Speech Communication 8 (1989) 125–135
North-Holland

TEMPORAL DECOMPOSITION OF SPEECH

Astrid M.L. VAN DIJK-KAPPERS and Stephen M. MARCUS
Institute for Perception Research, 5600 MB Eindhoven, the Netherlands

Received 4 March 1988
Revised 12 January 1989

Abstract. In articulatory phonetics speech is described as a sequence of distinct articulatory gestures, each of which produces an acoustic event that should approximate a phonetic target. Due to the overlap of the gestures these phonetic targets are often only partly realized.

Atal (1983) has proposed a method for speech coding based on so-called temporal decomposition of speech into a sequence of overlapping target functions and corresponding target vectors. The target vectors may be associated with ideal articulatory positions. The target functions describe the temporal evolution of these targets. This method makes no use of specific articulatory or phonetic knowledge. We have extended and modified this method to improve the determination of the number and the location of the target functions and to overcome the shortcomings of the original method. With these improvements temporal decomposition has become a strong tool in analysing speech, from which researchers working on speech coding, recognition and synthesis may profit.

Zusammenfassung. In der artikulatorischen Phonetik wird die Sprache als eine Folge einzelner Artikulationsgesten beschrieben, die jeweils ein akustisches Ereignis zur Folge haben, das eine Annäherung an ein phonetisches Ziel darstellt. Aufgrund der Überlappung der Gesten werden diese phonetischen Ziele oft nur teilweise verwirklicht.

Atal (1983) hat eine Methode zur Sprachcodierung vorgeschlagen, die auf der sogenannten Temporalen Dekomposition der Sprache in eine Folge von überlappenden Zielfunktionen und den entsprechenden Zielvektoren beruht. Die Zielvektoren können idealen Artikulationsstellungen zugeordnet werden. Die Zielfunktionen beschreiben den zeitlichen Entwicklungsverlauf dieser Ziele. Bei dieser Methode werden keine spezifischen Artikulations- oder Phonetik-Kenntnisse angewendet. Wir haben die Methode erweitert und abgeändert, um die Bestimmung von Anzahl und Position der Zielfunktionen zu verbessern und die Nachteile der ursprünglichen Methode zu umgehen. Durch diese Verbesserungen wurde die Temporale Dekomposition zu einem wertvollen Hilfsmittel in der Sprachanalyse, wovon Forschungen in der Sprachcodierung, -erkennung und -synthese profitieren werden.

Résumé. En phonétique articulatoire, la parole est décrite comme étant une séquence de gestes articulatoires distincts, produisant un événement acoustique qui devrait se rapprocher d'une cible phonétique. Du fait du recouvrement des gestes, ces cibles phonétiques ne sont souvent atteintes qu'en partie.

Atal (1983) a proposé une méthode permettant le codage de la parole, basée sur ce que l'on appelle la décomposition temporelle de la parole en une séquence de fonctions-cibles de recouvrement et de vecteurs-cibles correspondants. Les vecteurs-cibles peuvent être associés à des positions articulatoires idéales. Les fonctions-cibles décrivent l'évolution temporelle de ces cibles. Cette méthode ne requiert pas de connaissances articulatoires ou phonétiques particulières. Nous l'avons élargie et modifiée pour mieux déterminer le nombre et la position des fonctions-cibles ainsi que pour corriger les défauts de la méthode originelle. Grâce à ces améliorations, la décomposition temporelle est devenue un outil robuste pour l'analyse de la parole, dont pourraient bénéficier les chercheurs travaillant au codage, à la reconnaissance et à la synthèse de la parole.

Keywords. Temporal decomposition, target positions, speech analysis.

1. Introduction

Articulatory phonetics is based on a description of speech as a sequence of overlapping articulatory gestures. Each gesture produces an acoustic event that should approximate a phonetic target. Adjacent gestures overlap one another, resulting in the characteristic transitions between phonemes that can be observed in almost any parametric representation of the acoustic speech signal. Due to coarticulation and reduction in fluent speech a target may not be reached before articulation towards the next phonetic target begins. It has long been assumed that such targets cannot be determined from the acoustic signal alone, detailed knowledge of the production of all component phonemes being required before the speech signal can be "decoded" (Liberman et al., 1967).

Atal (1983), however, has proposed a so-called *temporal decomposition* method for analysing the speech signal without recourse to any explicit phonetic knowledge. This method takes into account the above articulatory considerations and results in a description of speech as a sequence of overlapping units of variable lengths and located at non-uniformly spaced time intervals.

The temporal decomposition method was developed for economical speech coding; Atal did not attempt to interpret the possibly phonetic meaning of the units. Subsequent work on temporal decomposition, however, focussed on the possibilities with respect to a phonetic interpretation of the units (Marcus and Van Lieshout, 1984; Niranjan and Fallside, 1987). Also, applications in the field of speech synthesis were reported (Chollet et al., 1986; Ahlbom et al., 1987; Bimbot et al., 1987).

The current research developed along the lines of Marcus and Van Lieshout (1984). They realized the possible applications of temporal decomposition in the field of automatic speech transcription or recognition, but also reported quite a few shortcomings from which the method still suffered. The objective of this paper is to propose some improvements and extensions to the original method to overcome these deficiencies. Although some of our choices are initiated by our future intentions with temporal decomposition, namely to derive phonetic information from the acoustic signal in an objective way, these modifications will also be favourable for other possible applications of this technique. As this paper aims at providing precise information about the way these modifications are implemented, we will start with presenting a rather detailed summary of Atal's original method as far as we need this for describing the modifications.

2. Temporal decomposition

Atal (1983) assumed that, given some suitable parametric representation of the input speech, coarticulation can be described by simple linear combinations of the underlying targets. This makes it possible to investigate speech using well-developed methods from linear algebra. Suppose that a given utterance has be n produced by a sequence of K movements aimed at realizing K acoustic targets. Let us denote the speech parameters corresponding to the kth target by a *target vector*, $\boldsymbol{a}(k)$, and the temporal evolution of this target by a *target function*, $\phi_k(n)$. The frame number n varies between 1 and N and is a discrete index of time. Atal's assumption is that we can approximate the observed speech parameters, $\boldsymbol{y}(n)$, by the following linear combination of target vectors and functions

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

or, in matrix notation

$$\tilde{\boldsymbol{Y}} = \boldsymbol{A\Phi}. \tag{2}$$

$\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{y}}(n)$ are approximations of the observed speech parameters. The set of acoustic parameters chosen by Atal to describe the speech signal $\boldsymbol{y}(n)$ are the log-area parameters. These parameters have a close relationship to the positions of the articulators, vary slowly in time and show a high mutual linear dependence, which makes them eminently suited for temporal decomposition (Van Dijk-Kappers, 1988b). These parameters are derived from the filter parameters of an LPC analysis; the source parameters do not play a role in temporal decomposition.

In equations (1) and (2), both the target vectors and functions are unknown and in order to find a suitable solution we have to impose some boundary conditions on the target functions $\phi_k(n)$. Each $\phi_k(n)$ should be non-zero only over a small range of time. Furthermore, at every instant in time only a limited number of target functions may be non-zero. For a moderate speaking rate the number of speech events varies between 10 and 15 per second, so we should expect about 13 target functions to be present in a time interval of 1 second. Given these restrictions, we will solve equation (1) for the $\phi_k(n)$; after that, the optimal acoustic target vectors can be computed.

Equation (1) can be inverted to give the kth target function $\phi_k(n)$ as a linear combination of the speech parameters $y_i(n)$

$$\phi_k(n) = \sum_{i=1}^{I} w_{ki} y_i(n), \qquad (3)$$

where the w_{ki} are a set of weighting coefficients and I is the number of speech parameters. In this equation, only the $y_i(n)$ are known and the w_{ki} have to be chosen so that $\phi_k(n)$ fulfils the requirements of a target function. Since most of the time $\phi_k(n)$ should equal zero, only a limited set of the w_{ki} are non-zero. This can be interpreted as putting a small window over the matrix of speech parameters $\boldsymbol{Y}$. A first useful step in determining the target function is to perform a singular value decomposition (e.g. Gerbrands, 1981; Golub and van Loan, 1983) on the windowed matrix $\boldsymbol{Y}_w$. In matrix notation this can be expressed as

$$\boldsymbol{Y}_w^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}, \qquad (4)$$

where both $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, their columns containing the singular vectors. $\boldsymbol{D}$ is a diagonal matrix of singular values, the square roots of the eigenvalues of $\boldsymbol{Y}_w^T\boldsymbol{Y}_w$. The singular values determine how much of the variance is accounted for by the respective singular vectors. Usually 3 to 5 singular vectors are enough to contain more than 95% of the variance, and we only use these to determine the target function. The operation described above is illustrated for a 210 ms analysis window in Fig. 1, where on the left side the I = 10 log-area parameters determined every 10 ms are shown, and on the right side the singular vectors $\boldsymbol{u}_i$ from the matrix $\boldsymbol{U}$. It can be seen that only the first few singular vectors are important and account for most of the variance.

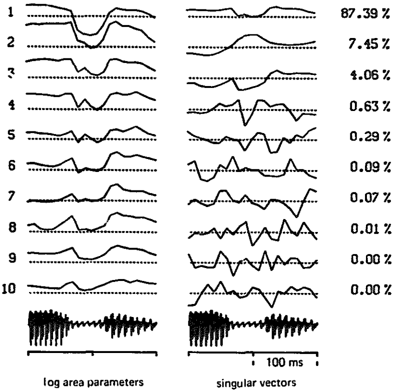

Fig. 1. Plot of the 10 log-area parameters, $y_i(n)$, of a 210 ms window and the singular vectors, $u_i(n)$, of the same speech segment.

It follows from equation (4) that the speech parameters $y_i(n)$ can be expressed as a linear combination of the parameters $u_i(n)$. Substituting this in equation (3) and taking only the s most significant singular vectors results in an important data reduction in solving equation (3). Thus, the target function $\phi_k(n)$ can be represented as

$$\phi_k(n) = \sum_{i=1}^{s} b_{ki} u_i(n), \qquad (5)$$

where the b_{ki} are a set of amplitude coefficients. In order to derive a target function, we have to choose a suitable set of coefficients b_{ki}.

2.1. Determination of the target functions

Atal defines a measure of spread $\theta(n_c)$ as

$$\theta(n_c) = \left[\sum_n \alpha(n)\phi_k^2(n) \Big/ \sum_n \phi_k^2(n) \right]^{\frac{1}{2}}, \qquad (6)$$

where $\alpha(n)$ is a weighting factor. The sum over n extends over the N_w frames of the analysis window of wh.ch n_c is the centre frame. To a certain extent the shape of the target function is determined by the weighting factor $\alpha(n)$. In fact, $\alpha(n)$ can be considered as a model for the target function. In the following section we will discuss Atal's weighting factor and an alternative one.

Depending on the choice of $\alpha(n)$, the spread measure $\theta(n)$ has to be minimized or maximized. In order to obtain the optimal target function, we replace $\phi_k(n)$ of equation (6) by the expression of equation (5), and set the derivatives of $\theta(n_c)$ (or ln $\theta(n_c)$, which gives the same results but with less computational efforts) with respect to the coefficients b_{ki} equal to 0. This results in the eigenvalue equation

$$\boldsymbol{Rb} = \lambda \boldsymbol{b} \tag{7}$$

with eigenvalues λ, where the coefficients r_{ij} of the matrix $\boldsymbol{R}$ are given by

$$r_{ij} = \sum \alpha(n) u_i(n) u_j(n). \tag{8}$$

The smallest (or largest) eigenvalue λ provides the optimal choice of the coefficients b_{ki}, and with equation (5) the target function $\phi_k(n)$ is determined (Lawley and Maxwell, 1971; Atal, 1983).

2.2. Weighting factor of Atal

Atal proposed a quadratic weighting factor:

$$\alpha(n) = (n - n_c)^2, \tag{9}$$

where n_c is the centre of the analysis window. With this $\alpha(n)$ the spread measure $\theta(n)$ should be minimized. Since $\alpha(n)$ is quadratic, it strongly focusses upon target functions centrally located. However, as the target functions are supposed to be related to articulatory gestures, they are in general not expected exactly in the centre of the analysis window. Furthermore, the target functions are forced to be as compact as possible, which impedes the search for speech events of long duration.

2.3. An alternative weighting factor

The weighting factor we propose provides a very simple, rectangular model for a target function:

$$\alpha(n) = 1, \quad \text{for } n_1 \leqslant n \leqslant n_2,$$
$$\alpha(n) = 0, \quad \text{elsewhere.}$$

In this case we have to maximize the spread measure $\theta(n)$ in order to determine the optimal $\phi_k(n)$. Since both location and length of the target function are unknown and differ for each analysis window, the optimal location of (n_1, n_2) is also unknown; an iterative procedure is used to determine the best choice.

The iterative procedure starts with a small rectangular model m_1 (first choice of (n_1, n_2)) in the centre of the analysis window, giving a first approximation $\phi_{k_1}(n)$ of the target function $\phi_k(n)$. The next model m_2 is located between the frames where $\phi_{k_1}(n)$ has the threshold value h_m. This procedure is repeated until the new model m_t equals

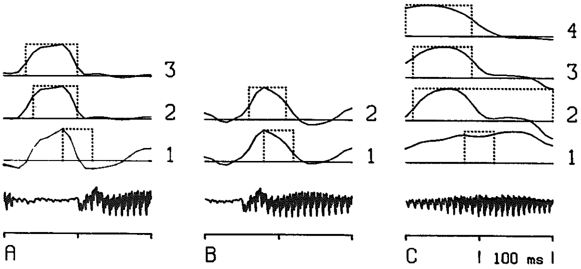

Fig. 2. Target functions resulting from various models for three segments of speech. The initial choice of the model is fixed in the centre of the window, and converges in successive iterations.

the previous model m_{t-1}. In practice, the iterative procedure converges after three to five iterations.

This iteration procedure, consisting of the successive models and the resulting target functions, can be seen in Fig. 2 for three different segments of speech. The choice of the initial model is not too critical, with the only mathematical restriction that it must not extend over the whole window. We found that a suitable length for m_1 is 5 frames around the centre of the window, although a length of 13 frames often gives the same results. A good choice for the value of h_m turned out to be 0.55.

With this procedure a single target function is found within each analysis window. The target function is always normalized to a peak value of 1. This is a reasonable choice, since if the target is reached a single function can describe the length of stay on that target. An unreached target is modelled by the overlap of two or even three target functions. As long as the target itself is unknown, normalization to 1 is the best solution.

2.4. Modification of the analysis window

The use of an analysis window with a fixed length has some serious drawbacks (Marcus and Van Lieshout, 1984). A target function should be only non-zero during a limited number of consecutive frames, but this requirement is not always fulfilled. Sometimes, as in Fig. 2b, the window size is too large, which results in edge effects due to neighbouring speech events. Atal solves this problem by simply truncating the sidelobes; we, however, prefer to adapt the window size to the length of the target function, since within an adapted window there might be a better solution of equation (5). At other times the resulting target function is not complete, as in Fig. 2c, because the window size is too small to accommodate the whole target function. This problem is solved by Atal at a later stage, where he selects a limited number of different target functions. Unfortunately, this procedure does not guarantee the selection of only well-shaped functions, so this presents an additional argument to adapt the window size.

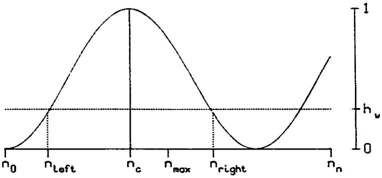

Fig. 3. Schematic overview of the variables used for modifying the window.

In order to adjust the window, the location of the maximum of the target function, $n_{\max}$, within the window (n_0, n_n), is determined. Next we determine the locations of n_{left} and n_{right}, the frames closest to $n_{\max}$ with a value less than the threshold value h_w, to the left and right side of $n_{\max}$ respectively, as shown in Fig. 3. If there is no frame which satisfies the conditions for n_{left}, the first frame, n_0, will be assigned to n_{left}; likewise the last frame, n_n, will be assigned to n_{right} if no frame to the right of $n_{\max}$ has a value smaller than h_w.

As a measure of the amount of (left) edge effects we use $\Sigma_{n=n_0}^{n_{left}} \phi^2(n)$. If this measure exceeds a certain threshold value S, the window needs to be shortened. The new location of n_0 is chosen relative to n_{left}. On the other hand, if $n_{\text{left}} = n_0$, the target function is not complete, and thus the window needs to be lengthened. In our experience the window size has to be increased in very small steps to make sure the adaptation procedure remains stable. However, if the window is really much too small a slightly bigger step provides a faster convergence. According to our measurements, the best choice of the values of the above-mentioned parameters h_w and S was 0.2 and 0.05, respectively (Van Dijk-Kappers and Marcus, 1987).

In this manner, left and right side of the window are modified independently. To prevent that the resulting target function is located completely outside the original window, the value of $\phi(n_c)$ is checked, n_c being the centre of the initial window. This value needs to be above h_w, otherwise the procedure is started all over again with an initial window somewhat smaller than the previous one. This will be repeated as often as necessary.

If one or both of the window sides has been changed, a new singular value decomposition is

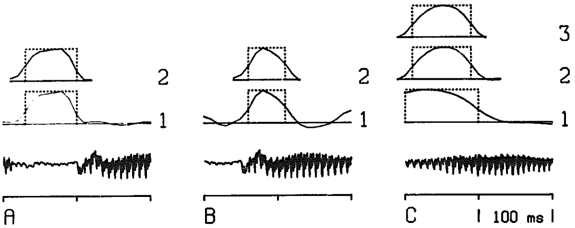

Fig. 4. Iterative modification of the analysis window size and position, for the same three segments of speech as in Figure 2.

performed on the original data within the new window. The most significant singular vectors are again used to construct a target function, but this time the initial model equals the model m_t as obtained from the previous iteration. This procedure is repeated until both sides converge, which usually takes place after two or three iterations.

The results of this window adaptation procedure are shown in Fig. 4, where the same speech segments are used as in Fig. 2. The target functions numbered 1 indicate the resulting target functions of Fig. 2; the higher numbers correspond to the successive results of the model iterations within the modified windows. In all three cases the final window is optimally adjusted to the target function.

3. Analysis of a complete utterance

So far we have only determined a target function for one particular analysis window. In order to analyse an entire utterance, the above procedure has to be repeated with windows located at intervals throughout the utterance. Atal's original method requires the window to be moved in very small steps, of about 10 ms, in order not to miss any functions. An example of the analysis of a number of successive windows is shown in Fig. 5. Although the length of the analysis window may seem flexible, this is only due to a truncation of the sidelobes after the analysis. Furthermore, in spite of the spread measure which attempts to force the function to be located in the centre of the window, the resulting target function regularly lies outside the centre or is not well-shaped.

For comparison, we also show the analysis of the same utterance derived with the modified temporal decomposition method described in the previous sections (Fig. 6). The target functions of a number of adjacent windows are very nearly identical, with only some negligible differences in edge effects. Moreover, an acceptable target function is found for almost every window location.

By shifting the analysis window in steps of 10 ms, the total number of target functions equals

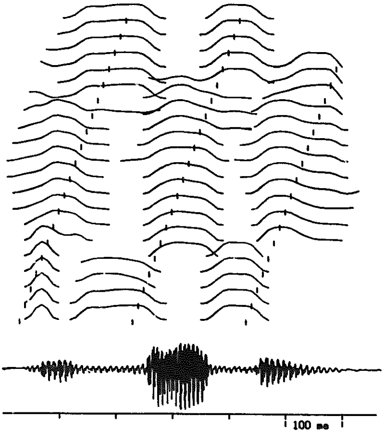

Fig. 5. Target functions determined within the successive analysis windows of the utterance /dəbabə/ using the original method of Atal. The vertical bars indicate the centres of the analysis windows.

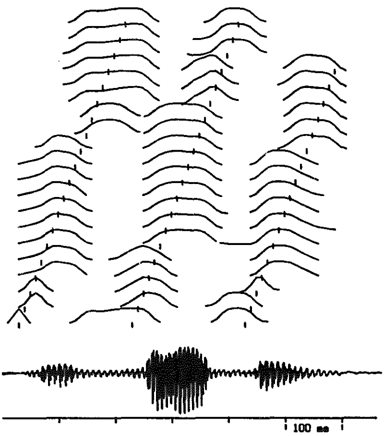

Fig. 6. Target functions determined within the successive analysis windows of the utterance /dɔbabɔ/ using our modified method.

the number of frames. Since many of the target functions describe the same speech event, it is obvious that their number can and has to be reduced. Atal's reduction algorithm will be discussed in the next section, followed by our alternative algorithm.

3.1. Atal's reduction algorithm

Atal has developed a very simple reduction algorithm, which, however, discards a great deal of the relevant information and does not guarantee the selection of only well-shaped functions (Marcus and Van Lieshout, 1984). To determine the locations of the target functions as a function of the centre n_c of the analysis window, Atal uses a timing function $\nu(n_c)$:

$$\nu(n_c) = \sum_{n=n_0}^{n_n} (n - n_c)\phi_k^2(n) \Big/ \sum_{n=n_0}^{n_n} \phi_k^2(n). \qquad (10)$$

The minimum and maximum values of $\nu(n_c)$ are, of course, bounded by the size of the analysis window. According to Atal a speech event occurs every time $\nu(n_c)$ crosses from positive to negative, and he uses this simple criterion to reduce the total number of target functions. Since there is not always a rapid shift from one $\phi_k(n)$ to the next, this timing function could remain nearly constant for some time, without making any zero crossings. This can result in a gap between two selected target functions. Furthermore, it is possible that an incomplete function is selected, while there are much better candidates. Finally, spurious crossings may result in finding the same function twice.

3.2. An alternative reduction algorithm

Although Atal's procedure for selecting the different target functions would probably work without any problems for the target functions determined with our modified temporal decomposition method, it seems a waste of computation time to determine twice or even more often the same target function. Therefore, we have developed a much more efficient method of analysing the whole utterance. Instead of shifting the analysis window by steps of 10 ms, the centre of the next analysis window is located where we expect to find a new $\phi_k(n)$, without skipping any target function. The best choice for this new location turned out to be the n_{right} of the previously found function. Since there exists a small chance of finding the same function once more, the similarity of the two subsequent $\phi_k(n)$'s is tested. As a similarity measure we used the cosine of the angle α between the two $\phi_k(n)$'s, considering them as vectors, where each frame represents a new dimension:

$$\cos\alpha = \sum_n \phi_{k-1}(n)\phi_k(n) \times \left[\sum_n \phi_{k-1}^2(n) \sum_n \phi_k^2(n)\right]^{-\frac{1}{2}}. \qquad (11)$$

The summation extends over the overlapping frames n. If cos α is more than 0.75, the $\phi_k(n)$'s are considered to be similar, and one of them is rejected. In that case the location of the centre of the analysis window is shifted two frames more. It is our experience that this procedure provides a fast determination of all different target functions.

3.3. Determination of the acoustic vectors

For the determination of the acoustic vectors we use the same procedure as proposed by Atal. 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:

$$E = \sum_n \left[y(n) - \hat{y}(n) \right]^2 , \tag{12}$$

or, by substituting equation (1)

$$E = \sum_n \left[y(n) - \sum_{k=1}^{K} \boldsymbol{a}(k)\phi_k(n) \right]^2 . \tag{13}$$

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 (Atal, 1983). This results in a set of acoustic target vectors $\boldsymbol{a}(k)$, each consisting of a frame of 10 log-area parameters.

3.4. Temporal decomposition of a speech utterance

Temporal decomposition of a speech utterance results in a new description of the speech parameters in terms of target functions and vectors which, we hope, will be related to a phonetic description. A few examples of the output of our modified method are shown in Fig. 7. The plot shows the amplitude-time waveform of the utterance, together with the phonetic transcription and the automatically extracted target functions. The 10 log-area parameters of the associated target vectors are transformed into the spectral domain and the corresponding log amplitude spectra are also shown in Fig. 7. In Fig. 7a there is a clear

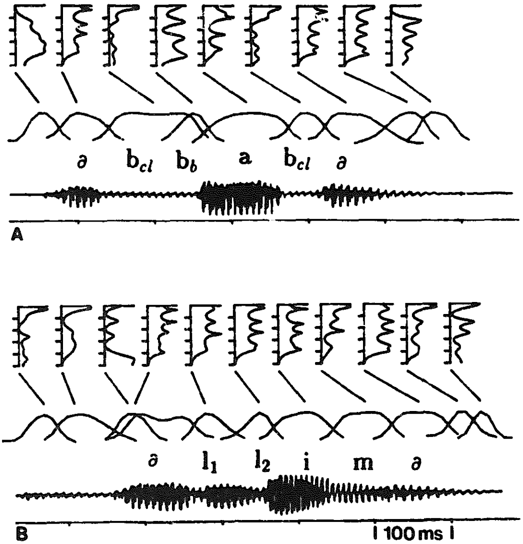

Fig. 7. Temporal decomposition of some *CVC* utterances: (a) /dəbabə/, (b) /dəlimə/. The subscripts *cl* and *b* stand for closure and burst respectively.

correspondence between the target functions and speech events, although a function associated with the burst of the second /b/ is missing. In Fig. 7b there is one speech event described by two target functions.

4. Evaluation and discussion

In several respects, our modified temporal decomposition method gives better results than the original method of Atal. A very important improvement is the fact that now in all situations target functions are found, while in the original method a gap sometimes occurred between two functions. It will be clear that this aspect need not be quantified, each gap being unacceptable irrespective of the intended application. Another improvement is that due to the window convergence procedure the target functions are guaranteed to be well-shaped. Comparing Figs. 5 and 6 will illustrate this. In spite of these modifications, the computation time has more or less remained the same, since the singular value decomposition is the most time-consuming part of the procedure.

4.1. Weighting factor

An improvement which can be quantified is the choice of weighting factor. We stated that the weighting factor or model of Atal tends to yield target functions as compact as possible. Since we want to relate target functions to speech events, this is undesirable. Speech events may have variable lengths and the shortest length is not necessarily the optimal one. In order to be able to compare the performance of Atal's model with our rectangular model, we have embedded both models within our modified method. Thus, resulting differences will only be due to difference in weighting factor.

The criterion for good performance will be the correspondence of target functions to speech events. The target vectors will be left out of consideration. A small database was constructed consisting of *CVC*-comoinations embedded in the context /dəC_1VC_2ə/. The consonants C_1 and C_2 were one of the phonemes /l/, /m/, /b/ or /p/ and the vowel V was one of the phonemes /a/, /i/ or /o/. Each of the 48 combinations was produced by a single male speaker. A phonetic labelling was carried out by hand, closure and burst of the stops being labelled separately. Temporal decomposition analysis using the modified method described above was carried out automatically. A few examples were already shown in Fig. 7, where use is made of the rectangular model.

A tentative phonetic labelling by hand of the target functions was made for each utterance, and Table 1 shows for each weighting factor the percentage of speech events described by 0, 1, 2 or more target functions respectively. Although a reasonable percentage of the speech events is described by only one target function, also an unacceptable percentage of the speech events is missed. However, this percentage is mainly due to missing bursts of the stop consonants which were labelled separately. It is not surprising that these bursts are poorly detected: they are poorly represented by the initial LPC analysis and the temporal decomposition itself results in further smoothing out of such short-duration events.

To get a better idea of the achievements of temporal decomposition, we show the results of the analysis of the same words leaving out the bursts in Table 2. As can be seen, the improvement is considerable; only a very small percentage

Table 1
Percentages of speech events described by 0, 1, 2 or more target functions

	No. of target functions			
	0	1	2	more
Rectangular	17.8	63.2	18.4	0.5
Atal	15.1	55.1	26.5	3.2

Table 2
Percentages of speech events (without bursts) described by 0, 1, 2 or more target functions

	No. of target functions			
	0	1	2	more
Rectangular	1.4	73.2	24.6	0.7
Atal	0.0	60.1	35.5	4.3

of the speech events is really missed and most of the speech events are described by only one target function. Furthermore, although the bursts of the plosives are not considered, this does not lead to problems for the plosives since their closures are always detected. In both Tables 1 and 2 it can be seen that Atal's weighting factor results in more target functions, as we already expected.

To understand why our simple and unrealistic model gives reasonable results, we have to consider equation (5). There, the target function $\phi_k(n)$ is expressed as a linear combination of only 3 to 5 singular vectors, and thus the possible shapes of the target function are limited. Furthermore, although the spread measure will be maximal whenever the target function has an exact rectangular shape, this situation will never be reached given this limited number of possibilities and the fact that the speech parameters vary smoothly in time. A more realistic exponential model gives similar results (Van Dijk-Kappers, 1988a).

4.2. Reduction algorithm

Atal will have discerned some of the shortcomings of his method. In his article he proposed, as an extension, an iterative refinement procedure to refine both target functions and vectors. Indeed, gaps between functions will be filled in by this procedure, but in our opinion the so-obtained target functions look rather distorted, which is an unwanted artefact. Still, this proposal has been followed by several other workers on temporal decomposition (e.g. Ahlbom et al., 1987). Of course, this iterative procedure could also be added to our modified method. Although we do not expect any improvements concerning our intentions with temporal decomposition, other applications, for instance the derivation of rules for synthesis, might profit from it.

Finally, in this respect, we would like to mention an interesting different approach, which unfortunately is not very well documented. Chollet et al. (1986) refer to a clustering technique, applied after the determination of all target functions. Without giving any details, they claim that this technique removes the shortcomings of Atal's selection criterion. It remains to be seen how the target functions obtained with this method compare to our target functions. In any case, the clustering technique causes a substantial increase in computation time.

5. Conclusions

The extended and modified temporal decomposition method makes the determination of the number and the location of the target functions more robust, and does not suffer from most of the problems of the original method of Atal (1983). It can be stated that with these improvements temporal decomposition has become a strong tool in analysing speech, from which researchers working on speech coding, recognition and synthesis may profit.

If we use as a criterion the correspondence of target functions to speech events, the weighting factor we have proposed performs better than the original measure of Atal, which tends to yield too many target functions. Of course, the choice of what is the best weighting factor really depends on the intended applications. For speech coding, more but shorter target functions may give a better speech quality (though less economical). For speech synthesis it might be profitable to have separate functions for transitions from one phoneme to the next. And finally, for speech recognition one target function per speech event might be the best starting point.

For all possible applications it is encouraging that the present outcomes are obtained without making use of any specific phonetic knowledge. Future studies, which may include this knowledge, are needed to examine the achievements of temporal decomposition in more detail and with respect to a particular application.

Acknowledgement

This research was partially supported by the Foundation for Linguistic Research, which is founded by the Netherlands Organization for the advancement of pure research, ZWO. The authors wish to thank L.F. Willems, D.J. Hermes, H.C. van Leeuwen, G. Bloothooft and

R.N.J. Veldhuis for their useful comments on the manuscript.

References

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

B.S. Atal (1983), "Efficient coding of LPC parameters by temporal decomposition", *Proc. ICASSP*, Vol. 2. No. 6, pp. 81–84.

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

G. Chollet, Y. Grenier and S.M. Marcus (1986) "Temporal decomposition and non-stationary modeling of speech", *Proc. 3rd EUSIPCO*, pp. 365–368.

J.J. Gerbrands (1981), "On the relationship between SVD, KLT and PCA", *Pattern Recognition*, Vol. 14, pp. 375–381.

G.H. Golub and C.F. van Loan (1983), *Matrix Computations* (North Oxford Academic, Oxford), pp. 16–20.

D.N. Lawley and A.E. Maxwell (1971), *Factor Analysis as a Statistical Method* (Butterworth, London), pp. 79–82.

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

S.M. Marcus and R.A.J.M. van Lieshout (1984), "Temporal decomposition", *IPO Annual Progress Report*, No. 19, pp. 25–31.

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

A.M.L. van Dijk-Kappers (1988a) "Temporal decomposition of speech: Compactness measures compared", *Proc. of the 7th FASE Symposium*, Edinburgh, pp. 1343–1350.

A.M.L. van Dijk-Kappers (1988b) "Comparison of parameter sets for temporal decomposition", *IPO Report MS 652*, also submitted to *Speech Communication.*

A.M.L. van Dijk-Kappers and S.M. Marcus (1987) "Temporal decomposition of speech", *IPO Annual Progress Report*, No. 22, pp. 41–50.