CONTINUOUS VARIATION OF THE VOCAL TRACT LENGTH IN A KELLY-LOCHBAUM TYPE SPEECH PRODUCTION MODEL.

Hui Yi WU   Pierre BADIN   Yan Ming CHENG   Bernard GUERIN

Laboratoire de la Communication Parlée (ICP, UA CNRS 368)
E.N.S.E.R.G. - I.N.P.G.
46, av. Félix Viallet - 38031 GRENOBLE Cédex, FRANCE.

ABSTRACT

The KELLY-LOCHBAUM reflexion-type line analog (K-L model) is a temporal speech production model which has the advantages of a low computational cost and of a simple and clear physical interpretation. But it has an important drawback : it is not designed to handle a continuous variation of the vocal tract length. In this paper we present a strategy to solve this problem : the vocal tract length variation is dealt with as a variation of the working sampling frequency and then this variable sampling frequency is converted into a constant one.

The sampling frequency conversion is achieved by means of a time-varying FIR filter, designed to minimize the computational cost. The performances of the algorithm are evaluated with simple sinewave signals and with synthetic vowels. Finally, since for a practical application we should use frames within which the sampling frequency is constant, we extend our algorithm to a frame to frame basis and solve the problem of the FIR filter definition when moving across frame boundaries.

INTRODUCTION

In human phonation, the phonological informations are encoded by both vocal tract configuration and its dynamic variations. For synthesis purposes, a vocal tract configuration is described as an area function, including naturally the information of vocal tract length. Several models have been proposed for the acoustic simulation of the vocal tract in the time domain (/2/, /4/). For computational cost reasons, we use a KELLY-LOCHBAUM (K-L) reflection-type line analog to develop our research. Since the K-L model and even its recent developments (LILJENCRANTS, 1985) do not take into account the possibility of variation of the vocal tract length, we have made a attempt to develop this feature. In the first section of this paper, we present the basic idea for the spatially continuous length variation of the vocal tract, i.e. sampling frequency conversion, and we test the method. In the second section, we extend this algorithm to a frame to frame based temporal variation.

1. SPATIALLY CONTINUOUS VARIATION OF THE VOCAL TRACT LENGTH

1.1 The Problem

Since the vocal tract length vary rough between 16 and 19 cm during speech, we need include this feature in any vocal tract acoust simulation. For a K-L type of vocal tract tempor simulation (or improved versions, /3/), all t tubes have the same length (spatial sampling step and the sampling frequeney of the temporal sign produced is inversely proportional to this length. continuous variation of the vocal tract length c be achieved by a continuous variation of the tub length around a given value, which leads to related variation of signal sampling frequenc Therefore, if we wish a signal sampled with constant frequency, we need a system to convert t signal sampled with the variable frequency into signal sampled with the constant frequency. In th first section, we reduce the problem to the me conversion from a constant input sampling frequen F_i to a constant output sampling frequency F_o.

1.2 The Sampling Frequency Conversion

In a classical way (/1/), we decompose th sampling frequency conversion into two steps : fir we convert the discrete input signal x(n) sampl with F_i into a continuous signal $x_c(t)$, and then sample this continuous signal with F_c.

To reconstruct $x_c(t)$ from $x(n)$, we only nee to low-pass filter $x(n)$ with a cutoff frequency F_c equal to $F_i/2$. The theoretical formula o interpolation by an ideal low pass filter is :

$$y_c(t) = \frac{1}{F_i} . \sum_{n=-\infty}^{+\infty} x(n).h(t-nT_i) \quad (1$$

where

$$h(t) = 2.F_c.\frac{\sin 2\pi F_c t}{2\pi F_c t} \quad (2$$

is the impulse response of the filter, an $T_i = 1/F_i$. This impulse response beeing infinit (I.I.R.) and non causal, we need to approximate th filter by a F.I.R. (Finite Impulse Response) filte by using a windowing function w(t). This leads to a certain distortion of the frequency response of the filter.

In the second step, we need to sample y(t) at F_o, and thus, to avoid aliasing we need to insure that F_c is lower both than $F_i/2$ and than $F_o/2$. Then we obtain the formula :

$$y(m)=\frac{2F_c}{F_i}. \sum_{n=N1}^{N2} x(n).w(mT_o-nT_i).\frac{\sin(2\pi F_c.(mT_o-nT_i))}{2\pi F_c.(mT_o-nT_i)} \quad (3)$$

where w(t) is the windowing function, and N1 and N2 are determined as functions of F_i and F_o, and of the length of the window. Finally, the global system described by eq. (3) is a time-varying low-pass digital filter (/1/), implemented as a F.I.R. filter. The method of windowing the impulse response of an I.I.R. filter for F.I.R. filter design provides the avantage that the F.I.R. length, and thus the computational cost of the filter, can be easily and independently varied. At the same time, it makes it easy to compute the coefficients of the filter for each frame.

We know that the type and the length of the window used influences the properties of the filter. Therefore we need to evaluate quantitatively this influence.

1.3 Evaluation of the Transformation

In order to evaluate the performance of the sampling frequency conversion, we have made tests with sinewaves of different frequencies, and with synthetic vowels generated by our K-L line analog.

Sinewaves analysis

The influence of the transformation on a single sinewave has been analyzed : for different fundamental frequencies, two sinewaves with the same fundamental frequency, amplitude and phase, S_{Fi}, sampled with the system input frequency F_i, and S_{Fo}, sampled with the system output sampling frequency Fo have been generated. Then S_{Fi} has been converted into S'_{Fo} by the system, and finally the following parameters have been compared for S_{Fo} and S'_{Fo} : (1) the difference of amplitude between the sinewaves, (2) the difference of phase, and (3) the Signal/Distortion (S/D) ratio.

Because of the nature of the low-pass filter, an undulation is introduced in the pass band of the filter transfer function : it is always smaller than ±1 dB, which can be considered negligeable. Since the window we use is symetric around the origin point, the impulse response is symetric and thus a linear phase filter is insured : the transformation has no effect on the signal waveshape.

As expected, the S/D ratio increases with the window length. An informal analysis (by visual inspection of the FFT spectrum of S'_{Fo}) has shown that the distortion is mainly due to harmonic components corresponding to frequencies such as $F_0 + n.(F_o-F_i)$ or $F_0 + n.(F_o-F_i)/2$, where F_0 is the frequency of the sinevawe, and that the non-correlated noise is very much below this distortion. Therefore the S/D ratio is defined as the ratio between the energy measured in a 300 Hz band centered on the sinewave fondamental frequency and the energy outside this band (up to 5 kHz). Fig. 1 shows the evolution of the S/D ratio as a function of the number of points for the window, for a rectangular and for a Hamming window, for two different sampling frequency conversions. For short windows (i.e. 4-5 points), there is less scattering in the S/D ratio for a rectangular window than for a Hamming window, and for longer windows, the opposite phenomenon happens : we conclude that rectangular windows lead to better results than Hamming windows for short windows, and that Hamming windows give better results for longer windows.

Synthetic vowel analysis

The transformation has also been tested with vowels synthesized with our K-L model. The signals for the synthetic vowels [a], [i] and [u] have been converted into signals sampled with various frequencies ; the corresponding spectra (obtained by the Cepstrum method) have been compared with the original spectra visually and by means of a "distance" defined by :

$$D = \sum_{N=1}^{1024} \frac{A_{dB}(N\Delta F) - A_{dBref}(N\Delta F)}{1024} \quad (4)$$

where the missing points of $A_{dB}(N\Delta F)$ are evaluated by linear interpolation (since the frequency steps for the two spectra are different, due to different sampling frequencies). On the curves (see example in Fig .2) we can see that the system retains the formant characteristics very well, the errors appearing mainly in the spectrum valleys.

The error measured by eq. (4) converges toward a non zero value when the window length increases, depending on the vowel configuration and on the sampling frequency change. Since we know that for a long window the error must be very small, we conclude that this bias is due to our "distance" and to the linear interpolation, and we normalize the results in relation to this convergence value for each case. Fig. 3 shows that finally, there is no obvious difference between a rectangular and a Hamming window. In every case, there is a rather abrupt decrease of the scattering of the normalized error for windows longer than 4 points : we conclude that a rectangular window with 5 points is optimal.

2. FRAME TO FRAME TEMPORAL VARIATION OF THE VOCAL TRACT LENGTH

2.1 The problem

In real speech, the length of the vocal tract varies continuously with time, for instance in transitions from rounded vowels to unrounded ones. Following a classical approximation, we suppose that the vocal tract area function is constant during a short time interval, and thus its length. This is the basis for a frame by frame simulation of the vocal tract in most of the models. Thus, we should apply the frequency conversion developped in the previous section on a frame to frame basis : the input sampling frequency F_i must be considered constant for each frame, but may vary from one frame to the next one, according to the length variation of the vocal tract. This leads to the problem of the realization of the filter defined by eq. (3) when the window w(t) overlaps the boundary between two frames.

2.2 The solution

We first try to solve theoretically the problem for a system with only two input sampling frequencies. We suppose that the input signal is given with a sampling frequency F_{i1} from $-\infty$ to 0, and with a sampling frequency F_{i2} from 0 to $+\infty$. Thus, we can suppose that the corresponding

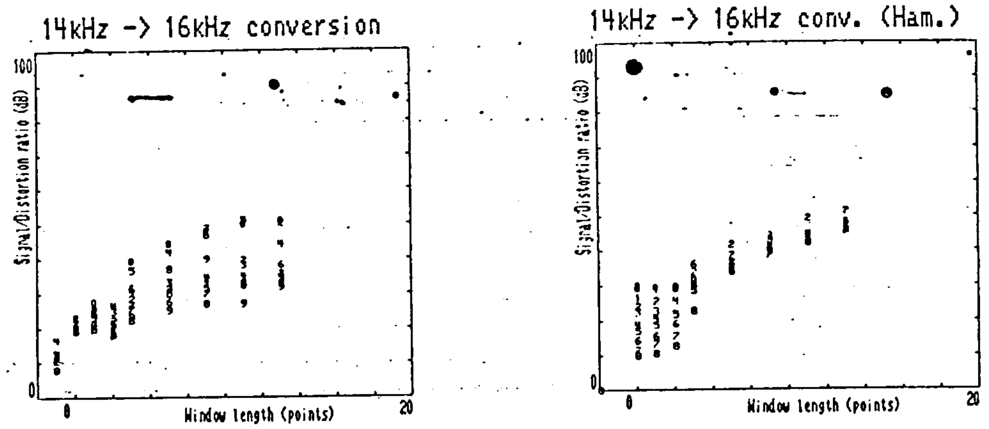

Fig.1 . Signal/Distortion ratio against window length (expressed as a number of points) for sinewaves wi frequencies ranging from 50 Hz (symbol 1) to 4.5 kHz (symbol 9) by 500 Hz steps (Ham. = Hamming windowing, otherwise rectangular windowing).

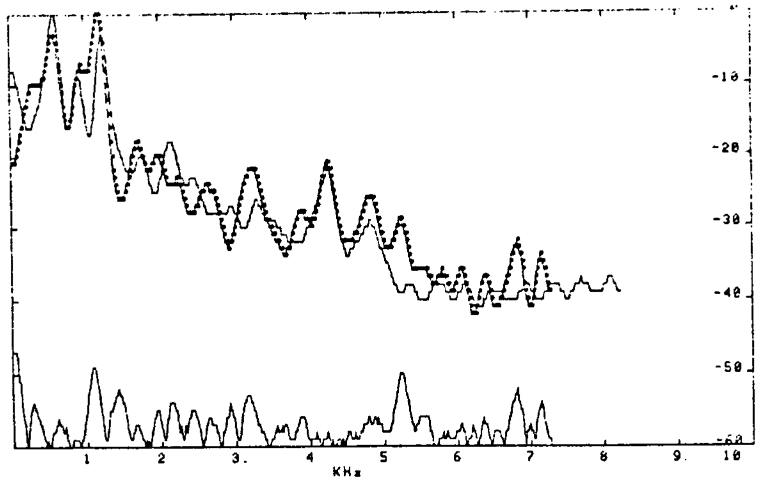

Fig.2 : Spectra of a synthetic vowel /a/ (continuous line, F_i=16.55kHz), of the signal resulting from th transformation (dotted line, F_o=14.55kHz), and difference of the spectra (bottom line).

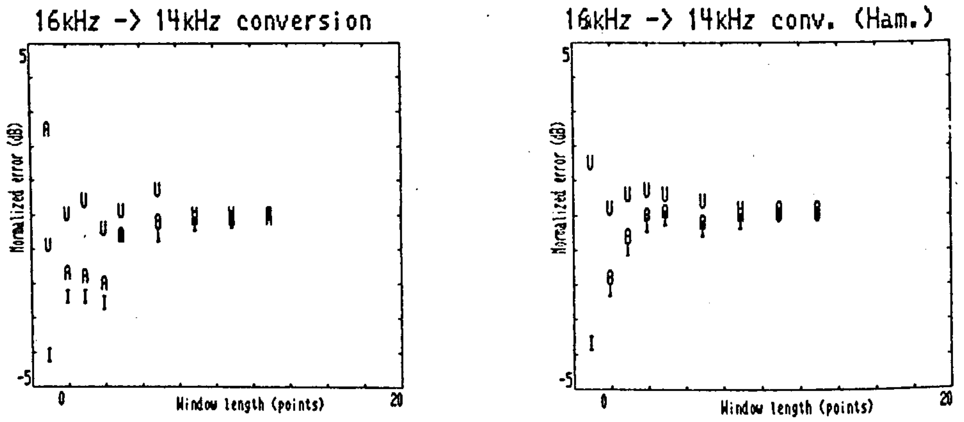

Fig.3 : Spectral error against window length (expressed as a number of points) for 4 vowels (Ham. = Hamming windowing, otherwise rectangular windowing).

continuous input signal $x_c(t)$ has been decomposed, by an appropriate-step windowing into two signals $x_{c1}(t)$ and $x_{c1}(t)$ defined by :

for $-\infty < t < 0$, $x_{c1}(t) = x_c(t)$ and $x_{c2}(t) = 0$,
for $0 \leqslant t < +\infty$, $x_{c1}(t) = 0$ and $x_{c2}(t) = x_c(t)$.

Then, we suppose that $x_{c1}(t)$and $x_c2(t)$ are sampled respectively with F_{i1} and F_{i2} into $x_1(n)$ and $x_2(n)$. Thus, we can apply eq. (3) to these two discrete signals, with the corresponding $2.F_c/F_{i1}$ and $2.F_c/F_{i2}$ factors,. and then sum up the results to reconstruct the complete signal, knowing that the filtering is a linear operation. In the case where the window is entirely included in one of the two frames only, eq. (3) applies directly. Otherwise, taking into account the instants where x_1 and x_2 are zero, we obtain the following equation :

$$y(m)=\frac{2F_c}{F_{i1}}.\sum_{n=N1}^{N0} x(n).w(mT_o-nT_{i1}).\frac{\sin(2\pi F_c.(mT_o-nT_{i1}))}{2\pi F_c.(mT_o-nT_{i1})}$$
$$+\frac{2F_c}{F_{i2}}.\sum_{n=N0+1}^{N2} x(n).w(mT_o-nT_{i2}).\frac{\sin(2\pi F_c.(mT_o-nT_{i2}))}{2\pi F_c.(mT_o-nT_{i2})} \quad (5)$$

where N0 is the last sample of the first frame and N0+1 the first sample of the second frame.

We see that eq. (5) means that for output samples close to the frame boundary, it is just needed to take into account the contributions from the samples in the two frames with the suitable coefficients. Nevertheless, we understand that the abrupt step windowing defined above will introduce some spectral distorsion in the two filters, which leads practically to some distortion for the signal near the boundary. We can anyhow check that eq. (5) reduces to (3) in the case where $F_{i1} = F_{i2} = F_i$, which could be expected.

The problem beeing solved for one boundary, we can extend the method to the frame to frame basis mentioned above, as far as the window length is shorter than the frame length, to avoid to include two boundaries in the same window. We have chosen to keep the cutoff frequency F_c of the interpolation filter independant of the input sampling sampling frequencies : we take the half of the smallest of all the input and output frequencies.

2.3 Evaluation of the method

In this section, we give describe the tests that we used in order to check the validity of the algorithm.

Triangle waveform

In order to evaluate the distortion of the signal at the frame boundaries, we have generated constant frequency triangle waveforms with sampling frequencies varying from one frame to the next one. Then, we have applied our frequency conversion algorithm to obtain a signal with a constant sampling frequency. Thus, whenever a segment of straigth line crosses a frame boundary, it is easy to measure the departure from linearity. We have made this evaluation directly on the processed signal, and also on its second derivative which can be expected to be an series of Dirac impulses corresponding to the inversions of slope in the triangle waveform : when the signal departs from a straigth line, the second derivative is not null any longer and shows up as a noise.

In different experiments, we have shown that the amplitude of the distortion at the boundary :
(1) depends very little on the window length (the length of the segment where distortion appears is longer for longer windows) ;
(2) is roughly proportional to the amplitude of the signal at the boundary ;
(3) is proportional to the sampling frequency variation between the two frames ;
(4) is roughly independant of the input and output sampling frequencies.

In the reality, the vocal tract length does not vary quickly : thus the sampling frequency variation from one frame to the next one never exceeds one or two percent. For these type of variation, the departure from linearity is less than one percent.

Vocalic transitions

Finally, we have elaborated a few vocalic transitions such as [i] → [u], [a] → [u]. The signal produced is high quality, and it is impossible to detect any boundary problems either by listening or by visual inspection of the signal.

We conclude that our algorithm is well suited to our practical speech application.

CONCLUSION

We have shown that it is possible to solve the problem of the spatially continuous variation of the length of the vocal tract by a sampling frequency conversion method. This method leads to good results even with rather short windows (4-5 points). It has been succesfully extended to a simulation based on a frame to frame decomposition. Thus our K-L model is not limited any longer by a constant vocal tract length. The study has been done for a "quasi-static" model : now it is needed to extend our algorithm to a fully dynamic model.

BIBLIOGRAPHY

/1/ CROCHIERE R.E. & RABINER L.R. (1983), "Multirate Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey.
/2/ KELLY J.L. & LOCHBAUM C.C. (1962), "Speech Synthesis", 4th Int. Congr. Acoust., G42.
/3/ LILJENCRANTS J. (1985), "Speech Synthesis with a Reflexion-Type Line Analog", Doctoral dissertation, R.I.T., Stockholm.
/4/ MAEDA S. (1982), "A Digital Simulation Method of the Vocal-Tract System", Speech Comm. 1, 199-229.