyses, with only lung volume changing from one repetition to the next.

Voiced fricative sequences from multiple repetitions on one expiratory breath of [paCa] were used also, from Corpus 3 of the SPEECH MAPS project.

ANALYSES

Four repetitions of each of the nine voicing types were analysed in several ways at the same mid-vocoid time point. The four channel data files were converted to ESPS "Waves" format. Signals were displayed using this software and the ESPS programs were modified at Leeds according to the analyses required. The signals Ut and Po were low-pass filtered for the aerodynamic analyses described in this section.

Estimation of the controlling parameters

Subglottal pressure PSG was assumed to equal intra-oral air pressure Po at the end of the initial rapid rise in Po during the [p] closures. Linear interpolation gave an estimate for PSG at mid-vocoid.The pressure drop across the glottis PDIFF was calculated as (PSG - Po).

Glottal area AG was estimated at mid-vocoid by assuming that the transglottal flow UG was equal to Ut here. The orifice equation [4] was used:

$$AG = k.UG/(\sqrt{PDIFF}) \qquad (1)$$

(where k = 0.00076 with AG in cm^2, UG in cm^3/s, PDIFF in cm H_2O).

Complexity of fundamental frequency F0 patterning is derived by assuming that a myoelastic component Q and an aerodynamic component PDIFF together control F0 from:

$$F0 = Q + KF.PDIFF \qquad (2)$$

KF is an empirical constant to be determined for a given speaker. For some of the voiced fricatives of Corpus 3, PSG and PDIFF were estimated as described above. KF values were calculated as dF0/dPDIFF, with a mean value of 4.1 Hz/cmH_2O for speaker PB. Q was obtained from F0 by equation 2.

Estimation of the voice source parameters

Inverse filtering was done using methods and software developed at Trinity College Dublin [5]. Three successive cycles of voicing were used, at or very close to the predefined mid-vocoid time point. Total output airflow Ut, unfiltered, was used to estimate total volume flowrate of air through the glottis Ut,g with its acoustic component Ug, and dc flow also. Simultaneously recorded Sp gave the differentiated glottal flow Ug' (dUg/dt in Figure 1). The number of pairs of conjugate complex poles the program was to find was set at 6. For a sampling frequency of 10 kHz the number is normally 5 for a male speaker; however, an extra pair was added, on the basis that the flow frequency response of the Rothenberg mask needed to be taken into account (see [6]).

The Leeds voice waveshape, based on Ug, is not mathematically equivalent to the LF model of voicing [7], based on Ug'. But comparisons of pairs of waveforms showed that some time points could be aligned; features of the Ug′ wave could be used to enhance the analysis of the Ug wave. Figure 1 shows both waveshapes. The LF model has four parameters in addition to T0 while the Leeds model has only three; it lacks a return phase after excitation. Three robust parameters were used here; they are the same three shown to characterise a speaker's voice using the LF model [8]. In our nomenclature they are VOIA, EE and T0. The three acoustic parameters required to define the voice source waveform in the Leeds model are: VOIA, the amplitude of the acoustic component of flow; TCR, the ratio of closed time over total periodic time (TCR = 1 - open quotient) and K, an asymmetry factor. The three parameters were obtained as follows :

1. The dc flow level was drawn as an averaged value where Ug was low. This is the 'closed' phase, which does not necessarily mean complete closure; indeed usually it does not.
2. VOIA was defined from the dc flow level to the peak flow. Any ripples remaining in the Ug trace were smoothed by eye.
3. For TCR, the start and end of the closed phase were both defined from Ug′. The time point for EE (the maximum negative value of Ug′) defined the start of the closed phase. The start of the rise of Ug′ up from its zero line defined the end of the closed phase. This time point was difficult to locate, whether Ug or Ug′ was used. T0 was obtained from EE points in successive cycles of Ug′.
4. In our previous work, the asymmetry factor K was measured from the gradient of Ug half way up the rising portion to give TB; the gradient at closure as seen on Ug gave TD. K was calculated [1] from the equation :

$$K = 0.5 + 0.125(TB/TD)^2 \qquad (3)$$

This was very difficult to do with any degree of confidence. Different measurements were made here, to approximate to this formula for K, as follows:

On Ug′, EI is the maximum positive value of the gradient of Ug.

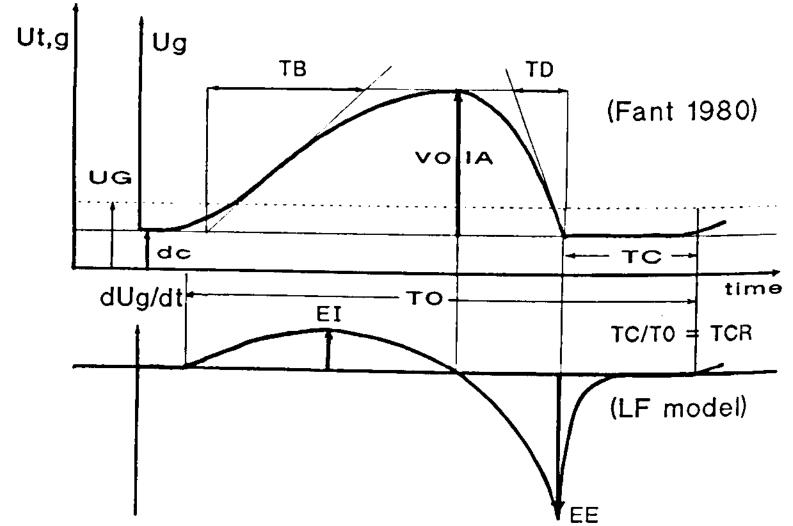

Figure 1. Waveshape representations of the voice source: flow Ug above, differential of flow Ug′ below.

It was found to be located in time quite near half way up the rising portion of Ug, so approximately:

$$TB = VOIA/EI \qquad (4)$$

Similarly, on Ug′, the absolute value of EE gave an approximation to the (negative) gradient for Ug near 'closure' in the Leeds model, so approximately:

$$TD = VOIA/|EE| \qquad (5)$$

$$\text{So} \quad K = 0.5 + 0.125(|EE|/EI)^2 \qquad (6)$$

CONSTRUCTION OF THE VOICING MODEL

Stepwise multivariate regression on the four repetitions of the nine voicing types was used to obtain the relationship between each voice waveshape parameter, the dependent variable, and the three physiological, independent controlling variables. The three equations obtained were :

$$VOIA = -197.00 + 68.80\ PDIFF + 982.00\ AG + 1.49\ Q \qquad (7)$$

$$TCR = +0.33 + 0.03\ PDIFF - 0.65\ AG \qquad (8)$$

$$K = +4.08 - 0.10\ PDIFF - 2.49\ AG - 0.02\ Q \qquad (9)$$

In addition to these mapping equations, upper and lower limits were set for the voice source parameters, based on the data for speaker PB.

SIMULATIONS OF THE SPEAKER

The Leeds composite forward model of speech production is being used to simulate speaker PB's productions. Comparisons between the model and the natural speech are made for articulatory paths, aerodynamics, acoustic sources and output speech signal, real or synthetic. As a first step, the adequacy of the voicing model can be assessed by simulating vowels produced with different voicing types. The power of the voicing model to go beyond the vowel data on which it is based is being investigated with sequences containing voiced fricatives.

Values for the dc flow through a glottal chink, obtained here for the nine voicing types, have been related to UG and the voice source waveform parameters [2]; the current analyses confirm the findings.

ACKNOWLEDGEMENTS

This work was partially funded by an ESPRIT basic research project, no. 6675, SPEECH MAPS. We thank the speaker, our colleague Pierre Badin.

REFERENCES

[1] Fant, G. (1980), "Voice source dynamics", *STL-QPSR*, Stockholm 2-3, pp.17-37.

[2] Scully, C. and Stromberg, K. (1992), "Physiologically-controlled voice source models for different speakers", *Proc. Inst. of Acoustics*, 14, pp.463-471.

[3] Ishizaka, K and Flanagan, J. L. (1972), "Synthesis of voiced sounds from a two-mass model of the vocal folds", *Bell Syst Tech*, 51, pp.1233-1268.

[4] Scully, C. (1986), "Speech production simulated with a functional model of the larynx and the vocal tract", *J Phonetics*, 14, pp. 407-414.

[5] Ni Chasaide, A., Gobl, C. and Monahan, P. (1992), " A technique for analysing voice quality in pathological and normal speech", *J Clinical Speech and Language Studies*, Vol. 2, pp. 1-16.

[6] Hertegard, S. and Gauffin, J. (1992), "Acoustic properties of the Rothenberg mask", *STL-QPSR* 2-3, pp. 9-18.

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

[8] Fant, G., Kruckenberg, A., Liljencrants, J. and Båvegard, M. (1994), "Voice source parameters in continuous speech: transformation of LF parameters", *Proc. ICSLP-94*, Yokohama, Vol. 3, pp. 1451-1454.