MULTIPULSE LPC MODELING OF ARTICULATORY MOVEMENTS: DETERMINATION OF MINIMUM PULSE SEQUENCES

Soumya BOUABANA and Shinji MAEDA
ENST - CNRS URA-820 - 46 Rue Barrault 75634 Paris 13, FRANCE
e-mail : soumya@sig.enst.fr

ABSTRACT

The frame-by-frame variations of tongue profiles derived from X-ray film data are described in terms of the temporal patterns of four articulatory parameters. The temporal variation of each parameter, i.e., movement, is assumed to be the output of a time-invariant auto-regressive filter. These filters are excited by a sequence of pulses, representing articulatory commands. The curve of synthesis error for each movement shows a rapid decrease up to the number of pulses corresponding to that of the syllables in the sentence and then the decreasing rate becomes distinctively slower. In this paper, the minimum number of pulses is determined by using acoustic criterion. It depends on the number of the phonetics features, in the sentence, of which their realization is related to particular parameters.

1. INTRODUCTION

Digitized lateral X-ray film data were used to monitor the temporal variation of tongue profiles in the mid-sagittal plane. The profiles are obtained by manually tracing, frame-by-frame, radio films shot at a rate of 50 frames per second during the production of 10 French sentences uttered by two female speakers. An articulatory model is derived as the result of a factor analysis on the measured tongue contours. In this model, the tongue profile is specified by one extrinsic parameter, jaw position **jw** (open/close), and three intrinsic parameters, tongue-body position **tp** (back/front), tongue-body shape **ts** (arched/flat) and tongue tip position **tt** (up/down). These four parameters suffice to specify the entire mid-sagittal tongue shape with reasonable accuracy, since they explain more than 90% of the variance of observed tongue profiles. The parameter values are calculated from each frame of the X-ray data.

The phonetics features of vowels F-patterns can be specified by the values of one or two dominant parameters. It appears, in a preliminary analysis of movements, that not the whole four parameters but only some selected parameters at time are involved in the production of a given phoneme. Our objective is to introduce a constraint for controlling individual articulator movements by using a simple source-model filter, in order to effectively describe the coordinated orchestration of the individual articulatory movements during sentences.

2. MODELING MOVEMENTS

Each articulatory movement is assumed to be the output of a time-invariant auto-regressive (LPC) filter. This hypotheses means that motor programs controlling muscular forces might be such as to produce a nearly constant stiffness condition because of the physical benefits of achieving movements which are both optimally smooth and energy efficient, as speculated by Nelson [1]. Actually, the comparison between the

optimum movements with respect to various physical performance constraints, such as energy and rate of change of acceleration (jerk), shows the remarkable similarity of movements predicted by the linear-spring invariant model and by performance with minimum-energy-cost constraint.

2.1. LPC analysis

The LPC analysis filter consists of two parts, the pre-emphasis and the filter to be identified. The role of the pre-emphasis, which is the first-order, is to flatten spectral tilt, in order to correctly identify the poles at high frequencies. An information criterion indicates a second order filter cascaded with first-order filter as optimum for modeling movements of the tongue parameters. A standard LPC technique is used to identify the values of filter coefficients [6]. The corresponding impulse responses including the de-emphasis filter exhibit highly da-mped characteristics with an effective duration of 140 to $200ms$ [2]. Because the command is represented by the train of pulses, these impulse responses behave as an elementary gesture.

2.2. Biomechanical interpretation of the filters

The biomechanical interpretation leads us to consider the second order filter as neuro-muscular system (a force generation) and the pre-emphasis filter as a passive mechanical system. This result is supported by the natural frequency calculations. The pole frequency of the identified second order system is about $3Hz$ [2]. However, the natural frequency of the human tissues seems to be much higher than $3Hz$. For example, the oscillation frequency of the lips during bilabial stops release is approximately 33 Hz [3]. The mechanical resonance of cheek tissues, measured for tensed or relaxed condition, was found in the frequency range of 30 to $60Hz$ [4]. These values are much higher than the calculated natural frequencies of the second order filter. Because in frequencies below the natural frequency of the second order system the motion is determined by the stiffness and resistance, the identified mechanical system would behave as a viscoelastic system corresponding to the first-order filter. Our filter approch corresponds to a Hill type muscular contraction model [5].

2.3. Multi-pulse synthesis

Each filter is excited by a sequence of pulses representing articulatory commands. The position and the amplitude of the excitation pulses are determined from the measured movements using an MLPC method (Multi-pulse LPC) proposed by Atal and Remde [7]. The accuracy of the MLPC synthesized movements monotonously improves with an increase in the number of excitation pulses [2]. So, the question we raise is how many pulses are needed to synthesize the observed articulatory movements. In the articulatory domain, it is difficult to establish a criterion for this. We, therefore, resort to an acoustic criterion.

3. ACOUSTICAL EFFECTS OF MOVEMENTS

We attempt to determine which parameters have dominant influence on the phonetics features of vowels in the sentence. We calculated the first four formants frequencies along 10 sentences from the measured articulatory movements as references, $\mathcal{F}_{ref}$. These references formants patterns are then compared with those calculated from only one parameter synthesized with the MLPC model, $\mathcal{F}_{syn,j}$, $j =$ **jw**, **tp**, **ts**, or **tt**. The values of the remaining six parameters are the measured ones. In these formant calculations, the measured lip movements are used.

The effect of each tongue parameter manifests when the number of pulses is equal to 0. In this paper, we present only $\mathcal{F}_{syn,jw}$ and $\mathcal{F}_{syn,tp}$ with $m = 0$, (*see* figure 1). The greatest distance error between $\mathcal{F}_{ref}$ (solid line) and $\mathcal{F}_{syn,j}$ (dashed line) patterns occurs with **tp** synthesized movement. The back/front feature is highly damage. The **jw** effect is more important for the close vowels than the open vowels. Except when the open vowel is located at the end of the sentence. The two patterns $\mathcal{F}_{syn,\mathbf{ts}}$ and $\mathcal{F}_{syn,\mathbf{tt}}$ are almost identical with $\mathcal{F}_{ref}$. In spite of the important articulatory activity of these two parameters, their contribution to the F-pattern vowels is small. These observations imply that the tongue articulatory parameters aren't controlled in the same importance. In the next section, the value of the minimum pulses m_j^* is determined for each parameter j.

4. MINIMUM NUMBER OF PULSES

The error between the first four formants frequencies of $\mathcal{F}_{ref}$ and $\mathcal{F}_{syn,j}$, $F(n), n = 1, ..., 4$, in terms of energy is calculated by

$$\mathcal{E}_{i,m,j} = \sum_{k=1}^{N}\sum_{n=1}^{4}(F_{i,m,j}(n,k) - F_i(n,k))^2,$$

where i is the sentence number, j is the synthesized parameter which depends on the number of the pulse m, and k is the frame number. An error tolerance, $\mathcal{E}_{i,m_j^*,j}$, is estimated from the difference limen (DL) of formant frequencies [8]. The value of m_j^* represents the number of pulses involved to synthesized movement j with the minimum accuracy necessary in synthesizing the perceptually acceptable formants patterns. This number is always less or equal to the syllable number in the sentence. Figure 2, illustrates one example for the sentence "une réponse ambiguë". The determined number of pulses depends on the type of parameters; thus, $m_{jw}^* = 5$, $m_{tp}^* = 6$, $m_{ts}^* = 3$ and $m_{tt}^* = 2$. Note that the number of syllables in this sentence is equal to 6. The observed and synthesized movements show large differences, especially for **tt** and **ts**. $\mathcal{F}_{syn}$ calculated with a small number of pulses exhibits the error tolerance less than DL. So, a great accuracy in synthesized movements isn't necessary to produce formants frequencies of vowels with reasonable accuracy. Otherwise, error in the articulation can be very high when the parameter isn't directly involving in the production. For example **ts** for the vowels $[aN,]$, **tt** for $[aN, i, y]$. The number m_j^* seems to be in relation with the number of the phonetics features inherent to the parameter.

5. CONCLUSION

The results show that the number of pulses necessary for each movement to synthesized formants frequencies of vowels is always less or equal to the number of syllables in the sentence. This is explained by the fact that the effective duration of the filter's impulse response is comparable to that of the average syllable duration in the 10 sentences, $180ms$. Moreover, it suggests that articulatory controls involve a unit of production whose length is about the size of syllable. Having the method that allows us to describe the relatively complex articulatory movements in terms of a small number of pulses, the problem now is to find out something about the nature of spatio-temporal organization, specifically the inter-articulator coordination during speech.

This work is supported by european project *Esprit/BRA, n° 6975, SPEECH MAPS*.

6. REFERENCES

[1] W.L. Nelson, Physical principles for economies of skilled movements. *Biol. Cyber.*, 46 (1983), pp. 135-147.

[2] S. Bouabana and S. Maeda, Effets acoustiques de la modlisation articulatoire. *Springer-Verlag*, 2nd Edition, (1994).

[3] O. Fujimura, Bilabial stop and nasal consonants : a motion picture study and its acoustical implications. *Journal of Speech and Hearing Research*, 4(3) (1961), pp. 233-247.

[4] K. Ishizaka, J. C. French and J. L. Flanagan, Direct determination of vocal tract wall impedance. *JIEEE. Trans. on ASSP*, 30(11) (1983), pp.828-832.

[5] A. V. Hill, The heat of shortening and the dynamic constants of muscle. *Proc. Roy. Soc. Brit.*, 19(6) (1974), pp. 136-136.

[6] F. Makhoul, A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, (1986) pp. 716-723.

[7] B.S. Atal and R. Remde, High-quality speech at low bit rates : Multi-pulse and stochastically excited linear predictive coders. *Proc. ICASSP*, (1986) pp. 614-617.

[8] J. Flanagan, Speech analysis synthesis and perception. *Springer-Verlag*, 2nd Edition, (1972).

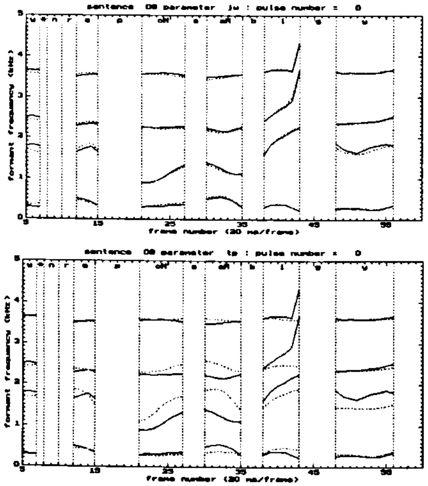

Figure 1: *Solid line: $\mathcal{F}_{ref}$. Dashed line: top, $\mathcal{F}_{syn,jw}$; bottom, $\mathcal{F}_{syn,tp}$, with $m =$* 0.

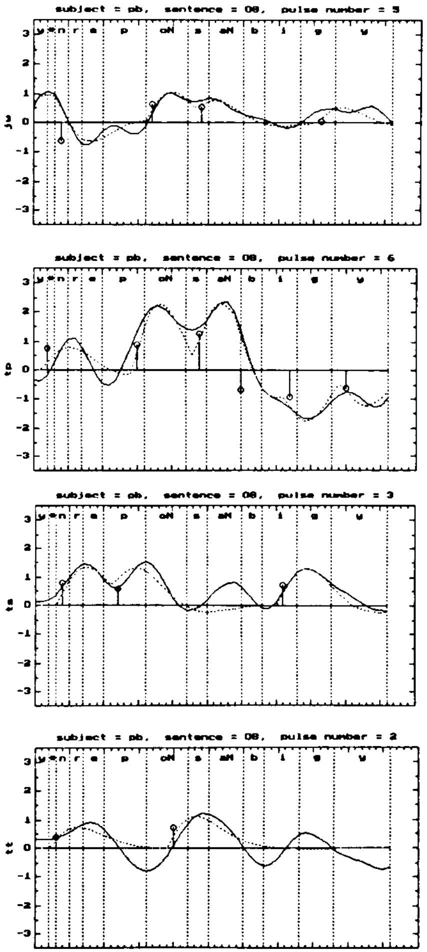

Figure 2: *observed (solid line) and synthesized (dashed line) movements with the minimum pulses for jw, tp, ts and tt parameters.*