Multi-pulse LPC modeling of articulatory movements

Soumya Bouabana, Shinji Maeda *

Ecole Nationale Supérieure des Télécommunications, Département Signal and Centre National de la Recherche Scientifique, URA 820, 46, rue Barrault, 75634 Paris Cedex 13, France

Received 7 May 1996; received in revised form 14 November 1997; accepted 25 February 1998

Abstract

The frame-by-frame variation of tongue profiles derived from X-ray film data is 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. Each filter is excited by a sequence of pulses, representing articulatory commands. The filter coefficients, and the position and amplitude of the pulses are determined by applying an MLPC method. The curve of synthesis error for each movement shows a rapid decrease up to a number of pulses corresponding to that of the syllables in the sentence and then the decreasing rate becomes distinctively slower suggesting the presence of syllable-size motor organization. The minimum number of pulses is determined by using an acoustic criterion. It depends on the number of the phonetic features, in the sentence, of which their realization is crucially related to their pertinent parameters. © 1998 Elsevier Science B.V. All rights reserved.

Zusammenfassung


Résumé

Les variations temporelles du profil de la langue enregistrées dans des images cinéradiographiques sont décrites par les mouvements de quatre paramètres articulatoires. La variation temporelle de chaque paramètre (mouvement) est supposée être la sortie d’un filtre auto-régressif invariant dans le temps. Chaque filtre est excité par une séquence d’impulsions qui représente la commande articulatoire. Les coefficients du filtre, la position et l’amplitude des impulsions sont déterminés par la méthode de la MLPC. Dans ce modèle, le geste syllabique apparaît comme l’unité de base dans l’organisation temporelle des mouvements de la langue lors de la production de la parole. Par ailleurs, l’erreur de synthèse décroit rapidement jusqu’au nombre d’impulsions de commande égal à celui des syllabes dans la phrase. Au delà de ce

*Corresponding author. E-mail: maeda@sig.enst.fr.

0167-6393/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PH: S0167-6393(98)00012-0
1. Introduction

Observed acoustic and articulatory phenomena reflect the biomechanical and physiological constraints of the complex articulatory system. Several investigators have recently proposed hypothetical principles for the articulatory coordination in speech production, particularly, how its temporal organization is designed, and what quantities resist different causes of variation.

For many years, the basic concept of temporal organization proposed by the majority of investigators has remained focused on concatenated series of stationary target gestures representing different units such as syllable and phonemes. Parallel to this point of view of phoneme concatenation, Kozhevnikov and Chistovich (after Kent and Minifie, 1977) introduced a sequential planning model based on a (C)V-type syllabic organizational unit. However, data on American English are not always compatible with the idea of a CV-type syllable as the basic unit of speech articulation. There are several indications that the anticipatory velar lowering for the production of a nasal consonant can extend over a vowel in the preceding syllable (Kent et al., 1974; Moll and Dainiolo, 1971; Krakow, 1993). In other words, the vowel in a syllable can be influenced by a consonant in the following syllable. In this sense, syllables cannot be a simple production unit. A theory of the motor control in speech proposed by Wickelgren (1969) is an interesting contrast to that of Kozhevnikov and Chistovich. He argued that speech units are coded allophonically as context-sensitive elementary motor responses. This approach has a serious shortcoming in that only the immediate context of a segment is taken into account. A look-ahead model in Henke’s dissertation work (Henke, 1966) explicitly pointed out anticipation of feature realization in an upcoming string of segments. This is related to the model of coarticulation by Öhman (1967), which suggests an explicit correspondence between distinctive features and different articulatory organs in their implementation. In his model, articulatory movements are specified by the superposition of a slow vowel-to-vowel transition and a rapid local perturbation of an articulator specific to the consonant production. More recently, Fujimura (1994, 1995) has developed the so-called CD (converter-distributor) model where the syllables act as the basic production units. Although the CD model incorporates the superposition principle employed by Öhman (1967), this model is built in a “generative” framework of phonology converting a phonological representation of speech into articulatory movements. For example, the consonant perturbation is specified by an impulse response and its magnitude is specified by the syllable strength in CD model.

In this paper, we attempt to model articulatory movements with the help of an MLPC (Multi-pulse Linear Predictive Coding) method proposed by Atal and Remde (1982). The objective of this paper is to demonstrate that observed articulatory movements can be very efficiently described by using a simple source-filter representation of the movement generation in connection with the multi-pulse excitation model. Since a time-invariant filter is used for a given articulator, its movements are specified only by the input excitation pulse train in which each pulse specifies the time and magnitude of the generated movement. We do not attempt here to implement a hierarchical command structure which can account for the superimposed vowel and consonant production as in the models of Öhman and Fujimura mentioned just above.

Instead we would like to show, using the well formulated technique in signal processing, that a train of sparsely distributed excitation pulses can specify complex articulatory movements and it...
can be automatically calculated in a straightforward manner. Our results indicate that a surprisingly small number of pulses per sentence is needed to adequately describe movements of each articulator. Such description, therefore, will be quite useful in the investigation of the coordinated orchestration of the individual articulators during sentence production. Since the procedure is automatic, it can handle a large body of data. In this sense, the technique we shall describe is suitable for the analysis of articulatory data acquired by such methods as the X-ray microbeam (Fujimura et al., 1973) and EMMA (Schönle et al., 1987; Perkell et al., 1992). With the large body of movement data described in terms of the compact excitation pulse form, it might be possible to statistically derive relational rules between the observed movements and the discrete phonological representation of speech.

2. Articulatory movement data

The movement data used in this study are derived from X-ray films of the vocal tract in the lateral view shot while two female speakers, PB and DF, read 10 short French sentences (Bothorel et al., 1986). Phonetic transcriptions of the 10 sentences are listed in Appendix A. The acoustic characteristics of speech depend on the shape of the entire vocal tract from the lips to the larynx. This simple fact motivates us to describe the tract shapes by an articulatory model with a small number of parameters. In this way, observed temporal variations of the entire vocal tract profile can be specified by the frame-by-frame variation of the articulatory parameters, which we call an articulatory movement. In the following, we shall briefly describe how such a model is formulated and how the value of articulatory parameters are determined from the X-ray film data. A more detailed description of our model and an improved version are found, respectively, in Maeda (1979, 1990, 1992) and in Gabioud (1994).

The vocal-tract profiles imaged on the X-ray films (with 50 frames per second) are manually traced frame-by-frame to obtain their contour descriptions, which are then digitized manually using a computer connected graphics tablet. About 500 frames, corresponding to 10 s worth of real speech were digitized in this way for each of the two speakers. The mid-sagittal vocal-tract shape in each frame is then measured by applying a semi-polar coordinate system which is fixed relative to the upper maxillary structure, as shown in Fig. 1(a). The coordinate grid lines are spaced by 0.5 cm in the two linear regions and by 11° in the polar region. The intersections between the tongue contour covering from the glottis to the apex and the coordinate grid lines, indicated by the circles in Fig. 1(a), are automatically detected by a binary search algorithm. A set of the coordinate values at the intersections from the glottis to the apex constitutes a vector representing the measured contour as shown in Fig. 1(c). The vector can be considered as the tongue shape sampled at the coordinate grid lines. The projection of the vector onto the semi-polar coordinate restores the mid-sagittal tongue contour.

It may be noted that the variables specifying the lip shape in Fig. 1(b) are also measured. But they are used only for acoustic calculations. We focus our attention on the movements of the tongue that is undoubtedly the most important articulator for speech.

In our approach tongue shapes are described by the vector having 28 elements. The values of its 28 elements are not independent, since they are constrained by the physical properties of the tissues, by the anatomical arrangement of the muscles, and so on. This implies that the observed 500 vectors would have some correlation structures. Using a factor analysis, such as a principal component analysis (PCA), we should be able to obtain a more concise specification of the tongue shape with a small number of orthogonal components. Such an analysis is not so interesting for us however, if the components cannot be interpreted in articulatory terms. For this purpose, we must have some idea about how an articulatory model should look like. In this regard, we follow a model proposed by Lindblom and Sundberg (1971), where tract shapes were determined as a function of parameters such as jaw, tongue-body and tongue-tip positions. The basic underlying hypothesis is that during speech production the complex
muscular activities of the vocal tract organs are organized into a small number of independently controllable but interlinked functional blocks. If this is the case, an appropriate factor analysis technique may be able to extract linear components which are related to these underlying functional blocks.

Since the mandible movement directly affects the tongue and lip positions, any realistic model must include jaw position as one of the parameters (Lindblom and Sundberg, 1971; Mermelstein, 1973). In our study, jaw position is measured in each data frame as the vertical distance between the tips of upper and lower front incisors. The measured jaw position is added to the corresponding tongue vector, which we shall denote \( y \). It is a standard practice in factor analyses to convert observation vectors into \( z \)-scores by the centering and normalization of each variable in \( y \) as follows:

\[
z_i = \frac{y_i - \bar{y}_i}{\sigma_i} \quad \text{or} \quad y_i = \sigma_i z_i + \bar{y}_i,
\]

where \( \bar{y}_i \) and \( \sigma_i \) are respectively the average value and standard deviation of the \( i \)th variable calculated over all observations of a given speaker. A factor analysis then determines a matrix of factor patterns (or loadings), \( A \). The \( z \)-scores are expressed as the linear sum of factor scores, \( x \), weighted by the factor patterns as follows:

\[
z = Ax.
\]

A factor score, i.e., an element of the vector \( x \), may be regarded as an articulator parameter, if its effects on the restored tongue shape correspond to those of a particular functional block described before.

The matrix \( A \) might be determined by a conventional PCA. The direct application of PCA showed that the first three components already explain more than 90% of the variance. We did not find, however, a component which could be interpreted to account for the influence of the jaw position upon observed tongue shapes. As explained before, it is not satisfactory to have an articulatory model without a separate jaw parameter. Therefore, we first calculate the correlation between the measured jaw position and each of the tongue variables, which determines the contributions of the jaw, i.e. the factor pattern for the jaw parameter, denoted henceforth \( \mathbf{jw} \). Note that we force the measured jaw position to play the role of one of the articulatory parameters. Second, the correlation structure due to \( \mathbf{jw} \) is subtracted from the covariances calculated over the observed \( z \)-scores.
The principal components then are determined from the residual covariances. We employed an algorithm, a so called arbitrary orthogonal factor analysis, formalized by Overall (1962) for these calculations, which ensures that the jaw component and subsequently determined principal components are orthogonal to each other.

The effects of each of the first four components upon tongue shapes are illustrated on the semi-polar coordinate as shown in Fig. 2 for the speaker PB. The three tongue contours in each figure are obtained by assigning the parameter values -3, 0 and 3 standard deviations. The middle contour in each figure corresponds to the zero value, which is the averaged tongue shape, $\bar{y}$, that is close to the neutral vocal tract. The model, in fact, describes the deviation of the tongue contour from its averaged profile. How and how much the tongue deviates due to an activation of a parameter are specified by its factor pattern, i.e., a column of parameters.

Fig. 2. The effects of the four articulatory parameters, $jw$ in (a), $tp$ in (b), $ts$ in (c) and $tt$ in (d), on the tongue shapes. The middle contour in each figure indicates the average tongue position and the two other contours correspond to the parameter value set at +3 and at −3 (in standard deviations).
extracted variance. One may notice, in Table 1, that the patterns of these extracted variances depend on the individual speakers. The speaker PB does not move the jaw very much suggested by the low extracted variance, 15.6% for $jw$. This speaker seems to use extensively the front–back tongue-body motion, indicated by the 45.8% of the variance explained with $tp$. The contribution of $jw$ and $tp$ is more even for the second speaker DF, 24% for $jw$ and 32.4% for $tp$. Interestingly, the sum of the variances of these first two components becomes much more similar for the two speakers, 61.4% for PB and 56.4% for DF. This suggests that the two parameters $tp$ and $jw$ could compensate each other, as demonstrated by Maeda (1990). Moreover, the variance explained by tongue tip position ($tt$) is the smallest for both speakers. The contribution of $tt$ is not preponderant to reconstruct the tongue profile as far as vowels are concerned.

Now, the matrix of factor patterns, $A$, in Eq. (2) is replaced by the truncated version, $\hat{A}$. The parameter values then are calculated from the measured z-score of each X-ray frame by

$$x = Bz,$$

where $B = \hat{A}(\hat{A}^T\hat{A})^{-1}$ (3)

and $B$ is called factor score coefficients. The superscripts, “$T$” and “$-1$” in Eq. (3) denote the transpose and the inverse of a matrix, respectively. The frame-by-frame variations of articulatory positions (i.e., the values of the four parameters, $jw$, $tp$, $ts$ and $tt$, that are the elements of $x$) can be regarded as articulatory movements. The units of parameters are standard deviations and the zero value corresponds to the mean for the total data domain of each articulator, which can be interpreted as its neutral position. Examples of calculated movements of the four articulators are shown by dashed lines in Fig. 3 for the sentence 08 “Une réponse ambiguë”. It may be noticed that the derived articulatory movements exhibit a considerable amount of jitter, which is a direct consequence of measurement noise. The jitter will affect the accuracy of the estimation of filter parameters in the following analyses. Therefore, the noise is filtered out before the movement modeling.

In order to obtain the clean movement traces, $s(n)$, out of noisy data $x(n)$, we extracted the signal

$A$, and the parameter value, respectively. Restored tongue contours, therefore, serve us to interpret the components in terms of the functional blocks of articulatory organs discussed earlier.

Fig. 2(a) shows the effects of the jaw parameter ($jw$). Since the measured jaw position is defined as the parameter, its interpretation is straightforward. The variation in the lower jaw position manifests as a high–low tongue movement. Notice that the jaw position affects cross-sectional areas not only in the buccal region but also in the pharyngeal region. The effect of the second component, shown in Fig. 2(b), seems to indicate a front–back tongue-body motion. We thus denote this parameter as tongue-body position, $tp$. The third component is shown in Fig. 2(c), where the deviation of the contours from the averaged position indicates the bulging or the flattening of the tongue-body. It specifies therefore the deformation of the tongue shape, $ts$. Maeda and Honda (1994) have demonstrated that $tp$ and $ts$ can be quantitatively related to electromyographic (EMG) activities of specific extrinsic tongue muscles. The fourth component represents the movement of the tongue tip as shown in Fig. 2(d). We name the fourth parameter as tongue tip position, $tt$. The higher order components, the fifth and ups, cannot be interpreted in articulatory terms and therefore they are simply discarded. Fortunately, the truncation of the higher order components will not degrade the modeling accuracy too much, since the jaw and first three principal components explain 91% of the variance.

The extracted variance by each of four components is summarized in Table 1 for the two speakers. It indicates that the four components suffice to specify the entire mid-sagittal tongue shape with a reasonable accuracy, since they explain in total 91% of the variance of observed tongue profiles for PB and 90.8% for DF.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>$jw$</th>
<th>$tp$</th>
<th>$ts$</th>
<th>$tt$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB</td>
<td>15.6</td>
<td>45.8</td>
<td>20.4</td>
<td>9.2</td>
<td>91</td>
</tr>
<tr>
<td>DF</td>
<td>24.0</td>
<td>32.4</td>
<td>23.4</td>
<td>11.0</td>
<td>90.8</td>
</tr>
</tbody>
</table>
subspace using an eigendecomposition procedure (Roy and Kailath, 1989), which is very similar to the PCA described in the preceding section. It turned out that a more familiar low-pass filtering resulted in almost identical results, except at the onset and offset of each utterance where an edge effect of filtering appears. This is due to the fact that our X-ray data are truncated utterance-by-utterance and always some degree of discontinuity occurs at the edges. The cleaned movement traces obtained by the eigendecomposition are shown by the solid lines in Fig. 3, where these curves are superimposed on the recorded noisy measurements. In the following section we shall use the cleaned movement data obtained by the eigendecomposition. Moreover, only movement data from the speaker PB are used, because the digitized X-ray tracings of DF lack the lip contours which are needed in the acoustic calculation described in Section 5.

3. Modeling movements

Biological and physiological systems are complex, being inherently nonlinear. The traditional way of handling such a system is to start with a simple model and then to gradually evolve to a more complex model as increased understanding is obtained and new questions are raised. Over the years, many models have been proposed to match the trajectory of biological movements.

![Fig. 3. Temporal patterns of the four tongue articulatory parameters (movements), from the top to the bottom, jaw position (jw), tongue-body position (tp), tongue-body shape (ts) and tongue tip (tt), in the sentence 08 “Une réponse ambiguë” read by the speaker PB. (Note that “oN” denotes the nasal vowel [5].) The dashed and solid ylines indicate, respectively, the original measured movements and the cleaned ones.](image-url)
They range from simple linear models with constant coefficients (e.g., Mannard and Stein, 1973) to higher order nonlinear models which incorporate the effects of a large number of physiological and biomechanical variables (e.g., Hatze, 1978). Each of these modeling approaches has its advantages and disadvantages, and the type of model used would depend strongly on the goal of the study (Winters and Stark, 1987). For the analysis of articulatory movements, which is the output of a neuromuscular control system, it may be appropriate to briefly describe the basic model of a muscular contraction.

3.1. Muscular contraction model

The structural and functional characteristics of striated muscle have been extensively studied. The experimental results have led many investigators to develop a visco-elastic model on the basis of the anatomical muscle structure and the sliding-filament theory as well as the simulations of resting and active muscles (Hill, 1938). Hill reported a muscle model with visco-elastic components as shown in Fig. 4.

Each component of the model in Fig. 4 corresponds to the lumped anatomical muscle structure (also in Akazawa et al., 1969). Note that $K_s$ corresponds to a series elastic component, $K_p$ to a parallel elastic component and a viscous component is presented by $B$. $C$ denotes an internal contractile force which is generated by the reactions between the filaments of the muscle. This basic model structure has been subsequently supported by a number of groups, both for isolated muscle and for muscle-joint systems (e.g., Winters and Stark, 1987). Ordinary differential equations are able to describe such system and then the linear system theory can be applied to solve the problem. We are aware of the fact that there exists another type of motor control model based upon the equilibrium point hypothesis (Asatryan and Feldman, 1965), which is not universally accepted, however. In fact, there seems to be no muscular control model accepted universally.

3.2. Time-invariant model

In this paper, we develop an approach based on a straightforward linear system theory. Each articulatory movement is assumed to be the output of an auto-regressive (LPC) filter. In this stage of study, a critical problem is to identify the variables that are controlled by the nervous system to affect changes in position of the articulators. If the stiffness of the muscles were controlled, then the filter must be adaptive. In fact, study of the tongue movements suggests that the transfer characteristics such as the stiffness of the mechanical system could vary as a function of magnitude and duration of movements (Sonoda and Kiritani, 1976; Ostry and Munhall, 1985; Smith et al., 1993). We assume, however, a time-invariant filter for modeling the tongue articulatory movements. Our assumption implies that motor programs controlling muscular forces might be such as to produce a nearly constant stiffness condition. A physical benefit of this hypothesis is that the movements are both optimally smooth and energy efficient, as suggested by Nelson (1983). Nelson has determined the optimal movements that satisfy different physical performance criterion, such as minimum energy and minimum jerk, without assuming an underlying physical device. Moreover, he demonstrated the remarkable similarity between movements generated by the time-invariant linear spring model and those determined with minimum-energy-cost and maximum smoothness constraints.
4. MLPC analysis

Let us assume that a command to each articulator is generated by its own motor control block and that the four articulators are controlled independently of each other. This implies that the movement of each articulator is represented by a single channel source-filter model. Fig. 5 shows a detail of the source-filter model. The source, corresponding to the command, is specified by a train of pulses. Each pulse specifies the onset time of a movement to occur and the strength of that movement. The LPC synthesis filter consists of two parts, the de-emphasis \( g \) and the auto-regressive \( h \) filters to be identified. The output corresponds to the observable movement of an articulatory parameter. We shall describe later why \( h \) corresponds to the transfer function from command to force and \( g \) to that from force to movement.

4.1. De-emphasis filter

In the synthesis model, the transfer function of the first order de-emphasis filter \( g \) is defined as follows:

\[
G(z) = \frac{1}{1 - az^{-1}}. \tag{4}
\]

The value of \( a \) is determined empirically so that the error of synthesized movements becomes small. Note that this value is not critical. Any value between 0.4 and 0.7 did not significantly alter the results. We have chosen \( a \) equal to 0.5 for the four articulatory parameters. In speech signal processing, a pre-emphasis (with the filter \( g^{-1} \)) in analysis is used to flatten the spectrum so that poles at high frequencies are properly identified. The motivation for using pre-emphasis and de-emphasis in our movement modeling is different. In a preliminary experiment, we observed that if pre-emphasis was not applied to the movements before the determination of an excitation pulse sequence, i.e., \( a = 0 \), then the number of pulses necessary to describe the movements becomes large. This is presumably due to the fact that a large number of pulses are needed to describe relatively stationary parts of movements with a good accuracy. A possible solution might be to differentiate movement signals before the determination of pulses, i.e., \( a = 1 \). A relatively smaller number of pulses suffice to describe the differentiated movement, i.e., a velocity pattern. At synthesis, however, we observed that a restored movement slowly drifted away from the observed movement toward the end of each sentence. The integration of the synthesized velocity pattern with \( a = 1 \) causes the drift, because it accumulates errors in velocity. As a compromise, we have, rather arbitrarily, chosen the value of \( a \) equal to 0.5, which seems to solve the problem of the excessively large number of pulses required and that of the drift in the synthesized movement.

4.2. Second order filter

A second order filter is used for \( h \), because the spectrum of movements often exhibits a single dominant peak. It can be specified by its transfer function:

\[
H(z) = \frac{1}{1 - \sum_{i=1}^{2} a_i z^{-i}}. \tag{5}
\]

The AR coefficients \( \{a_i, i = 1, 2\} \) are identified using the autocorrelation method in the standard LPC technique (e.g., Makhoul, 1975) for each of the ten sentences. The coefficients are converted into the real \( \text{Re} z_j \) and imaginary \( \text{Im} z_j \) values of the poles. The subscripts \( "j" \) and \( "l" \) denote respectively the parameter identifier (\( jw, tp, ts \) or \( tt \)) and the sentence number. The pairs of the poles derived from \( jw \) and \( tp \) movements are plotted on the \( z \)-plane as shown in Fig. 6 for the ten sentences. Poles exhibit fairly tight clusters.

These poles are then described in terms of their frequencies \( F_j \), bandwidths \( B_j \), and quality factor \( Q_j \), whose values are averaged over the ten sentences, using the following relations:

![Fig. 5. Synthesis model of articulatory movements.](image-url)
\[ F_j = \frac{1}{10} \sum_{i=1}^{10} \left( f_s/2\pi \right) \tan^{-1} \left( \text{Im} \ z_{j/i}/\text{Re} \ z_{j/i} \right), \]

\[ B_j = -\frac{1}{10} \sum_{i=1}^{10} \left( f_s/2\pi \right) \ln |z_{j/i}|, \quad (6) \]

\[ Q_j = \frac{1}{10} \sum_{i=1}^{10} \left( F_{j/i}/B_{j/i} \right), \]

where \( f_s \) is the sampling frequency which is equal to 50 Hz.

Table 2 lists calculated values and standard deviations of \( F_j, B_j \) and \( Q_j \) for each of the four “articulatory” filters. Although standard deviations in frequency and in bandwidth are not so small, we decided to use the averaged value of poles for each articulator regardless of utterances for the sake of simplicity and of coherence with our time-invariant filter hypothesis. We therefore employ the strictly time-invariant filter to determine the position and amplitude of excitation pulses for each articulator.

It is worthwhile to mention that the order of filter \( (p=2) \) was validated using Akaike’s Information Criterion (AIC) which is defined as follows (Kay and Marple, 1981):

\[ \text{AIC}(p) = \log \left( \frac{V_p^2}{R_0} \right) + 2p/kN, \quad (7) \]

where \( N, V_p^2 \) and \( R_0 \) are respectively the number of frames (or samples) in a sentence, the LMS error obtained for \( p \) poles and the signal variance. The constant \( k \) equals to 0.4, which is valid for Hamming windowed signals. The calculations of AIC on each of the four parameters for the 10 individual sentences indicated that the minimum always occurs at \( p=2 \), as shown in Fig. 7 as an example. This means that the best “precision/complexity” is achieved with \( p=2 \).

4.3. Impulse response

The impulse responses of the filter with the pole positions averaged over ten utterances are shown...
in Fig. 8 for the two dominant parameters, \( jw \) and \( tp \). These impulse responses are regarded as an elementary gesture excited by a single pulse command. The response curves tend to exhibit a sinusoidal oscillation with high dissipation, especially in the case of \( jw \). Because of the high dumping factor, the major effect of the filter excitation occurs during its first half cycle with the duration 140–200 ms depending on individual articulatory parameters. Note that the effective duration of excitation is on the average about 180 ms, which roughly corresponds to the syllable duration. The first peak in the response curve can be interpreted as the target, which would manifest as a turning point in the observed articulatory movements.

4.4. Biomechanical interpretation of the filters

In our movement modeling, each pulse of the command sequence activates a muscular contraction source that generates force. This force then is applied to the muscle’s mechanical system (a visco-elastic system), in turn creating an observable articulatory movement. The conversion from command to movement, therefore, can be described by two distinct processes, the physiological process generating force and the mechanical process creating movement. The question is how the two filters, \( g \) and \( h \), for synthesis could be interpreted in physiological and biomechanical terms. It is interesting to note that a linear mass-spring system (corresponding to a second-order AR-filter) often underlies descriptions of articulatory movement data in the literature (e.g. Sonoda and Kiritani, 1976; Ostry and Munhall, 1985; Smith et al., 1993).

Yasuahara (1983), in analysing the hand writing system, postulated that a second-order filter represented the transfer characteristics of the force generation and a first-order filter corresponded to the mechanical hand system following the force...
generator in series. He identified the neuromuscular system to be a force generator having a higher $Q$ than the mechanical hand system. This was based on the assumption that the hand system could be regarded as a highly damped second-order process and approximated by a first-order filter. It is interesting to note that the frequency $F$, bandwidth $B$ and quality factor $Q$ of the hand neuromuscular system were respectively 3.4 Hz, 1.2 Hz and 2.8, which are comparable to the corresponding values of our second-order filter as shown in Table 2. In addition the real pole of the first-order filter $g$, which is about 0.7, is also close to our arbitrarily chosen value of $\alpha$, 0.5.

One might conjecture that the second-order system represents the natural frequency of the human tissues involved. However, since we identified the pole frequency of the second-order filter $h$ to be approximately 3 Hz (also in Bouabana and Maeda, 1994), this can be hardly the case. For example, the oscillation frequency of the lips during a bilabial stop release is approximately 33 Hz (Fujimura, 1961). The mechanical resonance of cheek tissues, measured for tensed or relaxed condition, was found in the frequency range from 30 to 60 Hz (Ishizaka et al., 1975). These values are much higher than the calculated natural frequencies of the second-order filter. Thus we favor Yasuhara’s identification and recognize the second-order filter $h$ as a representation of the force generation process. Incidentally, Bobet et al. (1993) have modelled force generation in skeletal muscle due to an electrical stimulation by a second-order system. Assuming that the natural frequencies of the mechanical structures that play a role in our problem are also well above 3 Hz, we may approximate the behavior of the mechanical system under consideration as visco-elastic (corresponding to our first-order filter $g$), since the motion of a second-order system below its natural frequency is primarily determined by the stiffness and resistance.

Before going to the next section, it may be interesting to note here about the nature of the pulse command. Obviously, there would not be such a single impulse command in the human motor system. Although we are well aware of the fact that our modeling is only an extremely simplified behavioral description of the immensely complex human speech motor control system, it is tempting to speculate that the impulse could be regarded as a code specifying the time instant and strength of a movement at the motor system in the brain. If this is the case, the transfer function represented by the filter $h$ covers characteristics of not only the physiology of muscular contraction force, but also the encoding and transmission of neuronal commands to the muscles.

4.5. Multi-pulse analysis and synthesis

After identifying the filter coefficients, we are ready to determine the locations and amplitudes of pulses representing articulatory commands for individual sentences. The excitation pulses are determined from the pre-emphasized movements using a multi-pulse LPC method (Atal and Remde, 1982; Atal, 1986). These authors have shown that the optimum position of a single pulse can be determined sequentially by a straightforward calculation, although a set of pulses cannot be determined at once. The algorithm, therefore, determines one pulse at a time from the movement signal. In order to determine the next pulse, a signal is synthesized using the just determined pulse as the input to the filter, $h$. The contribution of the first pulse to the observed signal is eliminated by subtraction of the synthesized signal from the observed one. The next pulse is then determined on the “residual” signal. This procedure is repeated until the residual becomes sufficiently small.

Let $E_k$ be the squared error between a signal, $s(n)$, which can be a measured pre-emphasized movement or the residual from the previous step, and the synthesized signal with a pulse located at the $k$th sample, $\hat{s}(n)^{(k)}$, which is defined as follows:

$$E_k = \sum_{n=1}^{N} (s(n) - \hat{s}(n)^{(k)})^2,$$

where $N$ denotes the number of samples corresponding to that of X-ray frames in a sentence. The synthesized signal with the single pulse excitation of the filter $h$ is expressed as the convolution of the input pulse $v(n)$ with the impulse response function of the system $h(n)$,
\[ \hat{s}(n)^{(k)} = \sum_{i=1}^{n} h(n-i)v(i), \quad 1 \leq n \leq N, \]  
(9)

and

\[ v(n) = \gamma_k e_n^{(k)}, \]  
(10)

where \( \gamma_k \) indicates the pulse amplitude and \( e_n^{(k)} \) an impulse with unit amplitude which occurs at the \( k \)th sample. The synthesized signal \( \hat{s}(n)^{(k)} \) is nothing but a shifted impulse response function scaled by \( \gamma_k \). The optimum position of \( k \) and amplitude \( \gamma_k \) that minimize \( E_k \) can be determined by letting

\[ \frac{\partial E_k}{\partial \gamma_k} = 0, \]  
(11)

which leads to

\[ \gamma_k = \frac{\sum_{i=1}^{N} s(i)h(i-k)}{\sum_{i=1}^{N} h(i)^2}, \]  
(12)

and then

\[ E_k = \sum_{i=1}^{N} s(i)^2 - \left[ \frac{\sum_{i=1}^{N} s(i)h(i-k)}{\sum_{i=1}^{N} h(i)^2} \right]^2. \]  
(13)

The optimum pulse location \( k_o \) can be determined by finding the maximum of the second term on the right side of Eq. (13). Notice that the numerator of the second term is a cross-correlation function between the signal and the impulse response of \( h \), and the denominator is constant. The desired pulse location therefore corresponds to the absolute maximum of the cross-correlation function. Moreover, the optimum amplitude of the excitation pulse corresponds to the normalized cross-correlation at \( k_o \) as indicated by Eq. (12).

Now a residual signal is computed by subtracting out the contribution of the just determined pulse. Then the location and amplitude of the next pulse is determined by minimizing \( E_k \) on the residual signal. The process of locating new pulses to reduce the mean-square error is continued until the error is reduced to an acceptable limit. In the actual calculation, the pulse amplitudes are jointly updated at each step by solving a set of linear equations derived from Eq. (8) with known \( k \) pulses’ positions, which increases considerably the accuracy of the synthesized movement.

The accuracy of the MLPC synthesized movements monotonously improves with an increase in the number of excitation pulses (Bouabana and Maeda, 1994). We calculated the synthesis error as a function of the number of pulses \( m \) for a given sentence as

\[ e(m) = \frac{\sum_{n=1}^{N} (s(n) - \hat{s}^m(n))^2}{\sum_{n=1}^{N} s(n)^2}, \]  
(14)

where \( \hat{s}^m(n) \) indicates the synthesized signal with \( m \) excitation pulses. Note that the error is normalized such that it becomes one when there is no excitation, i.e., \( m = 0 \). Fig. 9 illustrates an example of the calculated error in the movement of the four articulators. It is seen that the error rapidly decreases with an increase in the number of pulses. When this number becomes about equal to the number of syllables in the sentence (five in this case), the rate of decrease becomes much less after that point. This “saturation” point depends on the particular articulator; \( m = 6 \) for \( \text{jw} \) and for \( \text{tp} \), \( m = 5 \) for \( \text{ts} \) and \( m = 4 \) for \( \text{tt} \). This saturation phenomenon is observed also for most of the articulatory movements in other sentences.

Although the error curves tend to exhibit saturation, it is difficult to determine exactly how many pulses are needed to adequately synthesize the observed articulatory movements. In the articulatory domain, it is not obvious how to establish a criterion for this. We, therefore, resort to an acoustic criterion.

5. Acoustical effects of movements

Before describing the determination of the minimum number of pulses necessary, let us examine which parameters have a dominant influence on the acoustic patterns of vowels in the sentence \( l \). Our strategy is to calculate the frequencies of the first four formants, \( F_n \), where \( n = 1, \ldots, 4 \), along each of 10 sentences from the measured articulatory movements. The set of formant frequencies calculated in this way serves as a reference and is denoted as \( F_{\text{ref}} \). These reference formant patterns are then compared with those calculated from only one parameter synthesized with the MLPC model, \( F_{\text{syn,j}} \), where \( j = \text{jw}, \text{tp}, \text{ts} \) or \( \text{tt} \). The values of the remaining three parameters are set to measured
ones. Moreover, in all following formant calculations, the measured lip and laryngeal movements are used.

Formant frequencies are calculated with the following sequence of procedures. First, a set of observed or synthesized movements, which are nothing but frame-by-frame variations of the four tongue articulatory parameters, determine the vector representation of a tongue shape. Second, the complete midsagittal vocal-tract profile of each frame is restored on the semi-polar coordinate system by the projection of tongue vector and other measured vectors representing the lips and the exterior vocal tract walls. Third, the vocal tract area function is derived using a power law $A = x^\beta$, where $A$ and $x$ are respectively the cross-sectional area in cm$^2$ and the corresponding midsagittal dimension in cm, that is calculated from the projected vocal-tract profile (Heinz and Stevens, 1964).

The values of the coefficients, $\alpha$ and $\beta$, vary along the length of vocal tract specific to a given speaker. The values of these coefficients were empirically determined in our study. Finally, the transfer function and then formant frequencies are calculated from the area function using a one-dimensional simulation of an acoustic tube (Maeda, 1982). The acoustic simulation includes the effect of yielding vocal-tract walls and the radiation load at the lip opening.

It must be noted here that we only calculate formant frequencies of oral and nasal vowel segments in a sentence. In the calculation of nasal vowels, the nasal coupling is not considered. We feel that this simplification is not so unreasonable since we compare formant frequencies from measured movements and those from synthesized movements without the nasal coupling in both cases. Consonant segments are not considered however. In order to compute the transfer function of a consonant, we need to know the location of the sound source in the length of the vocal tract. The source location can be estimated, more or less, from the constriction location and shape. But the acoustic calculation requires a relatively high accuracy concerning the geometry of the constriction. We judge that due to the measurement noise mentioned ear-

---

Fig. 9. Normalized synthesis errors calculated for the movements of the four articulators, $\text{jw}$, $\text{tp}$, $\text{ts}$ and $\text{tt}$, in a function of the number of pulses for the sentence 09.
lier, our X-ray data do not always have an accurate enough specification of the vocal tract configuration, especially of the constriction for consonants. For these reasons, we deal only with the oral and nasal vowel segments in the following acoustic error analysis.

The acoustic effect of each tongue parameter becomes evident when the value of a selected parameter is kept at zero by letting $m = 0$, i.e., without any excitation and thus without any contribution of that parameter on tongue shapes. In Fig. 10, we present formant patterns $F_{\text{syn,jw}}$ calculated with zeroed jaw parameter, $F_{\text{syn,tp}}$ with zeroed tongue-body position parameter, $F_{\text{syn,ts}}$ with zeroed tongue-body shape parameter and $F_{\text{syn,tt}}$ with zeroed tongue tip parameter for the sentence 08 “Une réponse ambiguë”. In each figure, the solid line shows the reference pattern $F_{\text{ref}}$ and the dotted line indicates the $F_{\text{syn,j}}$ pattern. The greatest discrepancy between $F_{\text{ref}}$ and $F_{\text{syn,j}}$ occurs with zeroed tp movement. The back-front feature is considerably damaged as indicated by the noticeable discrepancy in the second formant frequencies. The effect of jw is less important than that of tp. For the other two parameters, $F_{\text{syn,ts}}$ and $F_{\text{syn,tt}}$ are almost identical to $F_{\text{ref}}$. In spite of the relatively important movements of these articulatory parameters, their contribution to the formant frequencies of vowels appears to be insignificant.

In order to quantify these observations, we calculated errors which are defined as the difference between the formant frequencies of the reference pattern $F_{\text{ref}}$ and the synthesized one, $F_{\text{syn,j}}$. First, we converted the formant frequencies $F_n$ in Hz into the Bark scale, $Z_n$, using the formula proposed by Schroeder et al. (1979) as follows:

$$Z_n = 7 \ln \left\{ \frac{F_n}{650} + \left[ \frac{F_n}{650} \right]^2 + 1 \right\}^{1/2} \text{ (Bark).}$$

(15)

Fig. 10. Reference formant patterns ($F_{\text{ref}}$) in solid lines and synthesized formant patterns ($F_{\text{syn,j}}$) in the dotted lines. Synthetic versions, jw at (a), tp at (b), ts at (c) and tt at (d), are calculated by setting the value of the corresponding parameter to zero, i.e., $m = 0$. Note that relatively large formant differences seen at tp indicate an importance of this parameter on vowel formants.
We feel that the use of Bark scale in acoustic error calculations takes into account, to a certain degree, the perceptual significance of the error in formant frequencies. A global acoustic error calculated for the sentence $l$ where the movement of parameter $j$ is synthesized with the excitation of $m$ pulses is defined by the following equation:

$$E_{l,j,m} = \sum_{k=1}^{N} \sum_{n=1}^{4} \left( Z_{n,l,j,m}^{\text{syn}}(k) - Z_{n,l}^{\text{ref}}(k) \right)^2,$$

(16)

where $k$ denotes the frame number. This error is, then, normalized with the maximal possible error $E_l$ that is obtained by letting the values of four parameters equal to zero as follows:

$$E_l = \sum_{k=1}^{N} \sum_{n=1}^{4} \left( Z_{n,l,0}^{\text{syn}}(k) - Z_{n,l}^{\text{ref}}(k) \right)^2.$$

(17)

The normalized global error is then calculated by the following equation and illustrated in Fig. 11:

$$e_{l,j,m} = \frac{E_{l,j,m}}{E_l}.$$  

(18)

The calculated error curves confirm that not all articulatory parameters have an equally important effect on the acoustic output. In Section 6, we shall determine the minimum number of pulses $m_j^*$ for each four articulatory parameters $j$.

6. Minimum number of pulses

The difference limen (DL) for formant frequencies can be used to determine the minimum number of pulses necessary to specify adequately formant patterns of a given sentence. The DL of $F_1$ and $F_2$ frequencies varies at the vicinity of 5% depending on formant and frequency (Flanagan, 1955). The value of DL is determined for stationary vowels, whereas, in our case, vowels are not al-

Fig. 11. Normalized formant frequency errors defined by Eq. (18) plotted as a function of the number of pulses for the four articulatory parameters, $jw$, $tp$, $ts$ and $tt$. 
ways stationary and they spread over a sentence. We, therefore, formulate a global difference between reference formant \((F_1\) or \(F_2\)) calculated from measured articulatory movements and the corresponding formant calculated by replacing the movement of a selected parameter, \(j\), by synthetic one with the varying number of pulses, \(m_j\). The minimum number of pulses, \(m_j^*\) is determined such that the global difference becomes less or equal to the DL of the formant frequency. The global difference is calculated by selecting all frames which corresponds to vowels in the sentence \(l\). We denote the selected frames by the index \(k\) having a range from 1 to \(k_v\). Then the global difference less or equal to DL is expressed by

\[
\text{DL}_m \geq 100 \frac{\sum_{k=1}^{k_v} |F_{n,l}(k) - F_{n,l,m}(k)|}{\sum_{k=1}^{k_v} F_{n,l}(k)} \quad (\%)
\]

The minimum value of \(m_j^*\) is the smallest number of pulses \(m_j\) that satisfies Eq. (19) for \(F_1\) and \(F_2\) simultaneously. Table 3 lists the determined values of \(m_j^*\) of the four articulators, \(jw, tp, ts\), and \(tt\), and the number of syllables, \(m_{syll}\), for the individual sentences. The minimum number of pulses for \(jw\), \(m_{jw}^*\) is often equal to the number of syllables and less for three cases. An exception occurs for the sentence 28 where \(m_{jw}^*\) is greater than \(m_{syll}\). For the tongue-body position parameter, \(tp\), that is most active in the vowel production for the speaker BP, \(m_{tp}^*\) are slightly greater than \(m_{syll}\) for six cases. We have suggested before that the number of pulses necessary to describe the observed articulatory movements roughly corresponds to the number of syllables in the sentence. The application of the DL as criterion, therefore, did neither eliminate nor add many pulses, which means the observed movements of \(jw\) and \(tp\) are acoustically relevant to the production of vowels. The situation is not the same for the remaining two parameters, \(ts\) and \(tt\), however. The values of \(m_{ts}^*\) and \(m_{tt}^*\) are always much less than \(m_{syll}\). In the extreme case, i.e., the sentence 18, no pulse is needed for these two articulators. The requirement of a much smaller number of pulses, which is less than that of syllables, indicates that \(ts\) and \(tt\) participate less than \(jw\) and \(tp\) in the vowel production. Presumably, they play an important role in the production of the consonants and to a certain extent, the onset and offset parts of vowel segments. In order to make this point clearer, let us show an example.

Fig. 12 illustrates the example for the sentence “Louis pense à ça” (the sentence 09). The determined number of pulses depends on the type of parameters as \(m_{jw}^* = 5, m_{tp}^* = 6, m_{ts}^* = 1\) and \(m_{tt}^* = 2\). Note that the number of syllables, \(m_{syll}\), in this sentence is equal to five. The observed and synthesized articulatory movements show large differences, especially for \(tt\) and \(ts\), whereas the formant frequencies calculated from the four synthesized parameters exhibit only small differences, as shown in Fig. 13. These two parameters are not involved in the production of these vowels. Errors in the articulation can be very high when the parameter does not directly contribute to the realization of the phonetic value \((F_1\) and \(F_2\)) of the vowel. In other words, the number \(m_j^*\) seems to be in a direct relation with the number of the phonetic features inherent to the parameter.

It is interesting to note that the pulse train determined for \(tp\) in Fig. 12 indicates two different types of idiosyncrasies specific to the multi-pulse excitation associated with the time-invariant filter. The first type is clearly seen at the first two pulses occurred during the first syllable. The corresponding \(tp\) movement exhibits a broad negative peak. The two pulses are necessary to adequately reproduce that movement. This is due to the fact that the impulse response of our non-adaptive time-invariant filter cannot cope with the broad peak by excitation of a single pulse. This is in contrast with the last strong negative \(jw\) peak and with the first

<table>
<thead>
<tr>
<th>Sentence</th>
<th>(m_{jw}^*)</th>
<th>(m_{tp}^*)</th>
<th>(m_{ts}^*)</th>
<th>(m_{tt}^*)</th>
<th>(m_{syll})</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>02</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>03</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>08</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>09</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>28</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Fig. 12. Measured (in solid lines) and synthesized (in dotted lines) movements of the four articulatory parameters ($jw$, $tp$, $ts$ and $tt$) with the minimum number of pulses and the corresponding excitation pulse trains for the sentence 09 “Louis pense à ça”.

Fig. 13. Formant patterns calculated from the measured movements (solid lines) and those calculated from synthesized movements (dotted lines) with the minimum number of pulses, which are shown in Fig. 12.
positive ts peak seen in Fig. 12, which are nicely modeled by the single pulse excitation, although the synthesized movement is shifted by one frame toward the left. This example illustrates the fact that we gain simplicity with the time-invariant filter, but we sometime pay the price in terms of a complication of the excitation. Alternatively, one can state that the time invariant filter can always be used to model the movements because its inadequacy can be compensated by an increase in the number of pulses. The second type is observed again in the tp movement toward the end of the sentence exhibiting a plateau with some irregularity. The pulse excitation is not so efficient in describing such a plateau in the sense that the three additional pulses following a strong excitation for the nasal vowel [ŋ] are necessary. In other words, in order to maintain a steady level without any significant movement, a number of excitation pulses (movements) is required. If a rectangular time function, i.e., a step response function having a finite duration, were used for the excitation, a single excitation could suffice to describe the plateau.

7. Concluding remarks

In this paper, we used the MLPC signal processing technique to describe the tongue movements. In particular, we applied a source-filter model which corresponds to an AR time-invariant filter excited by a train of pulses. We recognize the impulse response of the synthesized filter representing a particular articulator as an elementary gesture. The observed temporal variation of the tongue is reconstructed by the superposition of these elementary gestures. In the articulatory domain, the number of pulses necessary to achieve a reasonable accuracy appears to be close to the number of syllables in a sentence for the four tongue parameters, since the error is always less than 30%. This could be explained by the fact that the effective duration of the filter’s impulse response is comparable to that of the averaged syllable duration of the ten sentences, 180 ms. The impulse response of identified filters exhibits a highly damped sinusoidal oscillation where the natural frequency is 3 Hz or less. Due to the high damping, the major movement occurs during the first half cycle having a duration of 167 ms or less, which is close to the average syllable duration (180 ms). If we think of the fact that most common syllables consist of at least two phonemes, a consonant and a vowel, it is most remarkable that in spite of relatively “sluggish” individual articulators with a slow response, the speech rate in terms of the number of phonemes per second can be maintained at least twice the rate expected on the basis of 83.5 ms per phoneme. If the articulators were synchronously activated, then the rate could be 167 ms per phoneme. Our data do not show such a synchronous activation; rather they show that the activation of individual articulators is asynchronous. For example as seen in Fig. 12, the identified pulses do not exhibit any synchronisation among different articulators. We speculate therefore that the speakers exploit this kind of asynchronous spatiotemporal organization to achieve the relatively high speaking rate despite of the slow individual articulators. This spatiotemporal organization then manifests as coarticulation. Liberman et al. (1967) coined the speech organization process as “high-speed performance with low-speed machinery”. Moreover, we suggest that the rate is not limited by the mechanical time-constant determined by the mass and stiffness of the muscular tissues, but rather by the “physiological” time-constant that governs the rate of the command generation, transmission and force generation with muscular contraction.

The search for the minimum number of necessary pulses with an acoustic criterion, i.e. the DL of formant frequency, suggests that it is not necessary to synthesize all the movements with a high accuracy during the vowel segments to obtain the corresponding formant patterns. When the articulatory parameter is not predominant to specify the vowel formant frequencies, its inhibition does not damage the acoustic results in our simulation experiment. What is the origin of this unnecessary movement? One possible reason is that not all articulatory movements are always controlled. The acoustically unnecessary movements are possibly generated by the effect of the physiological constraints which are related to the
tongue structure, for example, its elastic continuity. The movements then can result from an articulator passively following those of the primary articulators and/or by the effect of the neural connections in the central nervous system which control the coordination of the set of the articulators as a functional unit. In such circumstances, the passive articulator is not involved directly to produce the acoustic characteristics of the vowels. For this reason, it may be interesting to consider a multi-channel modeling, as opposed to a set of independent single channels as in the present study. The multi-channel modeling might be able to reduce a redundancy due to the interdependent organization of the articulatory movements during speech.

Admittedly, our pulse excitation model is not always efficient in the synthesis of the articulatory movements. Studies in behavioral aspects of motor control recognize two distinct classes of movements: one consists of movements whose duration is less than about 200 ms and the other longer than 200 ms (Schmidt, 1988). Although these studies were carried out for limb movements, the duration of 200 ms roughly coincides with that of syllables, 180 ms. When the movement time is short or fast, a feedback correction cannot be operated. The movements are ballistic. It appears then that our multi-pulse excitation, which is regarded as a pure ballistic model, should be effective in describing the fast movements associated with syllables. The measured articulatory movements indicates a mixture of fast and slow movements, however. A slow or steady portion of an articulatory movement can be much longer than a syllable when the two consecutive syllables have the same or similar vowel and the intervocalic consonant does not require an adjustment of the articulator in question, as “the look ahead principle” of Henke (1966) might predict. As discussed before, when the position of an articulator is maintained at a steady level, an excitation by a rectangular time function could be more appropriate than by a series of impulses. In the literature (e.g., Kiritani, 1986) the rectangular function is often used to model movements. It might be interesting, then, to consider a mixed excitation model consisting of pulse and rectangular commands.

Acknowledgements

We would like to thank the two referees, Bert Cranen and Osamu Fujimura, for their insightful remarks and constructive critiques on previous versions of this manuscript. In addition, their extensive editorial suggestions and textual corrections certainly made this paper more readable than otherwise. We are also grateful to the Journal Editor Christel Sorin for her patience and encouragement. This work was supported, in part, by European project Esprit/BRA, No. 6975, SPEECH MAPS.

Appendix A

Phonetic transcriptions of the 10 short French sentences: The sentence numbers correspond to the utterance indexes in the list used during the X-ray film recording.

01 Ma chemise est roussie /məʃmizəusii/
02 Voilà des bougies /vwələdebuʒi/
03 Donne un petit coup /dɔnəcjptku/
08 Une réponse ambiguë /ynrepɔsɔbɪgy/
09 Louis pense à ça /luipɔsasa/
13 Ce mignon bout de chou /səmiŋɔbudʃu/
15 Mets tes beaux habits /metəbozabi/
18 Prête-lui seize écus /pʁɛtlɥisəeky/
24 Chevalier du gué /ʃəvaljedyɡe/
28 Il fume son tabac /ilfymstaba/

References


