Speech Communication 8 (1989) 203–220
North-Holland

COMPARISON OF PARAMETER SETS FOR TEMPORAL DECOMPOSITION

Astrid M.L. VAN DIJK-KAPPERS
Institute for Perception Research, 5600 MB Eindhoven, The Netherlands

Received 26 September 1988
Revised 29 March 1989

Abstract. Temporal decomposition of a speech utterance results in a description of speech parameters in terms of overlapping target functions and associated target vectors. The former may correspond to articulatory gestures and the latter to ideal articulatory positions. Although developed for economical speech coding, this method also provides an interesting tool for deriving phonetic information from acoustic speech signals.

The speech parameters used by Atal (1983) in proposing this method were the log-area parameters. Our modified temporal decomposition method (Van Dijk-Kappers and Marcus, 1987, 1989) also works with log-area parameters as input. However, the method is not restricted to these; in principle, most commonly used parameter sets can be used. In this paper we compare the results obtained with nine different sets of speech parameters, including log-area parameters, formants, reflection coefficients and band-filter parameters.

The main criterion for good performance will be the correspondence between target functions and phonemes or subphonemes. The phonetic relevance of the target vectors will also be considered, but in less detail. Speech signal resynthesis supplies yet another criterion; for those parameter sets which are transformable into the same parameter space, a reconstruction error will be defined and evaluated.

From these experiments it can be concluded that log-area parameters form the most suitable parameter set available for temporal decomposition. In some respects band-filter parameters yield better results, but this set is not classified as the best due to properties related to resynthesis.

Zusammenfassung. Die temporäre Auflösung von Sprache führt zu einer Beschreibung von Sprachparametern durch sich überschneidende Targetfunktionen und assoziierte Targetvektoren. Die erstgenannten können Artikulationsbewegungen, die letzteren idealen Artikulationspositionen entsprechen. Obwohl diese Methode zur effizienten Sprachkodierung entwickelt wurde, bildet sie auch ein interessantes Hilfsmittel zur Ableitung von phonetischen Informationen aus dem akustischen Sprachsignal.

Die von Atal bei der Veröffentlichung seiner Methode (1983) verwendeten Sprachparameters waren log-area Parameter. Auch die von uns modifizierte Methode der temporären Auflösung (1987) benutzt log-area Parameter zur Eingabe. Allerdings beschränkt sich diese Methode nicht auf log-area Parameter; im Prinzip können die gebräuchlichen Parametergruppen verwendet werden. In dieser Studie vergleichen wir die Ergebnisse, die mit neun verschiedenen Parametergruppen, darunter log-area Parameter, Formanten, Reflexionskoeffizienten und band-filter Parameter erzielt wurden.

Das wichtigste Kriterium für ein gutes Ergebnis ist die Entsprechung der Targetfunktionen zu den Phonemen oder Subphonemen. Außerdem wird – allerdings weniger ausführlich – die phonetische Relevanz der Targetvektoren behandelt. Die Resynthese des Sprachsignals liefert ein weiteres Kriterium; sowohl Verständnisfehler als auch physikalische Fehler vermitteln Erkenntnisse über die Eignung der Parametergruppe für die temporäre Auflösung.

Aus diesen Versuchen kann gefolgert werden, daß die log-area Parameter die geeignetste Parametergruppe darstellen, die für die temporäre Auflösung zur Verfügung steht. In einigen Punkten wurden mit band-filter Parametern bessere Ergebnisse erzielt, jedoch wurde diese Gruppe aufgrund ihrer Resynthese-Eigenschaften nicht als die Beste eingestuft.

Résumé. La décomposition temporelle d'un message vocal fournit une description des paramètres vocaux sous la forme de fonctions-cibles se recouvrant et des vecteurs-cibles correspondants. Les premières peuvent correspondre à des gestes articulatoires et les seconds à des positions articulatoires idéales. Bien que développée pour un codage économique de la parole, cette méthode constitue également un outil intéressant pour extraire des informations phonétiques du signal vocal acoustique.

Les paramètres vocaux utilisés par Atal lorsqu'il a proposé cette méthode (1983) sont les paramètres "log-area". Notre méthode de décomposition temporelle modifiée (1987) utilise ces paramètres comme information d'entrée. Toutefois, la méthode ne se limite pas aux paramètres "log-area"; en principe, on peut également choisir les groupes de paramètres plus

courammeni utilisés. Dans ce présent article, nous comparons les résultats obtenus avec neuf groupes différents de paramètres vocaux, notamment les paramètres "log-area", les formants, les coefficients de réflexion et les paramètres de filtre de bande.

Le principal critère de qualité du résultat sera la correspondance entre les fonctions-cibles et les phonèmes ou sous-phonèmes. On considérera aussi l'importance phonétique des vecteurs-cibles, quoique de manière moins détaillée. La resynthèse du signal vocal fournira un autre critère; tant les erreurs de perception que les erreurs physiques fournissent des informations sur l'efficacité du groupe de paramètres pour la décomposition temporelle.

On peut conclure de ces expériences que les paramètres "log-area" constituent le groupe de paramètres le plus approprié pour la décomposition temporelle. Les paramètres de filtre de bande donnent de meilleurs résultats à certains égards, mais compte tenu des propriétés relatives à la resynthèse, ce groupe n'est pas considéré comme étant le meilleur.

Keywords. Temporal decomposition, parameter sets, target positions, articulatory gestures, speech analysis

1. Introduction

In articulatory phonetics, speech production is considered as a sequence of overlapping articulatory gestures, each of which may be thought of as a movement towards and away from an ideal, but often not reached, articulatory position. The sound which is produced by such an articulatory movement and which corresponds to a phoneme or subphoneme will be called a phone. It has long been assumed that such targets cannot be determined from the acoustic signal alone, detailed knowledge of the production of all the component phonemes being required before a speech signal can be decoded (Liberman et al., 1967). However, the so-called temporal decomposition method, proposed by Atal (1983) for economical speech coding, decomposes a speech signal into overlapping units, each of which is described by a target function and a target vector. Although no use is made of any explicit phonetic knowledge, our hope is that these units can be related to phones. Indeed, we have shown (Van Dijk-Kappers and Marcus, 1987, 1989) that with some necessary modifications and extensions, promising results can be obtained with this method. With a restricted database consisting of *CVC* combinations embedded in a neutral context, 74% of the phonemes were described by only one target function and the associated target vector. Furthermore, only 1% of the phonemes was missed. The remaining phonemes were described by 2 or more target functions, and this was possibly due to the fact that these phonemes were produced by two consecutive articulatory gestures and thus, in fact, consisted of two phones.

These results were obtained using our modified temporal decomposition method, which is more robust than Atal's original method with respect to its sensitivity to the variation of parameter values. With our method the values of these parameters, such as the length of the analysis window, are optimized. Up to now the input parameters have always been log-area parameters. Although these parameters, possibly due to their close relationship to the positions of the articulators, yielded the reasonably satisfying results mentioned above, it is not inconceivable that better candidates exist. Indeed, recent papers (Chollet et al., 1986; Ahlbom et al., 1987; Bimbot et al., 1987) report temporal decomposition results obtained with alternative parameter sets.

In this paper we will compare temporal decompositions using nine different sets of speech parameters. Except for the band-filter parameters, all of these parameters are LPC-derived and are often used for other purposes, such as speech coding and synthesis. The main criterion for good performance will be the correspondence of target functions to phones, since this gives a good indication of the phonetic relevance of the decomposition. In addition, the possible phonetic meaning of the associated target vectors will be considered for some of the parameter sets. Another criterion for good performance is supplied by the quality of the resynthesis after temporal decomposition. Both perceptual and reconstruction errors yield information about the suitability of a parameter set for temporal decomposition. This last criterion will be applied to those speech parameter sets, for which resynthesis is possible.

The work reported upon here is part of a project to study the relationship between the target functions and vectors determined by means of

temporal decomposition on the one part, and a phonetic classification of the same utterance on the other. Such knowledge will not only provide deeper insight into the composition of the speech signal, but may also have applications with respect to even more economical methods of speech coding based on phonetic or subphonetic classes, or as a preprocessor for automatic speech recognition and transcription.

In the following sections we will first give a brief description of the temporal decomposition method. Next, we will devote a section to the speech parameters used and their relation to one another. Then, we will analyse the performance of the speech parameters according to the above-mentioned criteria. Finally, we will discuss the results achieved and draw some conclusions about the most convenient parameter spaces in which target functions and vectors should be determined.

2. Temporal decomposition

Temporal decomposition of speech is based on the assumption that, given some suitable parametric representation of the input speech, coarticulation can be described by simple linear combinations of the underlying targets. If we represent the kth target by a target vector $\boldsymbol{a}(k)$, and the movement towards and away from this target by a target function $\phi_k(n)$, the observed speech parameters $\boldsymbol{y}(n)$ can be approximated by the following linear combination of target vectors and functions

$$\tilde{\boldsymbol{y}}(n) = \sum_{k=1}^{K} \boldsymbol{a}(k)\phi_k(n), \quad 1 \leq n \leq N \tag{1}$$

where $\tilde{\boldsymbol{y}}(n)$ is the approximation of $\boldsymbol{y}(n)$. The frame number n represents discrete time and varies between 1 and the total number of frames N of the utterance. The total number of targets within the utterance is given by K. For the speech parameters $\boldsymbol{y}(n)$ any kind of parameter set can be chosen, as will be done in the following sections. In this equation, not only are the target vectors and functions unknown, but also their number and locations.

In solving this equation, all the different target functions $\phi_k(n)$ are first determined using the method described by Van Dijk-Kappers and Marcus (1987, 1989). Next, the target vectors $\boldsymbol{a}(k)$ associated with the target functions $\phi_k(n)$ can be determined by minimizing the mean-squared error E, defined by:

$$E = \sum_{n} [\boldsymbol{y}(n) - \tilde{\boldsymbol{y}}(n)]^2, \tag{2}$$

or, by substituting eq. (1):

$$E = \sum_{n} [\boldsymbol{y}(n) - \sum_{k=1}^{K} \boldsymbol{a}(k)\phi_k(n)]^2. \tag{3}$$

This equation can be solved for the $\boldsymbol{a}(k)$, by setting the partial derivatives of E with respect to $\boldsymbol{a}(k)$ equal to zero. This results in a set of target vectors $\boldsymbol{a}(k)$, each of which consists of a frame of I speech parameters of the same dimension as $\boldsymbol{y}(n)$.

According to eq. (1), the target functions and vectors together give a new representation of the speech parameters which, we hope, will be related to a phonetic representation. An illustration of the decomposition of a speech utterance is given in Fig. 1. The plot shows the amplitude-time waveform of an utterance, the phonetic transcription and the automatically extracted target functions. The log-amplitude spectra corresponding to the target vectors can also be seen in Fig. 1.

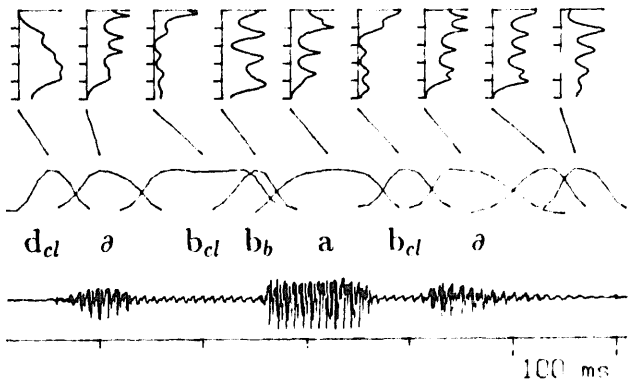

Fig. 1. Temporal decomposition of the *CVC* utterance /dəbabə/. The subscripts *cl* and *b* stand for closure and burst respectively.

3. Speech parameters to be compared

Eight parameter sets were derived from the prediction coefficients a_i obtained with LPC (e.g. Viswanathan and Makhoul, 1975; Markel and Gray, 1976; Vogten, 1983). For the temporal decomposition analysis, the source parameters (the filter gain, the pitch, and the voiced/unvoiced parameter) were left out of consideration. In this paper the prediction order I is always 10, except when specified otherwise, resulting in 10 speech parameters per frame. One parameter set was based on the output of a filter bank. In this set, amplitude information was integrated in the parameters.

3.1. Parameter sets

Four of the LPC-derived parameter sets were directly related to the physical parameters of a model in which the vocal tract consisted of an acoustic tube of I sections, each with the same length but a different cross-sectional area. They were all convertible into one another through linear or non-linear transformations. In this paper they are presented in terms of the LPC coefficients a_i.

(1) Reflection coefficients (RC). RC are often used for speech coding and transmission purposes (e.g. Viswanathan and Makhoul, 1975). The reflection coefficients indicated with the symbol k, have the following recursive relations with the prediction coefficients:

$$k_i = a_i^{(i)},$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \quad 1 \leq j \leq i-1, \tag{4}$$

where the index i takes the decreasing values I, $I-1$, ..., 1 and initially $a_j^{(I)} = a_j$, $1 \leq j \leq I$.

(2) Area coefficients (A). Area coefficients are the cross-sections of the I successive sections of the vocal tube. Equation (5) relates these parameters, A, to the reflection coefficients k:

$$A_{I+1} = 1,$$

$$A_i = A_{i+1} \frac{1 + k_i}{1 - k_i}, \quad 1 \leq i \leq I. \tag{5}$$

If a frame of I A coefficients is considered as a vector in an I-dimensional space, the length of this vector can be varied (within certain limits) without affecting the formant frequencies. The advantage of this property will become clear in one of the following sections. These parameters have been used for speech transmission (Markel and Gray, 1976). Bimbot et al. (1987) have also used the A for temporal decomposition.

(3) Log-area parameters (LA). These parameters, originally proposed by Atal, represent the logarithms of the areas of the cross-sections of the vocal tube, and are thus given by:

$$\log A_i, \quad 1 \leq i \leq I. \tag{6}$$

(4) Log-area ratios (LAR). The frequently used log-area ratios, indicated with the symbol g, can be expressed in terms of the reflection coefficients k:

$$g_i = \log \frac{1 + k_i}{1 - k_i}, \quad 1 \leq i \leq I, \tag{7}$$

or, by substituting eq. (5) into eq. (7), in terms of the area coefficients A:

$$g_i = \log \frac{A_i}{A_{i+1}}, \quad 1 \leq i \leq I, \tag{8}$$

thereby immediately explaining their name. Along with the LA, the LAR are the most frequently used parameters in temporal decomposition and other related techniques (Ahlbom et al., 1987; Bimbot et al., 1987; Chollet et al., 1986; Marteau et al., 1988; Niranjan et al., 1987).

Although eq. (7) suggests otherwise, the relationship between the LAR and the RC is almost linear within a large range of the possible data, as can be seen in Fig. 2. Viswanathan and Makhoul have shown that the LAR provide an approximately optimal set for quantization.

The following five parameter sets are related to the spectral contents of the speech signal.

(5) Formant frequencies (F). The formant frequencies (F) are defined as the acoustical resonances of the vocal tract. In addition to their frequent use by phoneticians, they are often employed in speech synthesizers (e.g. Flanagan,

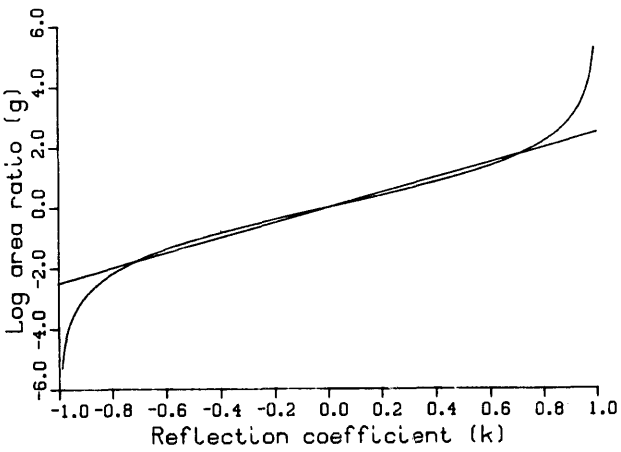

Fig. 2. Log-area ratio (LAR) plotted as a function of the reflection coefficient (RC). For comparison the linear characteristic $g_i = 4k_i$ is also shown.

1972). To determine the F and the associated bandwidths from the LPC coefficients, we used Willems' (1986, 1987) robust formant analysis method. This method, based on the Split Levinson Algorithm, always yields *I*/2 ordered formant tracks. The optimal bandwidth values can then be found in a table. Only the F will be used as input parameters for temporal decomposition.

(6), (7), (8) Three sets of spectral coefficients (S_*). These spectral coefficients were calculated by means of a discrete Fourier transform (DFT) of the prediction coefficients a_i. The transfer characteristics of the prediction filter can be described by a number of log-amplitude Fourier coefficients (I'). Both the prediction order and the number of these Fourier coefficients may vary, yielding the following three sets: $I = 10$ and $I' = 16$ (S_{10-16}), $I = 10$ and $I' = 32$ (S_{10-32}) and $I = 16$ and $I' = 16$ (S_{16-16}). These three different sets (S_*) were used to vary the amount of detail in the input parameters.

(9) Filter bank output parameters (BF). The last set of speech parameters consisted of the band-filter parameters (BF) derived directly from the digitized speech signal. There exists a wide variety of possible filter banks to determine these data, based on different models, each with its own specific advantages. We actually used a 1-Bark bandwidth auditory filter as described by Sekey and Hanson (1984) and yielding 16 parameters per frame.

3.2. Time variations of the speech parameters

As temporal decomposition is based on the assumption that speech production can be described by linear combinations of articulatory target positions, the speech parameters should somehow reflect this linearity. There should therefore be a high linear dependency between the time variations in the different parameters of a single set.

The time variations in the parameters of the sets used in our experiment can be seen in Fig. 3. It can be seen that the time variations of the RC and the LAR (Figs. 3a and b) are almost identical, which is explained in Fig. 2. Clearly, the three S_* sets (Figs. 3c, f and i) are closely related, with S_{16-16} showing more details than S_{10-16} and S_{10-32}. The BF parameters (Fig. 3e), and especially the low-frequent ones (upper lines), also show a resemblance to the coefficients of the S_* sets. However, the BF parameters, vary much more smoothly. The time variations of the *A* coefficients show distinct peaks between rather long periods of almost constant value (Fig. 3d). On the other hand, the LA vary constantly though rather smoothly (Fig. 3g). Finally, the F tracks vary somewhat capriciously but not entirely independently of each other (Fig. 3h).

3.3. An example of the temporal decomposition of a speech utterance

Figure 4 gives an example of the decompositions of the speech utterance /dəbabə/ using five different sets of input parameters. From the top downwards, RC, LAR, A, BF and LA are used. It can be seen clearly that different sets of input parameters yield different results. Not only the number of target functions but also their locations vary considerably from one set to another. Consequently, also the corresponding target vectors may be rather distinct. This variation, however, is far from random. The decompositions of LA and BF seem more closely related to the phonetic structure of the speech signal. RC and LAR nearly always yield more target functions than LA and BF. Furthermore, the decompositions of RC

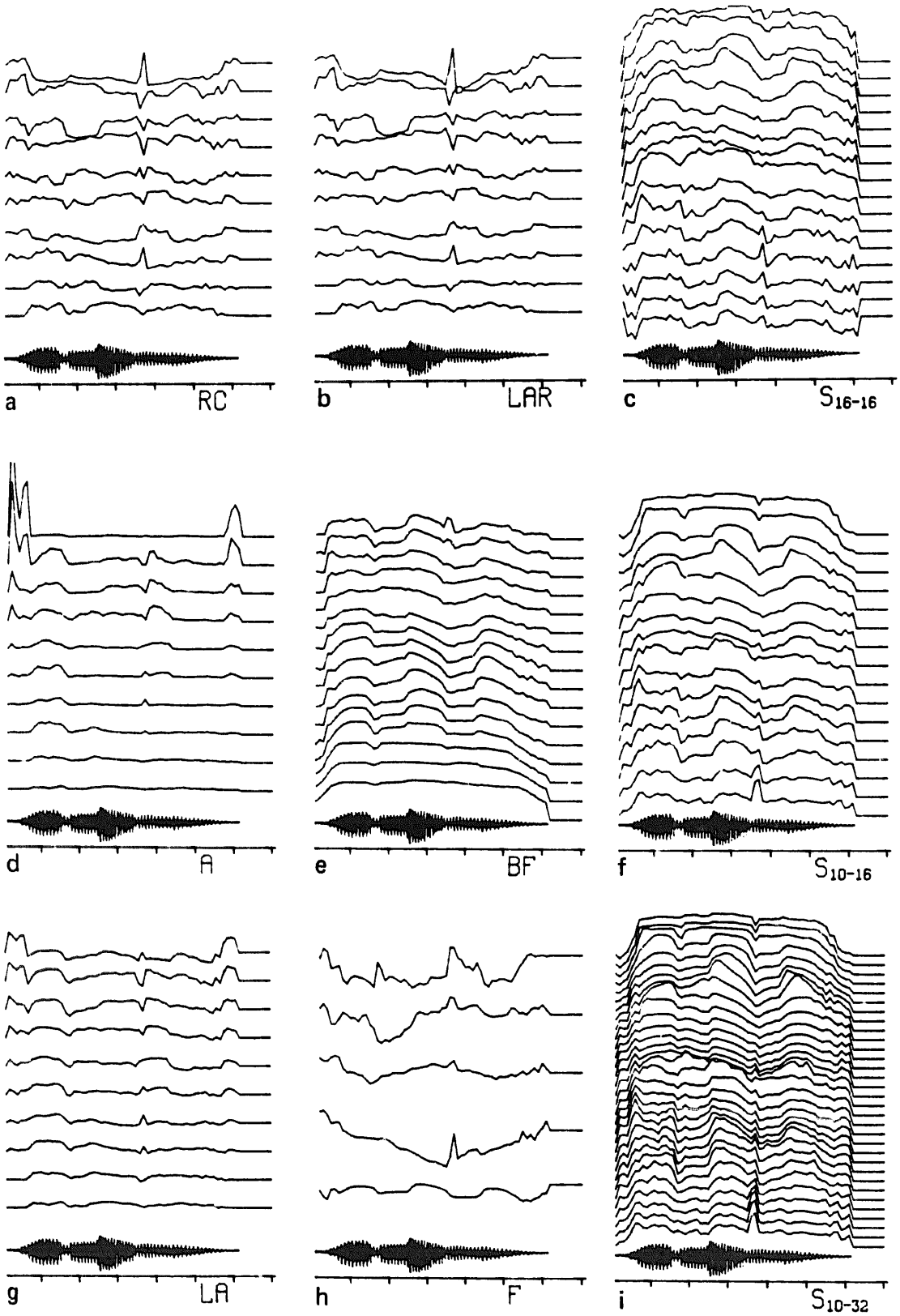

Fig. 3. Time variations of the parameter of several sets and the waveform of the speech signal. In all nine cases the speech utterance is /dələomə/. The time marks are 100 ms apart. (a) RC, (b) LAR, (c) S_{16-16}, (d) A, (e) BF, (f) S_{10-16}, (g) LA, (h) F and (i) S_{10-32}.

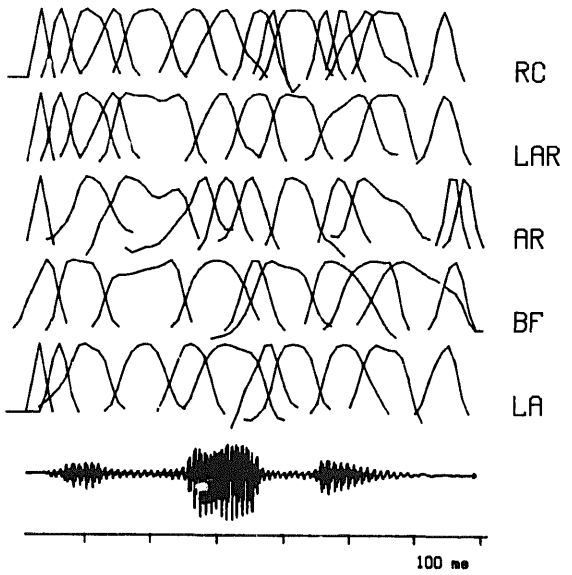

Fig. 4. Temporal decomposition of the speech utterance /dəbabə/ using five different sets of input parameters: reflection coefficients (RC), log-area ratios (LAR), areas (A), band-filter parameters (BF) and log-area parameters (LA).

and LAR are often quite different (as in this example), in spite of the fact that they are nearly identical apart from a scaling factor (see Fig. 2). Thus, the performance of the temporal decomposition method is very sensitive to minor differences in the input parameters. The following sections will deal extensively with these phenomena.

4. Phonetic relevance of the target functions

Since we aim at decomposing the speech signal into phone-like units, one important criterion for good performance is that the number of target functions is approximately equal to the number of phones for each utterance. Also, in the ideal case, the duration of a target function should equal the duration of the corresponding phone and their moments of occurrence should coincide. Here, each target function will be associated with a particular phone, and for each phone the number of target functions associated with it will be counted.

4.1. Experimental procedure

In order to perform this experiment, a small database was constructed consisting of *CVC* combinations embedded in a neutral context: /dəC_1VC_2ə/. The consonants C_1 and C_2 were taken from the phonemes /l/, /m/, /b/ or /p/ and the short vowel *V* was one of the phonemes /a/, /i/ or /o/. The phonemes used in this database are not, of course, representative of all possible phoneme classes, but for practical reasons the size of the database had to be restricted. Each of the 48 possible combinations was produced by a single male speaker.

The phonetic labelling of the *CVC* combinations was carried out by hand. As plosives can be considered as being made up of two articulatory gestures (two phones), closure and burst were labelled separately. Temporal decomposition analysis with the nine parameter sets was carried out for all 48 utterances. Subsequently, every target function was assigned to a phoneme, or, in the case of the plosives, to a subphoneme (closure or burst). As may be gathered from Fig. 4, this assignment was not always a straightforward matter. Sometimes, a target function was located at the transition of two consecutive phones and was thus difficult to classify. As we opted not to label transitions, a decision had to be made in each case as objectively as possible. However, since we were mainly interested in the number of target functions describing a phone, a wrong decision would not substantially influence the results as they were averaged out over all the phones. Furthermore, these transition-describing target functions appeared to occur much more frequently when the overall number of target functions was also relatively high. Thus, incorrect classifications would occur specifically for parameter sets which were obviously not very suitable for temporal decomposition.

4.2. Results

The number of associated target functions for each phone was counted as described above. Next, the percentage of the phones associated with 0, 1, 2 or more target functions was determined for each set of input parameters. The re-

sults are shown in Table 1a. The order in which the data of the various sets are presented may give an indication of the quality of the performance.

Since we were aiming for a one-to-one correspondence of target functions to phonemes (or in case of the plosives, to subphonemes), the second column gives the best indication of the performance. It can be seen that both the BF and the LA give the fair result that about 64% of the phones were described by only one target function and vector. At a distance of some 6% they are followed by the S_* sets and the F. The remaining three sets, consisting of the A, the LAR and the RC, only attained 44%. However, the other columns also reveal important information. A high percentage of phones associated with one single target function is useless if the remaining phones are not associated with any target function at all, and are thus not detected. As can be seen in the first column, all the sets show an unacceptably high percentage of missed phones. However, since the bursts were of only a very short duration and already spread out by the LPC analysis, the question arises whether these phones can be correctly modelled by a target function which, necessarily, has a longer duration. It might be expected, that a fair amount of them would not be detected. Thus it is important to examine whether the high percentages in the first column of Table 1a are indeed due to missed bursts.

In Table 1b the results are shown for the same phones, but excluding the bursts of the plosives. It can be seen clearly that the overall results are much improved, with all the percentages in the first column showing a dramatic decrease while all the percentages in the second column have increased in comparison with the results of Table 1a. For most of the parameter sets, the percentage of missed phones is as low as 0%. In particular the relatively good parameter sets such as BF and LA in Table 1a profit from this altered way of presenting the results, since their percentages of phonemes associated with one single target function increase to 79.0 and 73.2, respectively. Again, the S_* sets and the F form a middle group, while the same three sets as before lag behind. Furthermore, high percentages in the second column correlate with low percentages in the fourth column, indicating that only a small number of phones is associated with more than two target functions.

It should be noted, that in both Table 1 and the following table the remaining percentages of missed phones do not always signify a gap in the sequence of overlapping target functions. Rather, this can be attributed to strong coarticulation, because of which two consecutive phones are associated with the same target function. Only in the case of the closures of voiceless stop consonants of relatively long duration is a real gap sometimes found. This is easily understandable since,

Table 1
Percentages of the phones associated with 0, 1, 2 or more target functions. The results given are averaged over all phones, including and excluding bursts.

	(a) Including bursts				(b) Excluding bursts			
Parameters	0	1	2	>2	0	1	2	>2
BF	19.5	64.9	15.1	0.5	0.0	79.0	20.3	0.7
LA	17.8	63.2	18.4	0.5	1.4	73.2	24.6	0.7
S_{10-16}	19.5	58.4	20.5	1.6	0.0	70.3	27.5	2.2
S_{10-32}	22.2	56.2	21.1	0.5	1.4	69.6	28.3	0.7
S_{16-16}	20.0	51.9	26.5	1.6	1.4	60.9	35.5	2.2
F	14.1	58.9	23.2	3.8	0.0	63.8	31.2	5.1
A	21.1	47.0	29.7	2.2	3.6	53.6	39.9	2.9
LAR	16.2	43.2	34.1	6.5	0.0	45.7	45.7	8.7
RC	14.1	41.6	37.3	7.0	0.0	40.6	50.0	9.4

Table 2
Percentages of the consonants (excluding the bursts) and the vowels associated with 0, 1, 2 or more target functions.

	(a) consonants				(b) vowels			
Parameters	0	1	2	>2	0	1	2	>2
BF	0.0	85.9	13.0	1.1	0.0	65.2	34.8	0.0
LA	2.2	77.2	19.6	1.1	0.0	65.2	34.8	0.0
S_{10-16}	0.0	72.8	23.9	3.3	0.0	65.2	34.8	0.0
S_{10-32}	2.2	69.6	27.2	1.1	0.0	69.6	30.4	0.0
S_{16-16}	1.1	62.0	34.8	2.2	2.2	58.7	37.0	2.2
F	0.0	69.6	23.9	6.5	0.0	52.2	45.7	2.2
A	4.3	58.7	34.8	2.2	2.2	43.5	50.0	4.3
LAR	0.0	46.7	44.6	8.7	0.0	43.5	47.8	8.7
RC	0.0	40.2	48.9	10.9	0.0	41.3	52.2	6.5

in these particular cases, hardly any speech signal exists.

Given the results of Table 1 it will be interesting to investigate whether these results apply for all categories of phonemes. The most obvious division is into consonants and vowels, and the results for these are shown in Table 2. The consonants once again show an increase in the percentages in the second column for almost all the parameter sets. The BF rises as high as 85.9% of the consonants described by only one target function. Here, the LA cannot match the BF, although their 77.2% achievement level is also relatively high. The other parameter sets follow in almost the same order as in the previous table; only the F and the three S_* sets sometimes change places.

From these results it can be concluded that, for the vowels, the percentages in the second column of Table 2b must be lower than the corresponding ones in Table 1b. It is interesting to notice that, for the vowels, the BF and LA results are identical. Also the sets of S_{10-16} and S_{10-32} show similar results.

The percentages given in Tables 1 and 2 are based on the analysis of a small database of carefully spoken isolated words. Thus, conclusions can only be drawn for this kind of speech material. Fluent speech might very well produce different results.

The order in which the results of the several parameter sets are presented in the two tables is representative for the order of performance. The BF are therefore the most suitable input parameters for temporal decomposition if the criterion is a one-to-one correspondence of target functions to phones. The historically most often used LA obtain a second place. A large middle group, consisting of the three S_* sets plus the F, still gives reasonable results. The A, LAR and RC, turn out to be unsuitable for temporal decomposition in this respect.

A possible explanation for the differences between the results of the LA and the BF lies in the fact that in the latter set, amplitude information is integrated in the parameters. In the LA and the other LPC-derived parameters the gain information is left out of consideration. However, although amplitude information might be a useful cue for better temporal decomposition results, it is not a straightforward matter to integrate this information into the parameters.

Although the differences are probably not significant, the order of the S_* sets is nearly always S_{10-16}, S_{10-32}, S_{16-16}. This suggests that if more detail is included in the input parameters, the phonemes tend to be split up into more target functions. Furthermore, it follows that temporal decomposition is sensitive to small differences in the input parameters. This also holds for the results of the LAR and the RC. Although the parameters of both sets are almost identical (see Figs. 2 and 3), the former set always provides a slightly better performance.

5. Phonetic relevance of the target vectors

The target vectors are assumed to model ideal articulatory target positions. It will be clear that this is only possible if for each phone only one target function and thus also one target vector is found. In the previous section we saw that this was not always the case since some of the phonemes in our database were associated with more than one target function while, except for the plosives, these phonemes are thought to be produced by only one articulatory gesture. For this reason, we will restrict ourselves in the following to evaluating only target vectors belonging to phonemes associated with one single target function. Even more restrictively, we will confine ourselves to vowels.

The target vectors have the same dimension as a frame of input parameters. As we started our research using the LA, we will first investigate the interpretation and phonetic relevance of the target vectors determined with the LA. Next, we will extend or adapt our findings to the other sets of parameters.

5.1. Phonetic interpretation of the LA target vectors

Once the target functions had been determined, a target vector was computed for each function by solving eq. (3). In the case of the LA, the target vectors in fact described the shape of

the vocal tract. The ideal articulatory target position or vocal tract shape is, of course, not available as a reference; thus the target vectors have to be tested on their own merits. As the model assumes identical target positions for articulatory gestures producing identical phones, target vectors belonging to the same phones should show a high resemblance. A convenient way of judging this resemblance is in terms of the first two formants.

In order to obtain more vowels associated with a single target function, the database was extended with two more productions of the same utterances by the same speaker. For all the vowels associated with only one target function, the target vectors were transformed from the LA space to the formant and bandwidth space. Subsequently, the first two formants F_1 and F_2 of all these vectors were plotted against each other since these two formants are usually considered as perceptually most relevant for the vowels. The result is shown in Fig. 5, where the target vectors belonging to an /a/ are represented by filled circles (●), to an /i/ by filled squares (■) and to an /o/ by filled triangles (▲). These three groups form three separate clusters of points at places where one might expect them if they really represented the specific vowels. To enable a better appreciation of the location of these clusters, we also show, by way of comparison, the points belonging to the middle frames of the same vowels. These frames, which we used as a reference, were extracted by hand from the original matrix of speech parameters and thus, in contrast to the target vectors, were actually realized in the speech signal. In Fig. 5 the original phonemes /a/, /i/ and /o/ are represented by open circles (○), open squares (□) and open triangles (△), respectively.

In this figure a few things should be noticed. As mentioned earlier, the target vector clusters form three separate groups of points. Also, the original phoneme clusters form separate, slightly more compact groups. However, what is most important is that the two clusters belonging to the

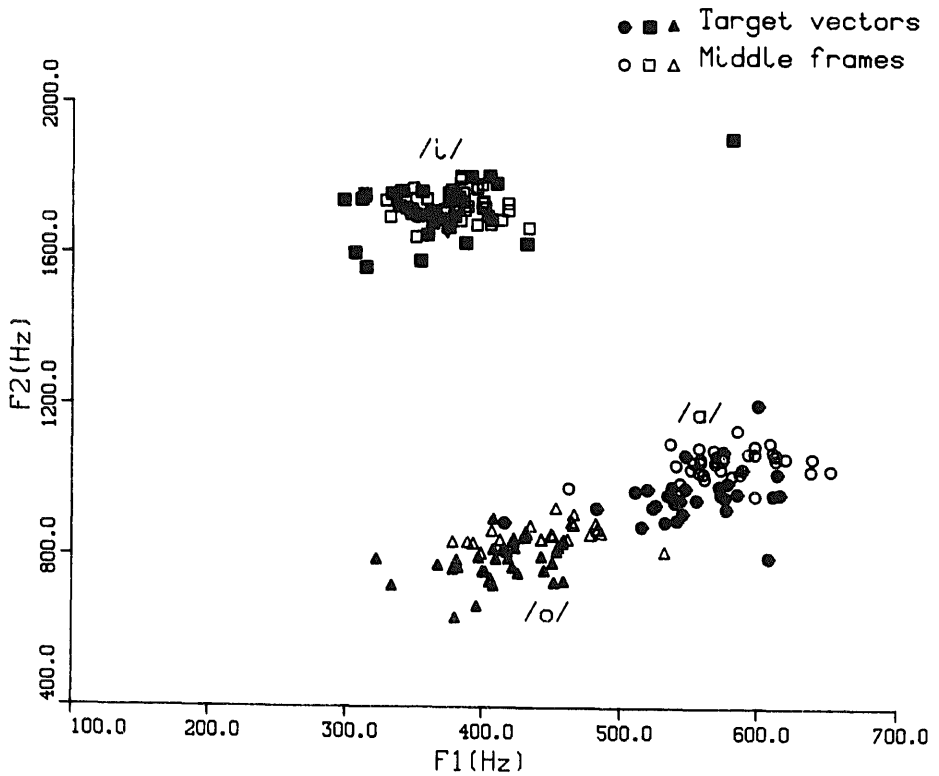

Fig. 5. The first two formants F_1 and F_2 plotted against each other for some target vectors and some middle frames of vowels. The target vectors are associated with the short vowels /a/ (●), /i/ (■) and /o/ (▲). The middle frames are taken from the same vowels: /a/ (○), /i/ (□) and /o/ (△).

same phoneme do not occur at exactly the same location, although there is a fair amount of overlap. This can be seen most clearly for the phoneme /o/; there is not much overlap between the groups of ▲ and △. One might argue that this is due to the fact that the target vectors represent idealized targets and thus are not necessarily realized in the acoustic speech signal. This argument is supported by the fact that the shift of the target vector clusters with respect to the middle frame clusters is away from a neutral vocal tube; that is, the target vector points are more pronounced than the phonemes actually realized. However, in this case one would expect more compact clusters since all the different realizations of the same phonemes are supposed to belong to the same target.

There is yet another plausible explanation. The $\boldsymbol{a}(k)$ are chosen subject to the condition that the product of $\boldsymbol{a}(k)$ and $\phi_k(n)$ approximates the original speech parameters $\boldsymbol{y}(n)$ as closely as possible. Thus, the length of the target factors is determined by both $\boldsymbol{y}(n)$ and $\phi_k(n)$. However, since the original speech parameters are given, the only variable factor can be $\phi_k(n)$. These target functions are normalized to 1, a choice which, although it can be defended, is in fact arbitrary. With this, the length of the target vector is also fixed. This can be seen quite easily in the following extension of eq. (1):

$$\hat{\boldsymbol{y}}(n) = \sum_{k=1}^{K} \boldsymbol{a}(k)\phi_k(n) = \sum_{k=1}^{K} \left(\frac{1}{x} \cdot \boldsymbol{a}(k)\right)(x \cdot \phi_k(n)), \tag{9}$$

where x is an arbitrary positive constant. Changing the normalization factor by a factor x yields target vectors with a length of $\frac{1}{x}$ times the standard length, while for all possible x the resulting approximation of the original speech parameters remains the same.

The effect a change of length of an LA vector has on the position of the vector in the F_1–F_2 plane is shown in Fig. 6. The F_1–F_2 points corresponding to the original length vectors are represented by filled circles (●). Increasing the

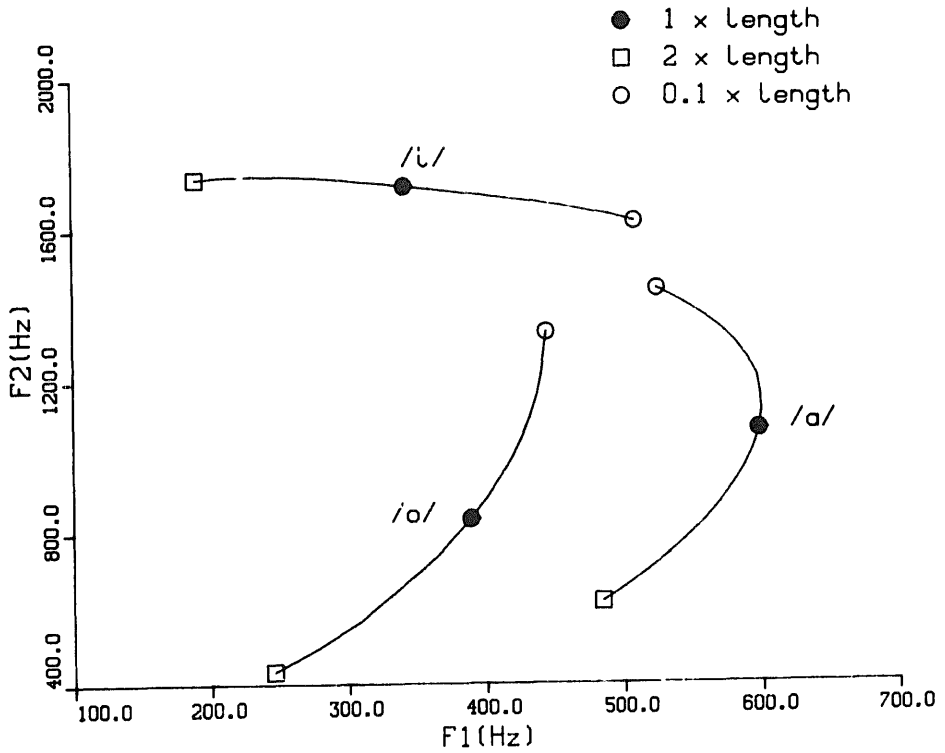

Fig. 6. Tracks in the F_1–F_2 plane of three vectors which are changed in length in the LA space. The three vectors correspond to the three vowels /a/, /i/ and /o/. The F_1–F_2 belonging to the vectors of original length are represented by the filled circles (●). If the length of the vector is doubled, F_1–F_2 take the place of the squares (□). Multiplying the length with 0.1 results in the F_1–F_2 at the places of the open circles (○).

length up to a factor 2 results in the tracks from ● to □, while decreasing the length up to a factor of 0.1 yields the track from ● to ○. Of course, the other formants and the bandwidths change as well, but for the sake of clarity only the effects in the F_1–F_2 plane are shown.

If temporal decomposition is used for coding, the actual choice of x is not important. However, in our case, it introduces an undesired extra degree of freedom and it is important to make a well-considered choice for x. The value of x actually used (namely $x = 1$) is of course based upon defensible grounds: if a target function does not have much overlap with neighbouring functions, the target can be reached. The target vector should resemble the input vectors since the input vectors at that particular point are approximated by the product of only one target function and vector, and this means that the target function has to be normalized to 1. However, in practice, consecutive target functions often show a considerable overlap. In these cases it is less clear what the normalization factor should be. The results given in Figs. 5 and 6 suggest that a normalization factor of slightly more than 1 would be a better choice. The lengths of the target vectors will then be slightly shorter, so causing a shift of the target vector clusters in the direction of the original phoneme clusters. It will be clear that all target vectors require different normalization factors in order to give optimal results. However, in the temporal decomposition method it is impossible to impose boundary conditions which solve this problem satisfactorily. A more detailed study of a wider range of target vectors will be necessary.

5.2. *Target vectors of the remaining parameter sets*

The conclusions with respect to the phonetic relevance of the LA vectors can be extended to RC, LAR and F. As a consequence, a comparative analysis of the target vectors in the F_1–F_2 plane makes no sense. Changing the length of A, S_* or BF target vectors has no effect on the values in the F_1–F_2 plane. However, sometimes the values of the A coefficients turned out to be negative and thus unphysical. Unphysical values were also found for RC and F. It may be clear that target vectors consisting of one or more physically uninterpretable parameters could never model a target position.

Unfortunately, it must be concluded that the phonetic relevance of the target vectors turned out to be an ineffective criterion for the comparison of performance of different speech parameters.

6. Resynthesis

Atal's temporal decomposition method was originally proposed for economical speech coding. Thus, after the decomposition of the speech signal in terms of target functions and vectors, the original speech signal has to be reconstructed or resynthesized. Reconstructed speech parameters, approximating the original ones, can be obtained by substituting the target functions and target vectors in eq. (1). Although it is not our purpose to use temporal decomposition for speech coding, it remains useful to analyse the quality of the resynthesized speech signal. Target functions and target vectors can only model the speech signal in a phonetically relevant way if the speech quality is not too much affected by temporal decomposition. Thus, the quality of the resynthesis gives a good indication of the usefulness of this model.

There are two ways of testing the quality of the resynthesis. First, the speech signal can be evaluated perceptually. However, for different reasons some of the parameter sets are unsuitable for speech resynthesis. In order to resynthesize the signal starting from the BF, a special synthesizer is needed (e.g. Pols, 1977) which, in general, yields unsatisfactory results. But such a synthesizer was not available for our experiments. The S_* coefficients can only be used for resynthesis if phase information is also available, but this is lost during the various stages of the analysis. Lastly, among the reconstructed speech parameter of RC, A and F sometimes unphysical values occur. Such unphysical values have to be corrected before they could be passed on to a speech synthesizer. Thus, only the LA and the LAR can be used without any problems for speech resynthesis.

The second possible way of evaluating the

speech quality consists of determining the difference between the original and the reconstructed speech parameters, using a suitable distance measure. For each frame of each parameter set such a difference or reconstruction error can be determined. However, the comparison of the errors of different parameter sets is only meaningful if these error signals are computed in the same parameter space. Since not all sets are transformable one into another, only LA, A, LAR and RC are compared in this way. The following section deals with this.

6.1. Reconstruction errors in the resynthesis

The LA space is used as reference; all the reconstructed speech parameters are transformed to the LA space. Next, for each frame the difference with the original frame is computed, subject to a suitable distance measure. Of course, a perceptually relevant error criterion would be the most appropriate, but although many attempts have been made (e.g. Gray and Markel, 1976; Nocerino et al., 1985; Applebaum et al., 1987), the definition of such an error does not yet exist. Therefore, we confined ourselves to a simple Euclidian error measure, the same as used in eq. (3). The error $E(n)$ for one particular frame is defined as:

$$E(n) = \left[\sum_{i=1}^{I} (y_i(n) - \tilde{y}_i(n))^2 \right]^{1/2}. \tag{10}$$

Both $y_i(n)$ and $\tilde{y}_i(n)$ consist of LA parameters, but the optimization of $\tilde{y}_i(n)$ (i.e. the determination of the target functions and vectors) has taken place in the various parameter spaces. The $E(n)$ of the various parameters sets can be compared directly. However, comparing these errors per frame is not the most convenient way; it seems better to sum $E(n)$ over a number of frames, obtaining an error measure E_m. As E_m will only be used to get an impression of the differences in reconstruction errors between the various parameter sets, the exact number and choice of frames over which the summation extends is not important. The actually used E_m is defined as follows

$$E_m = \sum_{n=10}^{50} E(n) \tag{11}$$

This choice of E_m is based on the consideration that it can be used for all the *CVC* utterances in the database, and that the summation extends over a relevant part of the utterance. In order to obtain a perceptually more relevant error measure, it is possible to weight the errors of a frame with the gain factor $G(n)$. This follows from the fact that if the amplitude of the speech signal is lower, the relative error will be less audible. This error is defined as:

$$E_A = \sum_{n=10}^{50} G(n)E(n)/1000. \tag{12}$$

The factor 1000 is only meant to bring the value of E_A into the same order of magnitude as E_m. Again, these values are only used to give an indication of the performance of the various parameter sets.

In Fig. 7, a representative example can be seen of the decompositions of the utterance /dəpalə/, using four different parameter sets: RC, LAR, A and LA. Like Fig. 4, this figure shows differences in the number, location and form of the target functions. Next to the target functions the Euclidian error $E(n)$ is shown. The vertical bars under the A-error signal indicate the locations where unphysical (i.e. negative) values were obtained. The numbers at the right side of this figure represent E_m and E_A respectively.

Temporal decomposition attempts to describe speech parameters with a linear model. A parameter set is really suitable for linear modelling if the error signal is small and varies little in time; peaks in the error signal indicate locations where this model is not satisfactory. In the example in Fig. 7 it can be seen that the error signal of the A shows considerable peaks, confirming once more that the A are not very convenient parameters for temporal decomposition. Although none of the other parameter sets produces a constantly small error signal, the achievements of the LA are satisfactory in this respect. Also, the absolute error measures E_m and E_A are smallest for the LA. Due to an almost identical decomposition,

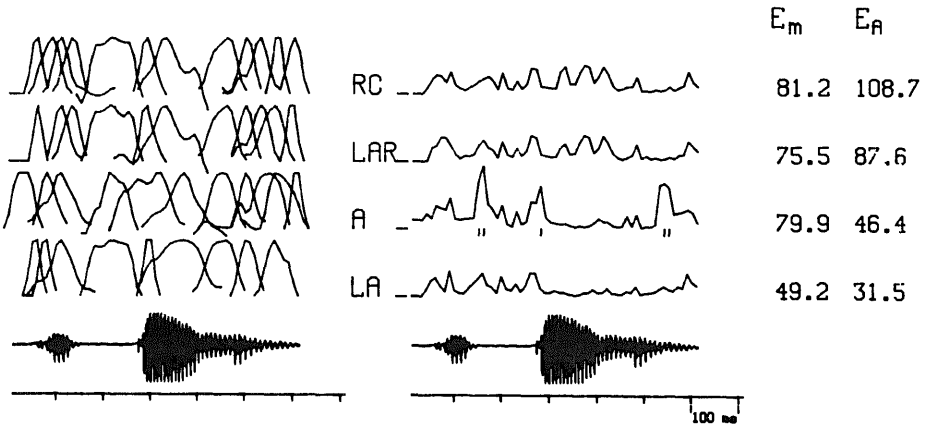

Fig. 7. Target functions belonging to the reflection coefficients (RC), the log-area ratios (LAR), the areas (A) and the log areas (LA), next to the resynthesis error of the *CVC* utterance /dəpalə/. A further explanation of this figure is given in the text.

the error signals of the RC and the LAR are very much alike in this example. However, even here, the error of the RC is the larger of the two.

Most of these observations hold for all examples studied, even when there is a considerable difference in the number of target functions; only in a few cases is the error signal of the LAR smaller than that of the LA. The decoded LAR always describe the speech signal better than the RC, and of the four sets the A usually perform worst in this respect, although this is not visible in this particular example.

6.2. *Reconstruction using mixed parameter spaces*

So far, the determination of the target functions and the subsequent computation of the target vectors has always taken place in the same parameter space. However, since the target functions are dimensionless, it is possible to use them in parameter spaces other than the one in which they have been determined, thus possibly combining the advantages of the two spaces.

Given the results of the previous sections, the LA space is an obvious choice for the determination of the target functions. The target vectors can be computed in the various parameter spaces and the resulting resynthesis errors can be compared using the same error measures as before. An example of this procedure can be seen in Fig. 8a.

It is interesting to note that the LA and LAR error signals are identical. This is no artefact of this example or of the procedure followed, but is rather due to the specific coherence of the two spaces. In the appendix it will be proved that the target vectors, and thus the error signals of these two parameter sets, are always identical if the same target functions are used for the computation. This means that after the determination of the target functions, these two spaces are equally suitable for the computation and interpretation of the target vectors. Moreover, in the previous section it has already been said that the LAR error signal is almost always larger than that of the LA, even if the description in the LA space consists of more target functions. It follows that better target functions can be obtained for the LAR if the decomposition takes place in another parameter space (i.e. the LA space).

Since identical target functions are used in all cases in Fig. 8a, it is possible to compare the RC and LAR error signals directly. Again, the RC error is the larger of the two. This is mainly due to the occurrence of unphysical values, but this is not the only reason. Apparently, the fact that these coefficients differ significantly in the region $0.8 \leq |k_i| < 1$ also plays a role here.

Another example of the decomposition of the same utterance /dəbabə/ is given in Fig. 8b to illustrate once more the described above effects. This time the target functions shown are derived in the RC space, and this yields considerably more target functions. A comparison of the error signals in the LA space demonstrates that the error E_m is almost identical in both cases, while

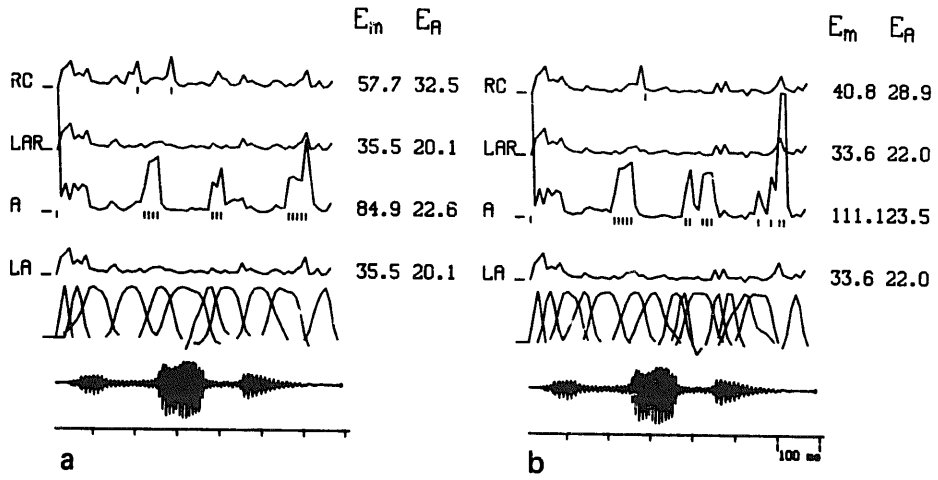

Fig. 8. (a) Target functions of the utterance /dəbabə/ determined in the LA space. Using these functions, the optimal target vectors are computed in the RC, LAR, A and LA space, yielding the plotted error signals. (b) Same as (a), only in this case the target functions are determined in the reflection coefficient space.

the amplitude-weighed error E_A is even larger if the RC functions are used. Of course, the same holds for the error signals in the LAR space. In this particular example the RC functions yield better results in the RC space than the LA functions, but quite often the opposite is true. In most cases, the number of unphysical values is decisive. Unaltered, these effects also apply to the A.

Of the four parameter sets compared in this section, the LA target functions gave the best results, not only here but also in the experiment comparing the phonetic relevance of the target functions. However, in the latter experiment, the BF parameters were found to yield even better results. In the hope of also achieving better results here, a final option that we examined was the use of BF target functions for reconstruction in the LA space. Although this did sometimes lead to better descriptions of the original speech signal, more often the error signals obtained were significantly larger. As it was impossible to transform the reconstructed BF parameters to the LA space we were not able to compare the error signals of both spaces.

7. Discussion and conclusions

In this article it has become clear that every speech utterance can be decomposed in a large number of ways by only varying the choice of input parameters (see e.g. Fig. 4). Variation of the other parameter settings of the method will lead to even more possible decompositions. The criterion to determine the best possible decomposition when optimizing the method was its phonetic relevance. Here, it is shown that speech parameters which yield a phonetically relevant decomposition, also give the best reconstruction results, even compared to decompositions which yielded much more target functions. Thus, it can be concluded that indeed the optimized temporal decomposition method is capable of decomposing the speech signal into units which are closely related to the composition of the speech signal.

One of the objectives of this article was to investigate whether there are speech parameter sets which yield better results than the LA. With respect to the phonetic relevance of the target functions, such a set has been found, namely BF. However, resynthesis of these speech parameters was not possible making them less suitable for temporal decomposition. Unfortunately, the reconstruction errors of this set could not be compared to those of the LA because these sets could not be transformed one into another.

A possible explanation for the differences between the results of the LA and the BF lies in the fact that in the latter set amplitude information is integrated in the parameters. In the LA and the other LPC-derived parameters the gain information is left out of consideration. Although

amplitude information could be a useful cue for better temporal decomposition results, it is not a straightforward matter to integrate this information in the parameters. A pilot experiment (not reported here) in which the LA were weighed with the gain factor to obtain a situation comparable with the BF, did not yield any better LA results.

The three sets of spectral coefficients were taken along in the comparison to study the effect of different amounts of detail in the parameters (i.e. a higher order of DFT or a higher number of parameters) on the decomposition. Although not significant, a tendency was found that more detail leads to more target functions.

The phonetic relevance of the target vectors turned out to be difficult to establish. In the F_1–F_2 plane the LA target vectors belonging to vowels associated with one single target function, formed a cluster of points which, compared to the cluster of middle frames of the same vowels, was slightly shifted and somewhat less compact. From this, it was concluded that apparently the target vectors do not really model idealized target positions. Moreover, nor do they represent the actually realized phone, which is possibly due to a non-optimal normalization of the target functions. Since the target vectors of the other parameter sets either consisted sometimes of unphysical values or could not be transformed to the F_1–F_2 plane, the phonetic relevance of the target vectors could not be used as criterion to distinguish between the various parameter sets.

Also, with respect to resynthesis, the LA turned out to be one of the most suitable parameter sets, often yielding the smallest reconstruction error, although compared to the other sets the reconstruction of the signal was achieved with the fewest number of target functions and target vectors. Better results could also be obtained in other parameter spaces when LA target functions were used. In the LAR space the results are then even identical (see also the appendix). This has to do with the fact that the temporal decomposition method was optimized while using LA. In principle, it must be possible to obtain the same target functions using the LAR. Although the RC are almost identical to the LAR, they invariably perform worse, mainly due to the occurrence of unphysical values. Viswanathan and Makhoul (1975) already reported that for speech transmission the optimal transformation of the RC were the LAR. The A performed worse than the LA, also when using the LA target functions. This can be understood by their logarithmic relationship; if LA parameters are suitable for linear modelling, as a consequence the A coefficients will not be suitable.

Recent literature reports succesful temporal decomposition results using the LAR (Ahlbom et al., 1987; Bimbot et al., 1987; Chollet et al., 1986; Marteau et al., 1988; Niranjan and Fallside, 1987). Although this is not in direct accordance with our results (Table 1), the resynthesis observations of section 6.2 have made clear that, in principle, the same results can be obtained with the LAR as with the LA. However, in their experiments the target vectors are assumed to be known, leaving only the target functions to be determined. As we show in the appendix, identical target functions yield identical target vectors in the LA and LAR space. This also holds the other way round: identical target vectors yield identical target functions. Thus, if temporal decomposition is used in this way, the LAR and the LA will perform equally well, independent of the way the temporal decomposition method is optimized.

Appendix

In this appendix it will be demonstrated that, if a fixed set of target functions and the same speech utterance are used, both calculations in the LA and LAR space yield identical target vectors. The input parameters in the LA space are given by:

$$y_i = \log A_i, \tag{A1}$$

and in the LAR space by:

$$y'_i = \log \frac{A_i}{A_{i+1}}$$

$$= \log A_i - \log A_{i+1} = y_i - y_{i+1}. \tag{A2}$$

In order to determine the target vectors in the LA space, the mean squared error defined by:

$$E = \sum_n [y(n) - \sum_{k=1}^{K} a(k)\phi_k(n)]^2, \quad \text{(A3)}$$

has to be minimized, and this yields the following set of equations (Atal, 1983):

$$\sum_{k=1}^{K} a_{ik} \sum_n \phi_k(n)\phi_r(n) = \sum_n y_i(n)\phi_r(n). \quad \text{(A4)}$$

The components a_{ik} of the target vectors can be determined from these equations. Using the same target functions $\phi_k(n)$ we now have to prove that these vectors are identical to the vectors determined in the LAR space. Following the same procedure we get an equivalent set of equations:

$$\sum_{k=1}^{K} a'_{ik} \sum_n \phi_k(n)\phi_r(n) = \sum_n y'_i(n)\phi_r(n), \quad \text{(A5)}$$

from which the components a'_{ik} can be determined. Expressing $y'_i(n)$ in terms of $y_i(n)$ (eq. (A2)) gives:

$$\sum_{k=1}^{K} a'_{ik} \sum_n \phi_k(n)\phi_r(n)$$

$$= \sum_n (y_i(n) - y_{i+1}(n))\phi_r(n) \quad \text{(A6)}$$

$$= \sum_n y_i(n)\phi_r(n) - \sum_n y_{i+1}(n)\phi_r(n). \quad \text{(A7)}$$

Substitution of eq. (A4) in the right terms of eq. (A7) gives:

$$\sum_{k=1}^{K} a'_{ik} \sum_n \phi_k(n)\phi_r(n)$$

$$= \sum_{k=1}^{K} a_{ik} \sum_n \phi_k(n)\phi_r(n) - \sum_{k=1}^{K} a_{(i+1)k} \sum_n \phi_k(n)\phi_r(n) \quad \text{(A8)}$$

$$= \sum_{k=1}^{K} (a_{ik} - a_{(i+1)k}) \sum_n \phi_k(n)\phi_r(n), \quad \text{(A9)}$$

which subsequently leads to:

$$a'_{ik} = a_{ik} - a_{(i+1)k}. \quad \text{(A10)}$$

Since the same relation holds between the target vectors (eq. (A10) as between the input parameters (eq. (A2)), identical target vectors are obtained for both the LA and the LAR.

Acknowledgements

This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organization for Scientific Research, NWO. The author wishes to thank E. van Mierlo of the University of Utrecht for making available computer programs for the band-filter analyses and F.J. Benning and L.W. Lemmens, students at the University of Technology, Eindhoven, for performing part of the experiments. Many IPO colleagues are also thanked for their comments on various versions of this manuscript.

References

Ahlbom, G., F. Bimbot and G. Chollet (1987), "Modeling spectral speech transitions using temporal decomposition techniques", *Proceedings ICASSP*, pp. 13–16.

Applebaum, T.H., A.H. Hanson and H. Wakita (1987), "Weighted cepstral distance measures in vector quantization based speech recognizers", *Proceedings ICASSP*, pp. 1155–1158.

Atal, B.S. (1983), "Efficient coding of LPC parameters by temporal decomposition", *Proceedings ICASSP*, pp. 81–84.

Bimbot, F., G. Ahlbom and G. Chollet (1987), "From segmental synthesis to acoustic rules using temporal decomposition", *Proceedings 11th ICPHS, Tallinn*, Vol. 5, pp. 31–34.

Chollet, G., Y. Grenier and S.M. Marcus (1986), "Temporal decomposition and non-stationary modeling of speech", *Proceedings 3rd EUSIPCO, The Hague, 2–5 Sept. 1986*, ed. by I.T. Young, J. Biemond, R.P.W. Duin and J.J. Gerbrands (North-Holland, Amsterdam), pp. 365–368.

Flanagan, J.L. (1972), *Speech Analysis Synthesis and Perception* (Springer-Verlag, Berlin, Heidelberg, New York).

Gray, A.H. and J.D. Markel (1976), "Distance measures for speech processing", *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 24, pp. 380–391.

Liberman, A.M., F.S. Cooper, D.P. Schankweiler and M. Studdert-Kennedy (1967), "Perception of the speech code", *Psychological Review*, Vol. 74, pp. 431–461.

Markel, J.D. and A.H. Gray (1976), *Linear Prediction of Speech* (Springer-Verlag, Berlin, Heidelberg, New York).

Marteau, P.F., G. Bailly and M.T. Janot-Giorgetti (1988), "Stochastic model of diphone-like segments based on trajectory concepts", *Proceedings ICASSP*, pp. 615–618.

Niranjan, M. and F. Fallside (1987), "On modelling the dynamics of speech patterns", *Proceedings European Conference on Speech Technology, Edinburgh*, pp. 71–74.

Nocerino, N., F.K. Soong, L.R. Rabiner and D.H. Klatt

(1985), "Comparative study of several distortion measures for speech recognition", *Proceedings ICASSP*, pp. 25–28.

Pols, L.C.W. (1977), *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*, doctoral thesis, (Amsterdam).

Sekey, A. and B.A. Hanson (1984), "Improved 1-bark bandwidth auditory filter", *J. Acoust. Soc. Am.*, Vol. 75, pp. 1902–1904.

Van Dijk-Kappers, A.M.L. and S.M. Marcus (1987), "Temporal decomposition of speech", *IPO Annual Progress Report*, Vol. 22, pp. 41–50.

Van Dijk-Kappers, A.M.L. and S.M. Marcus (1989), "Temporal decomposition of speech", *Speech Communication*, Vol. 8, No. 2, pp. 125–135.

Viswanathan, R. and J. Makhoul (1975), "Quantization properties of transmission parameters in linear predictive systems", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. 23, pp. 309–321.

Vogten, L.L.M. (1983), *Analyse, Zuinige Codering en Resynthese van Spraakgeluid*, doctoral thesis, (Eindhoven).

Willems, L.F. (1986), "Robust formant analysis", *IPO Annual Progress Report*, Vol. 21, pp. 34–40.

Willems, L.F. (1987), "Robust formant analysis for speech synthesis applications", *Proceedings European Conference on Speech Technology*, Edinburgh, pp. 250–253.