GESTURAL SUPERVISOR FOR THE VOCAL CORDS OF A SPEAKING MACHINE

Christophe Vescovi, Eric Castelli

Institut de la Communication Parlée U.R.A. - CNRS N° 368
I.N.P.G./E.N.S.E.R.G. - Université STENDHAL
46, Avenue Félix Viallet, 38031 GRENOBLE Cedex 1

ABSTRACT - Until now, inversion methods of the speech signal mainly deal with the representation of the vocal tract. This work extends those inversion models to the voice source using the two-mass model of the vocal cords. In a first step, the characterisation of the glottal flow produced by this model is done in order to define an appropriate control space. Then a forward model of the two-mass model is built, giving relations between the commands of the voice source and the control parameters, taking into account the alterations of the glottal flow caused by vocal tract configurations. The back propagation algorithm used with this forward model can predict the commands to use in order to reach a specified point in the control space. The validation of the algorithm is done inverting synthetics signals, thus results of the inversion can be compared with the commands used for the synthesis. Finally, first results on the inversion of natural speech are presented.

INTRODUCTION

The speech production model used at I.C.P. (Bailly, Castelli & Gabioud 94, Castelli & Badin, 93) can be divided in three parts (Figure 1): two acoustic modellings based on the Kelly & Lochbaum model (Kelly & Lochbaum, 62) for the lungs and the vocal tract, and a physical model of the vocal cords, the two-mass model (Ishizaka & Flanagan, 72). The interactions between the subglottal cavities, the supraglottal cavities and the model of the glottis, have been particularly studied (Trinh Van & Al, 92). In order to simplify the articulatory strategies, the model of production is coupled with the articulatory model of Maeda (90). Thus, the number of commands of the whole model is reduced to eleven.

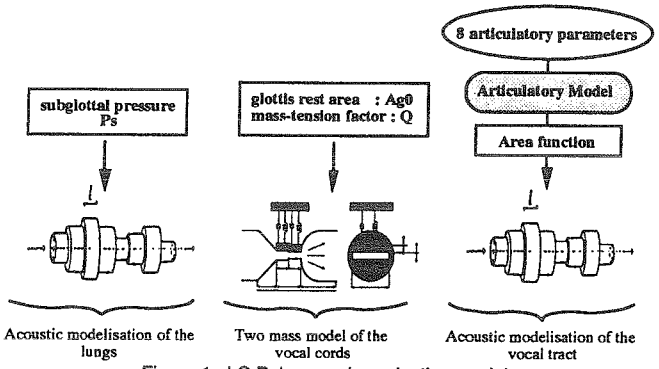

Figure 1 . I.C.P. 's speech production model

The robotics approach supporting this work consists in building a gestural supervisor able to learn the relations between a given phonological sequence and the articulatory trajectories that must be provide to the speaking machine. This problem is known under the general term of inversion. The inversion of the vocal tract as already been studied at I.C.P. (Bailly & Al, 91) and the back propagation approach used is here extended to the voice source. This approach is based on the forward modelling of the system under control. The forward model must provide an approximation of the output of the voice source model knowing the applied commands.

CHARACTERISATION OF THE GLOTTAL FLOW

The first problem to solve is that the glottal flow is filtered by the vocal tract in the speech signal. So as to make measurements on the glottal flow, the speech signal has to be inverse filtered. The inverse filtering technique used here is based on the IAIF algorithm (Alku & Al, 91). Then, the evaluation of the glottal flow must be characterised by a restricted number of parameters in order to minimise the complexity of the gestural supervisor. Works on the characterisation of the glottal flow using the LF model (Fant, Liljencrants & Lin, 85) have point out that the glottal flow of one speaker may be characterised by only three parameters (Fant & Al, 94). The parameters used here (Figure 2) are the fundamental frequency of the oscillations of the vocal cords Fo, the energy of the speech signal E and a wave shape parameter, the declination ratio Rd.

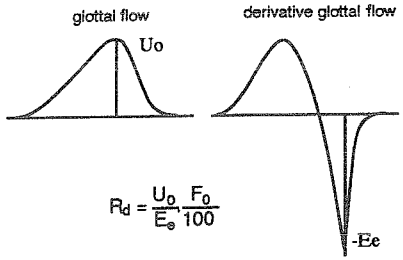

Figure 2 : Characterisation of the glottal flow

SOURCE-TRACT COUPLING

The first interaction between the source and the tract implemented in our speech production model is explained by the Figure 3. The glottis is considered as an elementary tube (similar to the other tubes of the vocal tract) with a variable area Ag (the instantaneous area of the glottis computed by the two-mass model). The main effect of this interaction is a change in the transfer function of the lungs-vocal tract filter during a glottal cycle, but have no preponderant effect on the glottal flow.

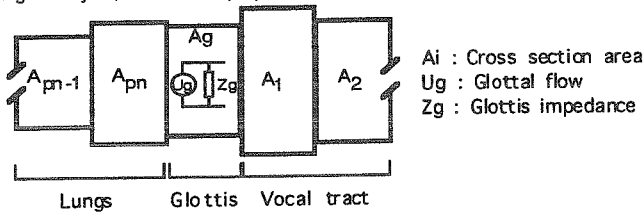

Figure 3 : Integration of the glottis in the vocal tract

The two-mass model is not driven by Ps but by the difference between the supraglottal and the subglottal pressures which has two main effects on the glottal flow. First, the pressure drop caused by marked constriction in the vocal tract reduces the intraglottal pressure and so as the energy and the fundamental frequency of the speech signal (Figure 4). The presence of acoustic pressure near the glottis has a non-negligible effect on the glottal flow, the interaction ripple on the glottal flow is an example of this, but the acoustic pressure in the vocal tract also causes changes in the oscillations of the vocal cords. This effect of the acoustic load of the vocal tract is shown in Figure 5. When the fundamental frequency or one of its harmonic corresponds to the first formant, the air consumption of the two-mass model is reduced, so at a constant lungs pressure, the amplitude of the glottal flow increases.

Considering this short study, the characterisation of the interactions must be done by characterising first, the aerodynamic pressure drop in the vocal tract and secondly, the acoustic load of the vocal tract.

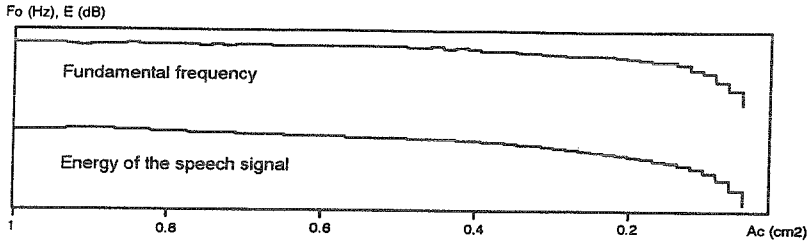

Figure 4 : Influence of the area of the constriction Ac in the vocal tract on Fo and E.

In our model the DC pressure drop in an elementary tube depends mainly on the inverse of the area of the tube. So a good parameter to characterise it is the inductance of the vocal tract lvt. The acoustic load of the vocal tract is then characterised by lvt an Xvt, the position of the half inductance in the vocal tract calculated from the glottis to the lips.

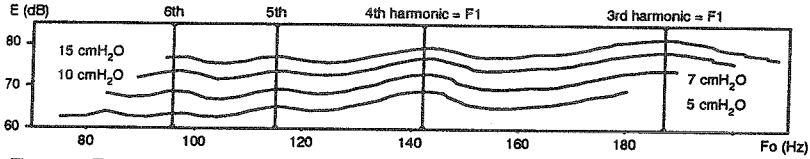

Figure 5 : Energy of the speech signal versus the fundamental frequency with a same vocal tract configuration. Vertical lines show the correspondence between one harmonic of Fo and F1 (565 Hz)

FORWARD MODEL

The command space is sampled in order to build a codebook between the inputs and the outputs of the model. The sampling is done so as to describe the whole working space of the two-mass model in the production of non pathological voices.

The forward model must be defined using an analytic function in order to simplify the inversion. The function choose here is an order-four polynomial function. Coefficients of the polynomial function are determined using the codebook. The RMS error caused by the polynomial approximation is respectively of 1%, 3% and 5% of the variation of Fo, E and Td. The approximation of E and Td seems quite poor, it can be explained by the influence of the acoustic load of the vocal tract on those parameters: the characterisation of the acoustic load of the vocal tract is very simple and the obvious non-linearity of this effect are not taken into account by the polynomial approximation.

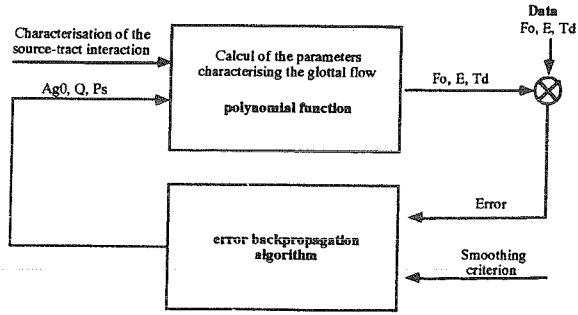

Figure 6 . Implementation of the inversion method of the vocal cords.

The gestural supervisor is based on a gradient backpropagation algorithm (Rumelhart, Hinton & Williams, 1986) that minimises the output error (configurational error) and also some smoothing errors in order to take into account kinematics properties.(Figure 6)

VALIDATION

The validation of the inversion method is done using signals generated by our speech production model. Thus, it is very simple to compare the commands given by the inversion method and those used for the synthesis. Moreover, the glottal flow given by the model is used, so there is no measurement errors due to inverse filtering imperfections.

For the first signal (PRESSURE) constant value for Ag0 and Q (0.05 cm^2 and 2) and a linear variation of Ps from 5 to 10 cmH_2O is used. For the second signal (TENSION), Ag0 and Ps are constant (0.05 cm^2 and 7 cmH_2O) and Q varies linearly from 1 to 3. The vocal tract geometry used for both signals corresponds to the vowel /e/.

- Inversion of PRESSURE

The main effect of an increasing subglottal pressure is the raise of Fo and E. A high subglottal pressure increases also high frequency components of the voice source: Rd decreases when Ps increases. The result of the inversion is shown of Figure 7. Ps is a little under-estimated by the inversion method. This may be caused by an error in the polynomial approximation of E. This defect of the approximation of E is certainly caused by the source-tract interaction.

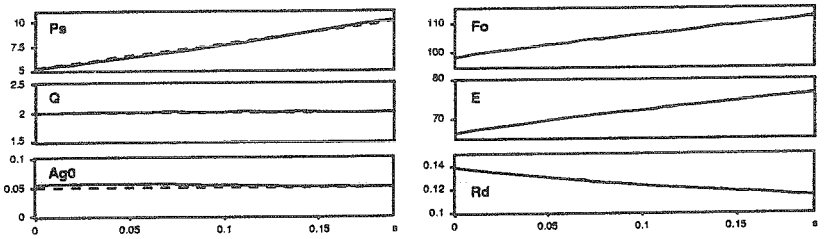

Figure 7. Inversion of PRESSURE : dotted lines are targets or effective commands used for the synthesis, plain lines are results of the inversion
(for Fo, E and Rd, dotted and plain lines are superposed).

- Inversion of TENSION

The main effect of the linear variation of Q is the raise of Fo. The effects on the other control parameters are not significant. The oscillations on E and Rd are the result of the source-tract interaction (see figure 5). The results of the inversion are quite good, but there are oscillations on Ps following the variation of E (figure 8). Once again this is caused by the non-linearity of the source-tract interactions that are impossible to take into account with our polynomial approximation.

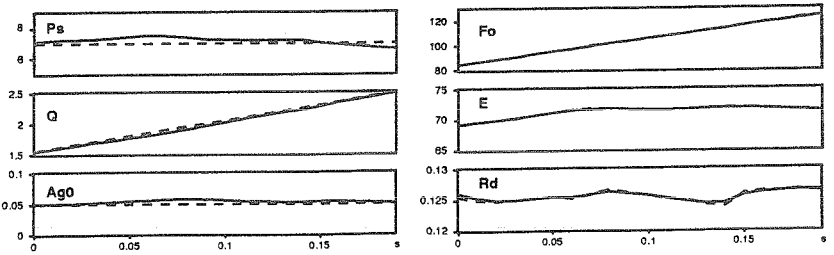

Figure 8. Inversion of TENSION.
(Ps in cmH_2O, Ag0 in cm^2 and Q no unit)

INVERSION OF NATURAL SPEECH

The signal used for the inversion is the logatome /iai/ with an emphasis on the vowel /a/, extracted from the french sentence "il y a immediatement .. ". The inversion of the vocal tract has already be done on those signals and this logatome presents two opposite positions of the tongue and so two very different interactions with the vocal tract. The characterisation of the source-tract interactions is calculated using the inverted vocal tract while the glottal flow is estimated using inverse filtering. The low frequency phase distortion present on the recorded speech signal change the shape of the inverse filtered signal which means that the measurement of Rd will not be precise. This is due to the bad quality of the recording equipment used (this signal was not recorded to do inverse filtering). So only the parameters Fo and E are used as targets in the inversion algorithm which shows the ability of it to solve "many to one" problems.

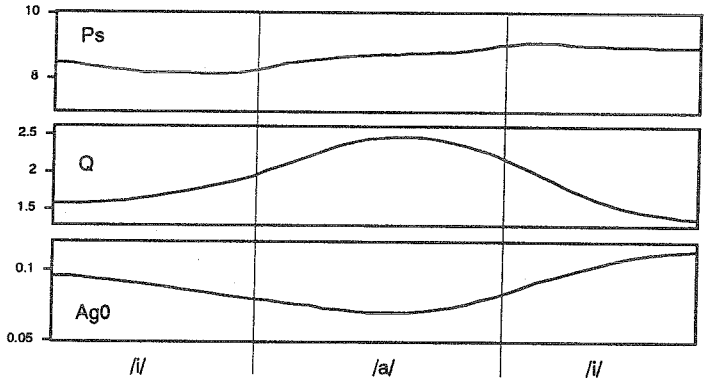

Figure 9. Results of the inversion of /iai/

The three commands parameters found by the inversion are shown on figure 9. Ps is quite constant, it raise a little bit during the vowel /a/ which can be explained by the emphasis on this vowel. But the main effect of the emphasis is a raise of the fundamental frequency caused by the increase of the mass-tension factor Q in the vowel /a/. Ag0 decreases slightly during the vowel /a/ which may be due to the muscular action used to change the vocal cords tension.

CONCLUSIONS

This first study shows that the inversion method gives coherent results. The inversion of synthetics signals shows that the error backpropagation algorithm is adapted to the problem of inversion. The polynomial approximation of the forward model gives a good description of the two-mass model working but is not able to describe correctly the influence of the source-tract coupling. The action of this coupling is not linear and it would be difficult to characterise it using polynomials. The inversion of natural voice shows that our forward model could be used with natural speech providing realistic evolution of the two-mass model commands.

However the approximation errors are not negligible and the characterisation of the source-tract interactions is not optimal. In order to improve the forward model, it might be necessary (1) to take in account the non-linear interactions; (2) to make dynamic inversion because the two-mass model is a mechanical model with a non-negligible time lag, this time lag appears clearly with important variations of Ag0 (devoicing) and must be taken into account to make the inversion of speech sequences containing unvoiced consonants or stops.

REFERENCES

ALKU P., VILKMAN E., LAINE U.K. (1991) *Analysis of glottal waveform in different phonation types using the new IAIF-method.* Proceed. of the XIIth International Congress of Phonetic Sciences, Aix en Provence, 362-365.

BAILLY G., CASTELLI E., GABIOUD B.(1994) *Building prototypes for articulatory speech synthesis.* Proceed. of the Second ESCA/IEEE Workshop on Speech Synthesis, New York, 9-12.

BAILLY G., LABOISSIERE R. & SCHWARTZ J.L. (1991) *Formant trajectories as audible gestures : an alternative for speech synthesis.* Journal of Phonetics, 19-1, 9-23.

CASTELLI E. & BADIN P. (1993) *Time and Frequency Domain Acoustic Models of the Vocal Tract.* Speech Maps - Mapping of Action and Perception in Speech - WP2 Deliverable n° 5, Appendix B, 1-25.

FANT G., KRUCKENBERG A., LILJENCRANTS J., BAVEGARD M. (1994) *Voice source parameters in continuous speech. Transformation of LF-Parameters.* In Proceedings of the International Conference on Spoken Language Processing, September 18-22, 1994, Yokohama, Japan, Vol.3, paper S25-4, 1451-1454.

FANT G., LILJENCRANTS J., QI-QUANG L. (1985) *A Four-parameter Model of Glottal Flow.* STL-QPSR 4/1985, 1-13.

ISHIZAKA K. & FLANAGAN J.L. (1972) *Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords.* B.S.T.J., 51, 1233-1268.

KELLY J.L. & LOCHBAUM C.C. (1962) *Speech Synthesis* Proc. Stokholm-Speech Communications Seminar - R.I.T. 127-130. and 4th Int. Congr. Acoust., G42.

MAEDA S. (1990) *Compensatory articulation during speech : evidence from the analysis and synthesis of vocal tract shapes using an articulatory model.* Speech Production and Speech Modelling, W.J. Hardcastle & A. Marchal eds., 131-149.

RUMELHART D.E., HINTON G.E. & WILLIAMS R.J. (1986) *Learning Internal Representation by Error Propagation.* in Parallel Distribued Processing : Exploration in the Microstructure of cognition. by RUMELHART D.E. & McCLELLAND J.L. (eds.), 318-362.
Cambridge, MA : MIT Press.

TRINH VAN L., GUERIN B. & CASTELLI E. (1991) *Source-Tract Coupling and the Subglottal System in an Articulatory Synthesizer.* Eurospeech 91, Genova, vol. 1, 267-270.