

VOCAL TRACT SIMULATION :  
IMPLEMENTATION OF CONTINUOUS VARIATIONS OF THE LENGTH IN A KELLY-LOCHBAUM MODEL,  
EFFECTS OF AREA FUNCTION SPATIAL SAMPLING

H.Y. WU P. BADIN Y.M. CHENG B. GUERIN

INSTITUT DE LA COMMUNICATION PARLEE (U.A. CNRS 368)  
L.C.P.-E.N.S.E.R.G. - I.N.P.G.  
46, av. Félix VIALLET - 38031 GRENOBLE Cédex, FRANCE

### ABSTRACT

The shape of the vocal tract and its dynamic transitions convey an essential part of the phonetic information in speech. Therefore the way to represent this shape is important for the quality of speech coding or speech synthesis. In this paper, we first study the influence of the spatial sampling of the area function upon the acoustic characteristics of the vocal tract, and determine an optimal sampling step minimizing the computational cost without perceptual degradation. In a second part, we deal with the problem of continuous variation of the vocal tract length for a KELLY-LOCHBAUM reflexion line analog simulation : we propose a method of temporal sampling frequency conversion.

### INTRODUCTION

The transformation of an idea into speech signal is realized through a set of semantic, syntactic, phonological, articulatory and acoustic operations. The production of the speech signal itself starting from the vocal tract shape is the last step in the chain : at that level, the information is coded as quasi static shapes of the vocal tract as well as dynamic transitions. Thus, for obtaining a good quality synthetic speech, it is important to have a good description of the area function. In this paper, we first attempt to assess the effects of the spatial sampling of the area function upon the corresponding acoustic characteristics, and we try to determine the optimal spatial sampling step for the area function. In a second part, we study the problem of vocal tract length continuous variation in the frame of a KELLY-LOCHBAUM (K-L) reflexion line analog simulation and derive a sampling frequency conversion method.

## 1. AREA FUNCTION SAMPLING EFFECTS

### 1.1 The Problem

In speech production simulations, the vocal tract is conventionally approached by a cascade of cylindrical tubes : the area function is then defined by the length of these tubes and their cross-sectional areas. In the frame of a K-L simulation [6] [7], we are bound to use the same length for all the tubes : this is a kind of spatial sampling of the area function. The problem is to study the effects of the sampling step (the common length of the cylindrical tubes) upon the

corresponding acoustic properties (e.g. formants and bandwidths), and to determine an optimal step to get the lowest computational cost without any perceptible degradation of the acoustic properties of the signal.

### 1.2 Tools, Method and Results

To assess the effects of the spatial sampling, we have defined as references 4 vowel and 9 fricative configurations (from [4]), of which area functions are sampled every 1 mm. Then, we have computed for each reference configuration, new area functions with a sampling step increasing from 2 to 30 mm.

This undersampling is realized by a transformation under two constraints : (1) keeping the total length of the vocal tract constant, (2) keeping the local volume constant, i.e. the volume of each new cylindrical tube is equal to the volume defined by the same coordinates along the vocal tract midline as in the reference tract.

The undersampled vocal tract configurations are then used to compute the transfer functions (between the lips output flow and the glottis input flow for vowels, or the constriction input pressure for fricatives) and their decompositions into poles and zeros, by means of a frequency domain vocal tract simulation [1]. In order to assess the errors induced by the undersampling, we determine the relative errors on the poles in relation to the reference values (see example on Fig.1), and the curve difference between the transfer function computed with the new spatial step and the reference transfer function (see example on Fig.2).

### 1.3 Discussion

#### Acoustic level

Generally, in the low frequency region, the transfer function is related to the global balance of the vocal tract main cavities [4] [2], whereas the high frequency region is more specifically defined by the geometrical details of the area function. More precisely, it is established (cf FANT's nomograms, [4], and STEVENS & HOUSE's model, [9]) that the characteristics (position, cross-sectional area, length) of the main constriction in the vocal tract are fundamental for the acoustic characterization of a given sound. The general effect of increasing the spatial sampling step is to decrease the precision of the vocal tract description. The errors on the main cavities modify rather the high frequencies (this phenomenon is clear in the case of our vowels), whereas the errors on the constriction have also consequences on the

low frequencies (the fricative consonants are more sensitive to this phenomenon). We have found that the errors due to undersampling are generally larger for the configurations having a smaller cross-sectional area at the constriction : when using STEVENS & HOUSE's model of vocalic configurations generation for the four vowels /a/, /i/, /u/ and /o/, an arbitrary decrease of the constriction area leads to an increase of the error on the formants. In the same way, the errors are larger for the high than for the low vowels. This can be explained by a gradual opening or smoothing of the constriction related to undersampling. This opening is very likely the cause of the roughly monotonous evolution of the errors on given formants : for instance, the formants "more associated" with "HELMHOLTZ resonator" type resonances, as F1 for /a/ and /i/, and F2 for /u/ and /o/ increases when the constriction area increases with the spatial step.

### Perceptual level

Our aim being to decrease the computational cost for the synthesis without any perceptible degradation of the output, the problem is to evaluate the spatial step corresponding to the Just Noticeable Difference (JND) between the original signal and the modified one. Since there are no JND studies about the perceptual effects of vocal tract undersampling, and since it is known that a slight change of formant frequency is more important than that of bandwidth or amplitude, we refer, in a first step, to two basic papers on the Difference Limens\* (DL) formant frequencies : FLANAGAN [5] has studied the DLs for vowel formant frequencies in isolated context, and MERMERLSTEIN [8] in CVC context.

Table 1 gives the values for the spatial sampling step corresponding to the DL for formant frequencies for static vowels in isolated context. We conclude that a 8 mm step leads to errors not perceptible according to FLANAGAN's DLs. Since these DLs are average values around which we must take scattering into consideration, since our reference configurations are also average configurations, and since we have neglected other variations (e.g. bandwidths and amplitudes), we must consider an optimal step smaller than 8 mm, in order to insure errors always below the perceptual level.

However we think that FLANAGAN's results represent the absolute lower limit for the DL, and that MERMERLSTEIN's ones in CVC context are closer to reality and continuous speech : from MERMERLSTEIN's data (cf also Table 1) we derive a 25 mm optimal step. But for the same reasons mentioned just above, we should consider practically smaller values.

Since there are no JND or DL studies about fricative consonants, we have applied FLANAGAN's DLs to the determination of the optimal sampling step for our fricatives. Table 1 gives the results. For /f/ and /f,/ there are no prominent poles in the frequency range of our analysis (0-5 kHz), and thus the method can not be used ; for /tʃ/, there is only one pole around 2.5 kHz. For /ʃ/, the low formants are rather sensitive to undersampling, because of

\* In general, the DL studies correspond to the restriction of JND studies to the controlled variation of only one specific parameter at a time, for instance a formant frequency or a formant amplitude.

the rapid opening of the constriction at the teeth level. For the whole set of the fricatives, if we do not take into account /ʃ/, the JND is reached for a step of 10 mm. In the case of the fricatives, our approach must be considered with caution since there is no evidence that the same formant errors on fricatives and vowels should be perceived in the same way : perceptual experiments are needed to check this first estimation.

### 1.5 Conclusion

In the frame of this study limited to a constant spatial sampling step (to be able to use a K-L type model), we have measured in an objective way the effects of the spatial sampling of the area function upon the acoustic properties of the vocal tract. From previous studies on DLs, we have made a first approach for an optimal sampling step in different situations : 8 mm for stationary vowels in isolated context, 25 mm for vowels in CVC context, and 10 mm for voiceless fricative consonants. We have seen that these values must be considered with caution : to determine a more precise and reliable optimal step, perceptual tests with sounds synthesized from the undersampled configurations should be realized first in a static context, and further in a dynamic context.

## 2. CONTINUOUS VARIATION OF THE VOCAL TRACT LENGTH

### 2.1 The Problem

Since the vocal tract length vary roughly between 16 and 19 cm during speech, we need to include this feature in any vocal tract acoustic simulation. For a K-L type of vocal tract temporal simulation (or improved versions, [7]), all the tubes have the same length (spatial sampling step), and the sampling frequency of the temporal signal produced is inversely proportional to this length. A continuous variation of the vocal tract length can be achieved by a continuous variation of the tube-length around a given value, which leads to a related variation of signal sampling frequency. Therefore, if we wish a signal sampled with a constant frequency, we need a system to convert the signal sampled with the variable frequency into a signal sampled with the constant frequency.

### 2.2 The Sampling Frequency Conversion

The sampling frequency conversion is realized by a time-varying low-pass digital filter [3], implemented as a F.I.R. (Finite Impulse Response) filter. The method of windowing the impulse response of an I.I.R. (Infinite Impulse Response) filter for F.I.R. filter design provides the advantage that the F.I.R. length, and thus the computational cost of the filter, can be easily and independently varied. Therefore, our time varying low-pass filter is obtained by windowing the I.I.R. of an ideal low-pass filter. The expression for the filter, including sampling with the new frequency, is :

$$y(m) = \sum_{N_1}^{N_2} x(n) \cdot W(mT_0 - nT_i) \frac{\sin\{\pi(mT_0 - nT_i)/T_i\}}{\pi(mT_0 - nT_i)/T_i} \quad (1)$$

where  $x(n)$  is the input signal sampled with the frequency  $f_i = 1/T_i$ ,  $y(m)$  is the output signal sampled with the frequency  $f_0 = 1/T_0$ ,  $W(t)$  is the windowing function, and  $N_1$  and  $N_2$  are determined as functions of  $f_i$  and  $f_0$ , and of the length of the window.

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.

### 2.3 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 reflexion 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  $f_o$  have been generated. Then  $S_{fi}$  has been converted into  $S'_{fo}$  by the system, and finally the following parameters have been compared for  $S_{fi}$  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  $\pm 1$  dB, which can be considered negligible. Since the window we use is symmetric around the origin point, the impulse response is symmetric 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 sine wave, 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 sine wave fundamental frequency and the energy outside this band (up to 5 kHz). Fig.3 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 :

$$D = \sum_{N=1}^{1024} \frac{| \text{AdB}(N\Delta f) - \text{AdBref}(N\Delta f) |}{1024} \quad (2)$$

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.4) 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. (2) 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.5 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.

In real speech, temporal continuous variation of vocal tract length usually happens. In a first attempt, we have not taken into account the dynamic effects of vocal tract length variation : we have restricted ourselves to a "quasi-static" model. Then we have extended this algorithm to frame-to-frame variations of the vocal tract length by solving the problem at the boundaries between two consecutive frames : this is not discussed here due to the page limitation. The temporal continuous variation of the vocal tract length can be approached by taking a frame width small enough.

### 2.4 Conclusion

We have shown that it is possible to solve the problem of the 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). This method has been tested for a "quasi-static" model : now it is needed to extend this algorithm to a fully dynamic model.

### BIBLIOGRAPHY

- [1] BADIN P. & FANT G. (1984), "Notes on Vocal Tract Computation", STL-QPSR 2-3/1984, 53-108.
- [2] BOE L.J. & ABRY C. (1986), "Nomogrammes et Systèmes Vocaliques", 15<sup>èmes</sup> JEP GALF, 303-306.
- [3] CROCHIERE R.E. & RABINER L.R. (1983), "Multirate Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey.
- [4] FANT G. (1960), "Acoustic Theory of Speech Production", Mouton, ('S-Gravenhague).
- [5] FLANAGAN J.L. (1955), "A Difference Limen for Vowel Formant Frequency", J. Acoust. Soc. Amer. 27, 288-291.
- [6] KELLY J.L. & LOCHBAUM C.C. (1962), "Speech Synthesis", 4th Int. Congr. Acoust., G42.
- [7] LILJENCRAFTS J. (1985), "Speech Synthesis with a Reflexion-Type Line Analog", Doctoral dissertation, R.I.T., Stockholm.
- [8] MERMELSTEIN P. (1978), "Difference Limens for Formant Frequencies of Steady-State and Consonant-Bound Vowels", J. Soc. Acoust. Amer. 63, 572-580.
- [9] STEVENS K.N. & HOUSE A.S. (1955), "Development of a Quantitative Description of Vowel Articulation", J. Acoust. Soc. Amer. 27, 484-498.

	V		CVC	
	F1	F2	F1	F2
/a/	30	15	30	30
/i/	20	08	30	25
/u/	08	10	30	XX
/o/	20	15	25	30
/f/	5	3		
/f/	10	20		
/x/	10	30		
/f/	XX	XX		
/f/	XX	XX		
/t/	XX	10		
/s/	10	15		
/s/	15	25		
/j/	20	15		

Reference DL for V context  
(FLANAGAN)

F1 : 12 Hz (F1=300),  
26 Hz (F1=500-700),  
F2 : 20 Hz (F2=1000),  
45 Hz (F2=1500),  
20 Hz (F2=2000).

Ref. DL for CVC context  
(MERMELSTEIN)

F1 : 49 Hz (F1=300),  
70 Hz (F1=600),  
F2 : 171 Hz (F2=2100),  
186 Hz (F2=1780).

Table 1 : Spatial sampling step leading to a given relative error corresponding to FLANAGAN et MERMELSTEIN's DLs on formant frequencies.

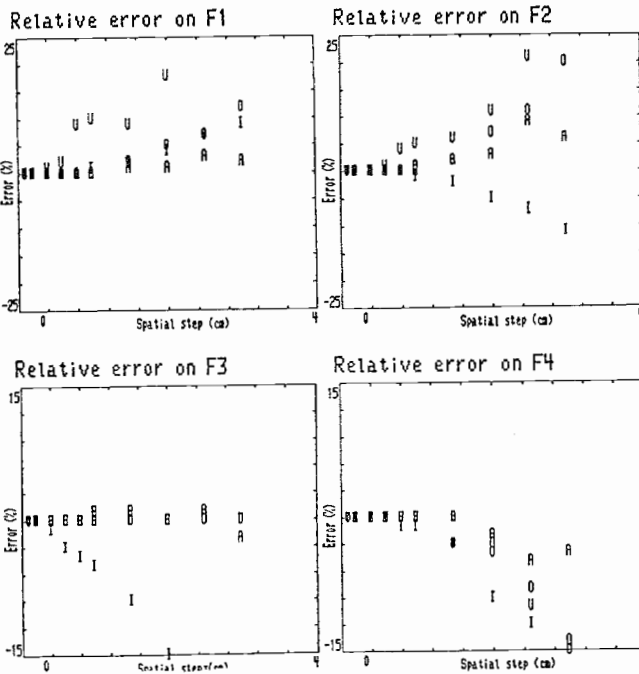


Fig.1 : Relative errors on formant frequencies for 4 vowels against spatial step.

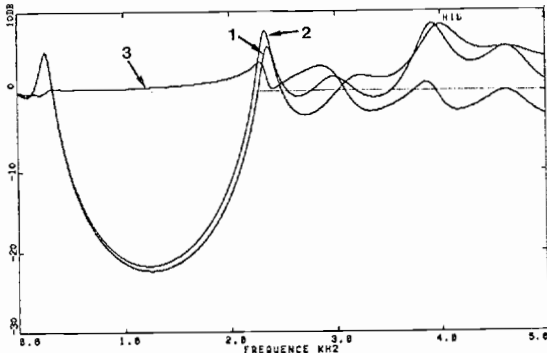


Fig.2 : Curve difference (3) between the transfer function for the reference configuration /i/ (1) and the transfer function for the /i/ with a 10 mm step (2).

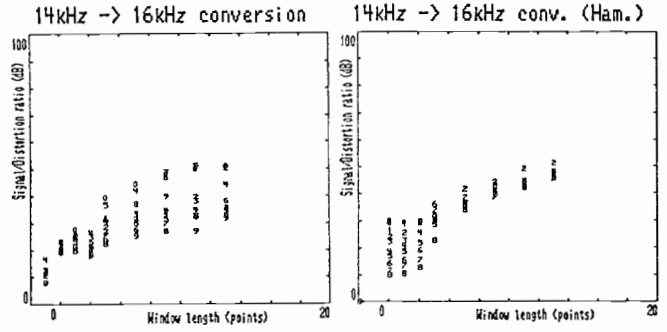


Fig.3 : Signal/Distortion ratio against window length (expressed as a number of points) for sinewaves with frequencies ranging from 50 Hz (symbol 1) to 4.5 kHz (symbol 9) by 500 Hz steps (Ham. = Hamming windowing, otherwise rectangular windowing).

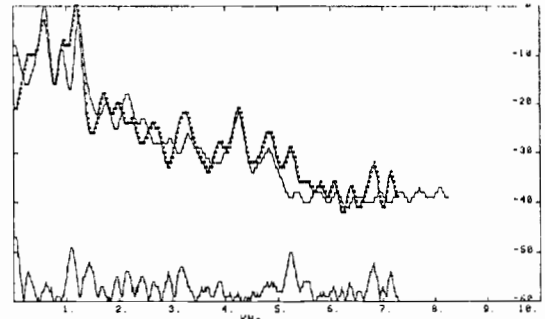


Fig.4 : Spectra of a synthetic vowel /a/ (continuous line, fi=16.55kHz), of the signal resulting from the transformation (dotted line, fo=14.55kHz), and difference of the spectra (bottom line).

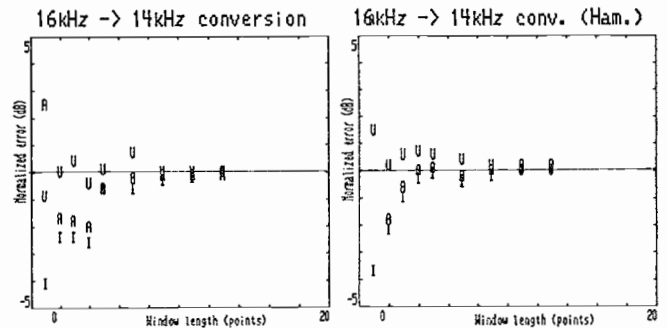


Fig.5 : Spectral error against window length (expressed as a number of points) for 4 vowels (Ham. = Hamming windowing, otherwise rectangular windowing).